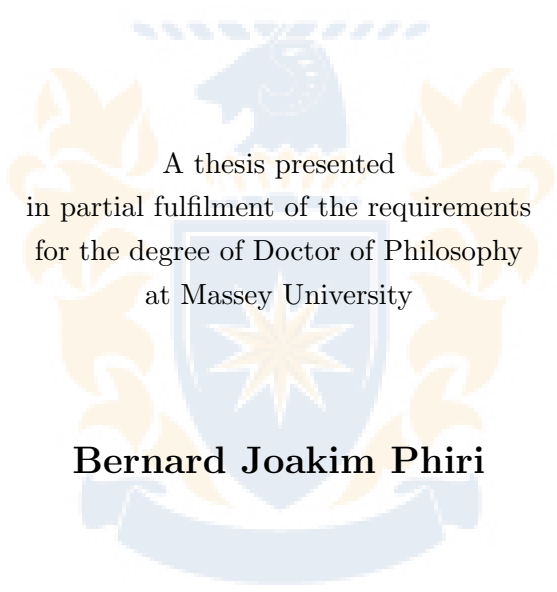


Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# Estimating the public health risk associated with drinking water in New Zealand

The crest of Massey University is centered in the background. It features a blue shield with a yellow star, flanked by yellow leaves and a blue banner at the bottom. Above the shield is a blue horse head facing left.

A thesis presented  
in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy  
at Massey University

**Bernard Joakim Phiri**

Institute of Veterinary, Animal and Biomedical Sciences  
Massey University  
Palmerston North, New Zealand

2015  
(submitted 20 January 2015)



# Preamble

It always seems impossible until its done.

— *Nelson Rolihlahla Mandela (1918–2013)*





# Abstract

This thesis is concerned with the application of both epidemiological and molecular tools to assess the drinking water safety in New Zealand. Compromised drinking water safety is commonly manifested as gastrointestinal illness. The studies in this thesis were motivated by the desire to find ways of reducing the burden of such illness in the human population. Although the studies were conducted in the New Zealand setting the methodologies can be readily applied elsewhere.

The first study investigated the factors associated with the presence of microbes in raw water intended for public consumption. Random forest, an established non-parametric statistical method, was used to model data with possible complex interactions and identified variables that were predictive of the presence of microbes in raw drinking water. *E. coli*, which is widely used as a microbial contamination indicator in the water industry, was found to be a better predictor of the presence/absence of *Campylobacter* (bacteria) than protozoan microbes (*Cryptosporidium* and *Giardia*). This suggests that alternative methods of determining the presence/absence of pathogens in water should be developed. In the second study, the relationship between river flow and reports of cases of gastrointestinal illness was described using the distributed lag modelling approach. This revealed a positive relationship that peaked around 10 days after high flow. Further, the river flow-gastrointestinal illness relationship was stronger in small drinking distribution networks than in large ones. The small drinking water distribution networks could be targetted for facility upgrade in order to enhance their ability to deliver microbiologically safer drinking water.

The third study utilised culture-dependent methods to assess the public health risk associated with drinking water supplied at outdoor recreation facilities — campgrounds. Water treatment using methods such as ultra violet and chemical treatment were found to be highly beneficial for the campgrounds to deliver drinking water that was microbiologically safe and compliant with water safety regulations. The profiles and functional factors of drinking water microbial communities are described in the fourth study. Techniques from the fast-growing field of metagenomics were employed for this purpose. The capability of metagenomic techniques to detect multiple pathogens in a single assay was demonstrated. This has the potential to greatly enhance the specificity and sensitivity of microbial water quality testing.



# Acknowledgements

I could not have accomplished the research work presented in this thesis without the excellent help and guidance that I received from my PhD supervisors. Thank you Nigel French for inviting me to carry out this research work and for teaching me a great deal of things about science. You always had a suggestion on how to move forward when faced with challenges. To Patrick Biggs, thank you for your kind and enthusiastic support even at short notice. You have been inspirational in my approach to bioinformatics and the presentation of genomic information. Thank you Mark Stevenson for your attention to detail and keeping me reminded of the need to apply the epidemiological principles appropriately in my work. To Deb Prattley, thank you for your unfailing support and kind guidance in presenting my research as a coherent story. To Paul Rainey, you provided that critical suggestion that got things moving again when the metagenomic DNA extraction was stalling, thank you.

Thank you to all my fellow postgraduate students at the Epicentre and Hopkirk Research Institute for being part of my journey and sharing your experiences with me along the way. To Christine Cunningham, Wendy Maharey, Jacque Mackenzie and Simon Verschaffelt, thank you for your administrative and computational support. Thank you to the mEpiLab team that provided me with the much needed laboratory support, in particular Angie Reynolds, Anthony Pita, Niluka Velathanthiri, Julie Collins-Emerson, Ann Midwinter, Neville Haack, Errol Kwan, Rhuksana Akhter, Lynn Rogers and Sarah Moore. Special thanks to the New Zealand Genomic Limited team, Lorraine Berry, Richard Fong and Trish McLenachan, for going beyond the call of duty to help me resolve the metagenomic sequencing issues.

I sincerely thank the Allan Wilson Centre and Massey University for funding this project. Thanks to the Department of Conservation for allowing me to conduct my research on their campgrounds. In particular, thanks go to the various campground managers that provided me with campground information and kindly showed me the routes to the water abstraction sites. Thank you to the National Institute of Water and Atmospheric Research as well as the sixteen regional councils that emailed me the river flow data. Thank you to the Institute of Environmental Science and Research Limited for providing me with disease case and drinking water supply data.

Most importantly, many thanks go to my family for being understanding and patient with me as I carried out this work. To my partner Eve, heartfelt thanks for making our home a warm and loving place to live. Completion of this work would have been extremely difficult

without your loving support. To Joe and Sam, thank you for letting your dad complete his PhD research trouble free. It has been a pleasure watching you blossom into fine young men during the four years of my PhD work, I love you very much.

# Acronyms

- mEpiLab*** molecular epidemiology and public health laboratory. 47, 48, 50, 110, 112, 114, 116, 138, 165
- aspA*** aspartase. 116
- glnA*** glutamine synthetase. 116
- gltA*** citrate synthase. 116
- glyA*** serine hydroxy methyl transferase. 116
- pgm*** phospho glucomutase. 116
- tkt*** transketolase. 116
- uncA*** adenosine triphosphate synthase alpha subunit. 116
- BLAST** basic local alignment search tool. 143, 161
- FLASH** fast length adjustment of short reads. 141, 143, 161
- MEGAN** metagenome analyzer. 143, 161
- PAUDA** protein alignment using a DNA aligner. 143, 161
- QIIME** quantitative insights into microbial ecology. 141, 161, 166
- AIC** Akaike information criterion. 86
- ATP** adenosine triphosphate. 8
- BIOM** biological observation matrix. 142
- BLUE** best linear unbiased estimator. 83
- CART** classification and regression trees. 53, 54
- CCA** canonical correspondence analysis. 79, 142, 145, 155, 156, 158
- DAF** dissolved air floatation. 20
- DAPI** 4',6-diamidino-2-phenylindole. 51, 52
- dATP** deoxyadenosine triphosphate. 5, 25
- dCTP** deoxycytidine triphosphate. 5, 25
- ddATP** dideoxyadenosine triphosphate. 5, 6
- ddCTP** dideoxycytidine triphosphate. 5, 6
- ddGTP** dideoxyguanosine triphosphate. 6
- ddNTP** dideoxyribonucleotide triphosphate. 5, 6
- ddTTP** dideoxythymidine triphosphate. 6
- dGTP** deoxyguanosine triphosphate. 5, 25
- DLM** distributed lag model. 84–86
- DLNM** distributed lag non-linear model. 85–87, 94, 95, 100–104, 174–180
- DNA** deoxyribonucleic acid. 5–10, 25, 26, 35, 40, 112, 114, 115, 136–140, 149, 157, 158, 160, 164, 165, 194–197
- dNTP** deoxyribonucleotide triphosphate. 5, 7, 8, 25
- DOC** Department of Conservation. 105–107, 121, 127–130, 137, 138, 155, 158, 159
- dsDNA** double stranded DNA. 6
- dTTP** deoxythymidine triphosphate. 5, 25
- DWSNZ** drinking water standards for New Zealand. 17, 28, 33, 41, 115, 120, 122, 128, 129, 160
- ELISA** enzyme-linked immunosorbent assay. 24–26, 42
- emPCR** emulsion polymerase chain reaction. 8
- ESR** Institute of Environmental Science and Research Limited. 17, 78
- ESRI** Environmental Systems Research Institute. 53, 108
- FC** faecal coliform. 34
- FISH** fluorescence *in situ* hybridisation. 25, 42
- GDH** glutamate dehydrogenase. 114
- GDP** gross domestic product. 105
- GLM** generalised linear model. 119, 120, 125, 127
- GLMM** generalised linear mixed model. 57, 58, 67, 68, 71, 73, 119, 120, 127, 128
- gp60** 60-kDa glycoprotein. 26
- GPS** global positioning system. 84, 107, 108, 110, 162, 163, 166
- HACCP** hazard analysis critical control point. 4
- HAdV** human adenovirus. 22
- HPyV** human polyomavirus. 22
- HSP** heat-shock protein. 35
- IMS** immunomagnetic separation. 51
- ISFET** ion-sensitive field-effect transistors. 8
- LAMP** loop-mediated isothermal amplification. 26, 42, 43
- LDA** linear discriminant analysis. 79
- LSU** large subunit. 34

- MANOVA** multivariate analysis of variance. 79
- MAV** maximum acceptable value. 115, 120, 122
- mCCDA** modified charcoal cefoperazone deoxycholate agar. 50, 51, 112, 164
- MfE** Ministry for the Environment. 1
- MFT** membrane filter technique. 23, 42
- MLST** multilocus sequence typing. 110, 115, 124, 128, 160, 194, 201
- MoH** Ministry of Health. 16, 17, 46, 78
- MPN** most probable number. 23, 33, 53, 60, 106, 116, 122
- MST** microbial source tracking. 29, 30, 42
- MTF** multiple-tube fermentation. 23, 42
- NCBI** National Center for Biotechnology Information. 143, 149, 156
- NGS** next-generation sequencing. 5, 11, 36, 140
- NIWA** National Institute of Water and Atmospheric Research. 78, 165
- NZGL** New Zealand Genomics Limited. 138, 165
- OD** optical density. 138
- OOB** out-of-bag. 55
- OTU** operational taxonomic unit. 141, 142
- PBS** phosphate buffered saline. 51, 114, 198, 199
- PCA** principal component analysis. 46, 79, 80, 90, 94, 95
- PCR** polymerase chain reaction. 5, 9, 25, 26, 29, 35, 42, 51, 110, 112, 114, 115, 120, 122–124, 128, 139, 140, 156, 165, 167, 194
- PSI** proportional similarity index. 142, 143, 149, 155
- QMRA** quantitative microbiological risk assessment. 22
- qPCR** quantitative real-time polymerase chain reaction. 22, 26, 43
- RAM** random-access memory. 166
- rDNA** ribosomal deoxyribonucleic acid. 34
- REC** river environment classification. 53
- RF** random forest. 46, 53–57, 65, 70–74
- RNA** ribonucleic acid. 25, 40, 138
- rRNA** ribosomal ribonucleic acid. 25, 26, 30, 34–37, 112, 136–140, 145, 155, 165
- SMRT** single molecule real-time. 9
- SNP** single nucleotide polymorphism. 136
- SPInDel** species identification by insertions/deletions. 35
- ssDNA** single stranded deoxyribonucleic acid. 5, 6, 9
- SSU** small subunit. 34
- ST** sequence type. 110, 112, 117, 124, 125
- STEC** shiga toxin-producing *E. coli*. 16, 72
- TC** total coliform. 34
- UK** United Kingdom. 37
- USA** United States of America. 16, 37, 45, 46, 75, 76
- USEPA** United States Environmental Protection Agency. 51, 114
- UV** ultra violet. 17, 21, 97, 120, 121, 127, 129, 130, 159
- VTEC** verocytotoxin-producing *E. coli*. 16, 72
- WGS** whole genome shotgun. 34, 136, 137, 139, 140, 143, 145, 149, 155, 157, 165, 166
- WHO** World Health Organization. 13, 15, 31, 41, 142, 149, 157
- YLL** years of life lost. 75
- ZMW** zero-mode waveguide. 9

# Contents

	i
<b>Preamble</b>	iii
<b>Abstract</b>	v
<b>Acknowledgements</b>	vii
<b>Acronyms</b>	ix
<b>General introduction</b>	
1.1 Background . . . . .	1
1.2 Water quality . . . . .	1
1.2.1 The chemical aspect of water quality . . . . .	2
1.2.2 The physical aspect of water quality . . . . .	2
1.2.3 The biological aspect of water quality . . . . .	3
1.2.4 Genomic sequencing . . . . .	5
1.3 The structure of this thesis . . . . .	10
<b>Literature review</b>	
2.1 Background . . . . .	13
2.2 Drinking water sources and supply in New Zealand . . . . .	16
2.2.1 Drinking water sources . . . . .	16
2.2.2 Drinking Water supply system . . . . .	16
2.3 Drinking water treatment processes . . . . .	20
2.4 Common methods for detecting indicator organisms in drinking water . . . . .	21
2.4.1 Organism isolation-based methods . . . . .	23
2.4.2 Immunological methods . . . . .	24
2.4.3 Gene sequence-based methods . . . . .	25
2.4.4 Microbial compliance criteria for New Zealand . . . . .	28
2.5 Microbial source tracking . . . . .	29
2.6 Indicator organism detection in recreational water . . . . .	30
2.7 Pathogens in drinking water — New Zealand . . . . .	31
2.8 Metagenomics . . . . .	34
2.8.1 Metagenomics in drinking water . . . . .	36
2.8.2 Metagenomic research trends . . . . .	36
2.8.3 Microbial community profiles . . . . .	39
2.8.4 Microbial community functional genes . . . . .	39
2.9 Summary . . . . .	40
<b>Factors associated with the presence of pathogens in drinking water sources of New Zealand</b>	
3.1 Background . . . . .	45
3.2 Materials and methods . . . . .	47
3.2.1 Study sites . . . . .	47
3.2.2 Sample collection . . . . .	47



## CONTENTS

---

3.2.3	Laboratory procedures . . . . .	49
3.2.4	Data . . . . .	52
3.2.5	Statistical techniques . . . . .	52
3.2.6	Data analysis . . . . .	55
3.3	Results . . . . .	58
3.3.1	Descriptive statistics . . . . .	58
3.3.2	Random forest analysis . . . . .	63
3.3.3	Regression analysis . . . . .	66
3.4	Discussion . . . . .	69
<b>The relationship between river flow and notified cases of gastroenteritis in New Zealand</b>		
4.1	Background . . . . .	75
4.2	Materials and methods . . . . .	77
4.2.1	Study units . . . . .	77
4.2.2	Data . . . . .	78
4.2.3	Multivariate data analysis . . . . .	79
4.2.4	Geostatistical exploration . . . . .	80
4.2.5	Statistical modelling . . . . .	84
4.3	Results . . . . .	87
4.3.1	Descriptive statistics . . . . .	87
4.3.2	Multivariate analysis . . . . .	90
4.3.3	Geostatistical analysis . . . . .	91
4.3.4	Distributed lag analysis . . . . .	93
4.4	Discussion . . . . .	94
<b>The culture-based microbiology of drinking water on campgrounds in New Zealand</b>		
5.1	Background . . . . .	105
5.2	Materials and methods . . . . .	107
5.2.1	Study campground selection . . . . .	107
5.2.2	Campground water catchment geospatial characteristics . . . . .	108
5.2.3	Sample collection . . . . .	109
5.2.4	Laboratory techniques . . . . .	110
5.2.5	Laboratory processing: Faecal samples . . . . .	112
5.2.6	Laboratory processing: Water samples . . . . .	114
5.2.7	<i>Campylobacter</i> MLST . . . . .	115
5.2.8	Public health risk assessment . . . . .	115
5.3	Data analysis . . . . .	116
5.3.1	Regression analysis . . . . .	117
5.4	Results . . . . .	121
5.4.1	Campground descriptive statistics . . . . .	121
5.4.2	Geospatial descriptives . . . . .	121
5.4.3	Water samples . . . . .	122
5.4.4	Faecal samples . . . . .	123
5.4.5	Multilocus sequence typing analysis . . . . .	124
5.4.6	Regression analysis . . . . .	125
5.5	Discussion . . . . .	127
<b>The metagenome of drinking water on campgrounds in New Zealand</b>		
6.1	Background . . . . .	135
6.2	Materials and methods . . . . .	138
6.2.1	Study sites and sample collection . . . . .	138
6.2.2	Laboratory processing . . . . .	138

6.2.3	Metagenomic DNA sequencing . . . . .	139
6.2.4	Sequence Data . . . . .	140
6.2.5	Data analysis . . . . .	141
6.3	Results . . . . .	145
6.3.1	Descriptive statistics . . . . .	145
6.3.2	Public health hazard assessment . . . . .	145
6.4	Discussion . . . . .	155
<b>General discussion</b>		
7.1	Background . . . . .	159
7.1.1	Types of data . . . . .	160
7.2	Challenges and pitfalls . . . . .	162
7.2.1	Sample collection . . . . .	162
7.2.2	Sample processing . . . . .	164
7.2.3	Data management and analysis . . . . .	165
7.2.4	Future research work . . . . .	166
<b>Appendix</b>		
A.1	Literature review . . . . .	169
A.2	River flow study . . . . .	172
A.3	Catchment study . . . . .	181
A.4	Campground study . . . . .	192
References		

## List of Figures

2.1	The Waitakere and Waikato public drinking water catchments . . . . .	18
2.2	Schematic representation of the Wellington area drinking water distribution network . . .	19
2.3	Schematic diagrams showing the three major parts of a nucleotide . . . . .	27
2.4	Schematic representation of the polymerase chain reaction process. . . . .	28
2.5	Number of 16S and metagenomic publications per calendar year . . . . .	37
2.6	Number of 16S and metagenomic publications in the top fifteen countries . . . . .	38
2.7	Top twenty peer-reviewed journals publishing articles on 16S and metagenomics articles .	38
3.1	Location of the twenty study drinking water sources . . . . .	49
3.2	Schematic representation of a basic decision tree . . . . .	54
3.3	Drinking water catchments with high <i>E. coli</i> concentrations . . . . .	61
3.4	Concentrations of <i>Cryptosporidium</i> and <i>Giardia</i> in study catchment samples . . . . .	62
3.5	Percentage of positive samples for the four study microbes for each season . . . . .	63
3.6	Variable importance scores for drinking water catchment geospatial attributes . . . . .	65
3.7	Random effects for the generalised linear mixed models . . . . .	68
4.1	Water distribution zones and abstraction points . . . . .	78
4.2	Number of gastrointestinal cases 1997–2006, New Zealand . . . . .	88
4.3	Twenty zones with the highest incidence rates during the study period . . . . .	90
4.4	Location of water distribution zones with the highest gastroenteritis incidence rates . . .	91
4.5	PCA biplot of gastrointestinal illness annual incidence rates . . . . .	92
4.6	Median annual gastrointestinal illness case incidence rates, 1997–2006, New Zealand . . .	98
4.7	Kriged median annual gastrointestinal illness case incidence rates . . . . .	99
4.8	Relationship between distributed lag river flow and gastrointestinal illness, New Zealand .	100
4.9	Relationship between distributed lag river flow and gastrointestinal illness, S00079 . . . .	101
4.10	Relationship between distributed lag river flow and gastrointestinal illness, S00118 . . . .	102
4.11	Relationship between distributed lag river flow and gastrointestinal illness, S00217 . . . .	103
4.12	Relationship between distributed lag river flow and gastrointestinal illness, S00735 . . . .	104
5.1	Map showing the location of study campgrounds in New Zealand . . . . .	109
5.2	Types of samples collected from the study campgrounds . . . . .	111
5.3	Flow diagram showing the <i>Campylobacter</i> taxonomic designation process . . . . .	113
5.4	Median most probable number of <i>E. coli</i> in campground water samples . . . . .	124
5.5	Minimum spanning tree of campground <i>Campylobacter jejuni</i> and <i>C. coli</i> . . . . .	126
6.1	Flow diagram showing how 16S rRNA gene metagenomes were analysed. . . . .	144
6.2	Flow diagram showing how whole genome shotgun metagenomes were analysed. . . . .	144
6.3	Taxa richness indices for 16S metagenomes . . . . .	146
6.4	Canonical correspondence plot of 16S metagenomes . . . . .	148
6.5	<i>Campylobacteraceae</i> phylogenetic tree constructed using 16S metagenomes . . . . .	150
6.6	NeighborNet trees illustrating divergence of metagenome sources . . . . .	152
6.7	Bubble plot showing the abundance of virulence factors found in WGS metagenomes . . .	153
6.8	Bubble plot showing the abundance of resistance factors found in WGS metagenomes . . .	153
6.9	NeighborNet tree illustrating divergence of virulence factors found in WGS metagenomes .	154

6.10	NeighborNet tree illustrating divergence of resistance factors found in WGS metagenomes	154
A.1	Schematic representation of table connections in a <b>MySQL</b> relational database	172
A.2	Bubble plots of gastrointestinal illness cases, 1997–2006, New Zealand	173
A.3	Relationship between distributed lag river flow and gastrointestinal illness, S00041	174
A.4	Relationship between distributed lag river flow and gastrointestinal illness, S00082	175
A.5	Relationship between distributed lag river flow and gastrointestinal illness, S00106	176
A.6	Relationship between distributed lag river flow and gastrointestinal illness, S00123	177
A.7	Relationship between distributed lag river flow and gastrointestinal illness, S00200	178
A.8	Relationship between distributed lag river flow and gastrointestinal illness, S00233	179
A.9	Relationship between distributed lag river flow and gastrointestinal illness, S00268	180
A.10	Location of drinking water abstraction sites used in the distributed lag analysis	181
A.11	Percentage of positive samples for the four study pathogens for each calendar month	185
A.12	Land cover for the first six study catchments supplying surface raw water	186
A.13	Land cover for the second six study catchments supplying surface raw water	187
A.14	Land cover for the last four study catchments supplying surface raw water	188
A.15	Lithology for the first six study catchments supplying surface raw water	189
A.16	Lithology for the second six study catchments supplying surface raw water	190
A.17	Lithology for the last four study catchments supplying surface raw water	191
A.18	Land cover for study campground catchments located in the North Island, New Zealand	192
A.19	Land cover for study campground catchments located in the South Island, New Zealand	193
A.20	Phred scores for 16S sequences	205
A.21	Phred scores for WGS sequences	205

## List of Tables

2.1	Bacterial pathogens associated with drinking water . . . . .	14
3.1	Description of the twenty study drinking water sources . . . . .	48
3.2	Description of variables used in both RF and regression analyses . . . . .	57
3.3	Percentage of positive samples from the twenty study drinking water sources . . . . .	59
3.4	Number of sampling occasions and positive samples for each drinking water source . . . . .	60
3.5	Random Forest predictions . . . . .	66
3.6	GLMM estimating the presence/absence of <i>Campylobacter</i> in raw water . . . . .	67
3.7	GLMM estimating the <i>E. coli</i> concentrations in raw water . . . . .	67
3.8	GLMM estimating the presence/absence of <i>Cryptosporidium</i> in raw water . . . . .	69
3.9	GLMM estimating the presence/absence of <i>Giardia</i> in raw water . . . . .	69
4.1	Gastrointestinal illness annual incidence rates per 100 000 population for New Zealand and countries of similar socioeconomic status, 2013. . . . .	75
4.2	Description of variables used in a principal correspondence analysis . . . . .	89
4.3	Median drinking water distribution zone populations and median annual cases reported . . . . .	92
4.4	Drinking water abstraction sites used in distributed lag non-linear modelling . . . . .	93
5.1	Description of study campgrounds operated by DOC . . . . .	121
5.2	Number of water samples collected from DOC campgrounds . . . . .	123
5.3	Number of faecal samples, stratified by animal source, collected from DOC campgrounds . . . . .	125
5.4	Multilocus sequence types for faecal and water <i>Campylobacter</i> isolates . . . . .	131
5.5	Multilocus sequence types for <i>Campylobacter</i> isolated from water . . . . .	132
5.6	GLMM estimating the presence of <i>Campylobacter</i> in campground faecal samples . . . . .	132
5.7	GLMM estimating the concentration of <i>E. coli</i> in campground tap water . . . . .	132
5.8	GLM estimating the concentration of <i>E. coli</i> in campground intake water . . . . .	133
6.1	Number of samples sequenced for 16S rRNA gene and whole genome shotgun . . . . .	147
6.2	Bacterial species deposited in the NCBI database matched with metagenome taxa . . . . .	151
7.1	Computer software used for data processing, data analysis and thesis compilation. . . . .	161
A.1	A description of shapefiles used for geospatial data and their sources. . . . .	182
A.2	Geospatial data for sixteen surface water sources monitored for microbial contamination . . . . .	183
A.3	Geospatial data for four groundwater sources monitored for microbial contamination. . . . .	184
A.4	Constituents of the <i>Campylobacter</i> and <i>Giardia</i> polymerase chain reaction master mixes . . . . .	202
A.5	PCR conditions for selected <i>Campylobacter</i> and <i>Giardia</i> . . . . .	203
A.6	Encoding for the four bases (A, C, T, G) and ambiguous DNA sequences. . . . .	204
A.7	The 1-proportional similarity index values used for <i>Campylobacteraceae</i> taxa divergence . . . . .	206
A.8	The 1-proportional similarity index values used for WHO-recognised pathogen taxa divergence . . . . .	206

# One

## General introduction

### 1.1 Background

Water is essential for virtually all forms of life and is highly abundant on Earth, occupying almost three quarters of the globe's area. However, 97 % of this vital resource is saline while only 3 % is freshwater. It is freshwater that is widely used for human consumption while both freshwater and saltwater are used for recreational purposes. Both drinking and recreational water are associated with public health risks due to the presence of contaminants such as chemicals and pathogens. While this thesis is focused on drinking water and its related public health risks, reference will be made to recreational water use where necessary.

### 1.2 Water quality

The quality of water can be characterised in terms of its chemical, physical, and biological properties. Based on these parameters water may be classified as suitable or not suitable for human consumption or recreational use. Like in many other countries worldwide, the New Zealand authorities use water quality parameters to set safety standards for both drinking and recreational water (New Zealand Ministry for the Environment, 2003; New Zealand Ministry of Health, 2008). It is important to monitor water quality parameters against safety standards in order that remedial measures are taken in case the safety standards are exceeded. An example of water quality data usage for monitoring purposes is a study by Ballantine and Davies-Colley (2014) which showed that from 1989 to 2009 the water quality in New Zealand's 35 major river systems was declining. However, in another example of water quality data usage the New Zealand Ministry for the Environment (MfE)<sup>1</sup> stated that in a ten-year period between 2001 and 2011 the water quality from 79 % of about 300 river monitoring sites was stable. Thirteen percent of these sites reported improving water quality indicators while eight percent reported deteriorating quality<sup>2</sup>.

---

<sup>1</sup><http://www.mfe.govt.nz/environmental-reporting/fresh-water/river-condition-indicator/summary-key-findings.html>

<sup>2</sup><http://www.mfe.govt.nz/environmental-reporting/fresh-water/river-condition-indicator/bacteria.html>

### 1.2.1 The chemical aspect of water quality

There is a wide variety of chemicals that are of public health importance in drinking water supply systems. These chemicals can be categorised by their source, e.g. agricultural, industrial, household or water treatment processes. Some chemicals naturally occur in the environment, particularly in rocks and soils (World Health Organization, 2011). Among the agricultural activities that contribute to chemical pollution of drinking water sources are application of fertilisers, manures and pesticides to land in the water catchment. Intensive animal husbandry has also been linked to chemical pollution (Parliamentary Commissioner for the Environment, 2004). Extractive industries, such as mining, and other industries that involve disposal of large amounts of water, such as the construction industry, are examples of possible sources of industrial pollution in waterways. Household-related activities involved in chemical pollution of waterways include inappropriate sewage disposal or sewage leakages, promiscuous solid waste disposal and urban runoff. Inappropriate application of the water treatment process can lead to water treatment chemical residues and by-products appearing in harmful concentrations in drinking water (World Health Organization, 2011).

Chemical contamination in drinking water rarely occurs acutely, instead, it tends to be low-grade over extended periods of time. For this reason, chemical contamination can be difficult to detect and quantify. Prolonged exposure to chemical contamination can lead to chronic illness such as cancer in the consuming public (Villanueva et al., 2014). Examples of chemicals that have been implicated in the occurrence of illness in humans include arsenic and iodinated or nitrogenated disinfection by-products. Arsenic has been linked to the occurrence of urinary, lung and skin cancers (International Agency for Research on Cancer, 2004) while disinfection by-products have been reported to be positively associated with urinary bladder, colon and rectal cancers (Costet et al., 2011).

### 1.2.2 The physical aspect of water quality

The physical aspect of drinking water quality includes colour, odour (off-flavours), taste, suspended solids and turbidity. These are generally considered to be the aesthetic aspect of drinking water with little or no risk to human health (Gray, 1994). However, coloured water with unpleasant odours and tastes is less palatable compared to water that is colourless, odourless and tasteless. Water that is aesthetically unappealing is likely to attract complaints or rejection from the consuming public. In addition, the presence of these factors is an indication of some kind of contamination. For example, algae growth in surface water supplies leads to production of earthy-musty odours. Water treatment chemicals, such as chlorine, and their by-products can cause off-flavours while iron from ageing pipework can impart taint and odour to drinking water.

Sources of off-flavours in drinking water include (Suffet and Rosenfeld, 2007):

- Natural products — these can produce grassy (hay, straw, woody) odours or fishy (rancid) odours.
- Industrial products — chemical (hydrocarbon) and medicinal (phenolic) odours are examples of odours produced by such products.
- Aerobic oxidation — this type of reaction produces chemicals such as geosmin and 2-methyl-isoborneol that in turn produce earthy-musty (mould) odours.
- Anaerobic degradation — this type of reaction leads to production of marshy, swampy, septic or sulphurous odours. Swampy odours come from products such as sulphides and amines while rancid odours are from fatty acids.

### 1.2.3 The biological aspect of water quality

The biological aspect of drinking water quality is based on the presence or concentration of microbes in water. Of most concern are microbes that cause illness. There are many such microbes of public health significance in drinking water and can be broadly categorised as bacteria, helminths, protozoa and viruses. Examples of bacterial pathogens associated with drinking water include *Campylobacter* spp., *Cyanobacteria* spp., *Legionella* spp., *Salmonella* spp. and *Vibrio cholerae*. *Dracunculus medinensis* and *Schistosoma* are examples of helminths of public health significance found in drinking water. *Cryptosporidium* spp., *Entamoeba* spp., *Giardia* spp. and *Naegleria fowleri* are among waterborne protozoal pathogens. Viral infection arising from drinking water can be caused by viruses such as adenoviruses, enteroviruses, noroviruses and rotaviruses (World Health Organization, 2011).

The chief source of infectious microbes in drinking water are human and animal faeces, which are deposited either directly into water sources or away from the water sources. For microbes in faeces deposited away from water sources, some kind of transportation is required to get them into water. During such transportation microbes often undergo a dilution process. Sometimes they are diluted to very low concentrations such that it is difficult to detect them, yet they might be in concentrations high enough to cause illness (Bridle, 2013). Further, some pathogenic microbes can persist in the environment for months or years while others are highly resistant to disinfection. A combination of these factors poses challenges to the effective treatment of drinking water. Therefore, an ideal drinking water treatment regime is one that takes a holistic approach in applying measures for preventing and eliminating microbial contamination throughout the supply system, i.e. from the catchment to the tap. In such an approach the concept of multiple barriers is applied in order to supply microbiologically safe drinking water. The key points within the drinking water



supply system at which preventive or eliminative barriers are placed are identified using an approach similar to the hazard analysis critical control point (HACCP) approach. HACCP is a systematic preventive approach used in the food production industry to achieve food safety (Alimentarius, 2003). In this type of approach, control measures and resources are focused on points within a production system where contamination can be measured and prevented efficiently.

In order to select the most suitable control points in a multiple barrier system, it is important to have an in-depth understanding of the contamination risks at the various stages of a given drinking water supply system. In the catchment, risk assessment includes characterisation of soils and rocks, mapping of both real and potential risk-associated activities such as agriculture, industrial practices or recreational activities. Within the treatment plant, contingency measures should be in place in case of unforeseen occurrences such as a higher than anticipated microbial load in raw water and also operational breakdowns. Ageing infrastructure in the distribution network can be a significant risk factor.

The multiple barrier approach is based on the principle that failure of one barrier can be compensated for by effective operation of the remaining barriers. This reduces the chances of pathogens surviving the treatment process and causing illness in the public. The multiple barrier approach is now regarded as the cornerstone of modern drinking water treatment systems. In general, five types of barriers can be used in a multiple barrier approach (Hrudey et al., 2006; Plummer et al., 2010):

- **Source protection:** This is aimed at keeping the source water as clean as possible with minimal microbial contamination. It may include measures such as fencing off waterways to prevent access by animals.
- **Treatment:** This is aimed at either removing or inactivating pathogens that have managed to find their way into source water. This process typically involves multiple stages, e.g. filtration followed by chlorination, ozonation or ultraviolet radiation.
- **Securing the distribution network:** The purpose of this barrier is to keep pathogens out of the distributed water. It also ensures that appropriate concentrations of treatment chemical residues are maintained throughout the network to kill or inactivate pathogens that survived through the treatment plant.
- **Monitoring programs:** These serve as warning signals when pathogen concentrations surpass acceptable limits. In modern treatment plants these may include equipment fitted with both warning and automatic control devices to remedy the situation.
- **Planning:** A well thought out and practised response should be in place in case of an emergency, adverse conditions or system failures.

In summary, the key features of a good multiple barrier approach are to manage risk in a preventive rather than reactive manner; having several preventive and/or eliminative measures in place; learn from experience; and having a contingency plan for out-of-ordinary eventualities.

### 1.2.4 Genomic sequencing

In order to enhance the preventive aspect of the multiple barrier approach, methods that are highly accurate in identifying the sources of pathogen contamination are required. Among the methods that have been shown to possess this capability are those based on genomic sequencing techniques. The sequencing technology was introduced in the mid-1970's and has continued to develop since then. To date three generations of sequencing technologies are identifiable: first-, second- and third-generation technologies.

#### First-generation

The Maxam-Gilbert (Maxam and Gilbert, 1977) and the Sanger (Sanger and Coulson, 1975) methods are regarded as the *first-generation* sequencing technologies. Newer sequencing technologies, currently in common use, are often referred to as next-generation sequencing (NGS) technologies, however, these may be appropriately referred to as second-generation, third-generation, e.t.c. sequencing technologies. Of the two first-generation technologies, the Sanger method (also known as dideoxy sequencing or chain termination) became the more widely used. It was designed to sequence single stranded deoxyribonucleic acid (ssDNA) in a process similar to that of polymerase chain reaction (PCR) that uses dideoxyribonucleotide triphosphate (ddNTP)s in addition to the normal deoxyribonucleotide triphosphate (dNTP)s to synthesize deoxyribonucleic acid (DNA) chains (refer to Section 2.4.3 on page 25). dNTPs include the four deoxyribonucleotide triphosphates: deoxyadenosine triphosphate (dATP), deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP) and deoxythymidine triphosphate (dTTP). The chemical structures of the four dNTPs are shown in Figure 2.3 on page 27. ddNTPs are similar to dNTPs except that they lack a 3' hydroxyl group (OH) in the chemical structures of dNTPs. When a ddNTP is incorporated in a sequence it prevents the addition of another nucleotide because a phosphodiester bond cannot form without the hydroxyl group on the 3' carbon.

In the original Sanger method, radioactive or fluorescent-labeled primers are first annealed to the target on template DNA strands. Then the solution is divided into four tubes labeled **A**, **C**, **G** and **T**. In each tube all four dNTPs are added together with DNA polymerase and a ddNTP specific to a particular tube. For instance, to tube **A** dATP, dCTP, dGTP, dTTP, DNA polymerase and dideoxyadenosine triphosphate (ddATP) are added. In this case ddATP is specific to tube **A**; dideoxycytidine triphosphate (ddCTP) would be added to tube **C** and so on. The function of the polymerase is to add the dNTPs to a growing

chain of DNA. Occasionally, a ddNTP is incorporated resulting in a chain-termination. This yields fragments of different lengths because the ddNTP is incorporated at random. Because the template DNA is synthesized numerous times, the new chains will terminate at all positions where a given ddNTP can be added. The end result of this DNA synthesis process is double stranded DNA (dsDNA) fragments which are again denatured into ss-DNA. The latter are separated according to their sizes using electrophoresis with contents of each of the four tubes in a separate lane on a polyacrylamide gel. Electrophoresis is sensitive enough to separate DNA fragments that differ by even a single nucleotide. The lengths of the fragments are then used to determine the sequence because the fragment were synthesized from the same starting point (i.e. the primer) and the last nucleotide is known (i.e. one of the four ddNTPs). For example, if two sets of fragments in tube **A** were 10 and 26 nucleotide long, respectively, it means the 10th and 26th nucleotides are **A**'s. This is so because all fragments in tube **A** end in **A**. Thus by determining the length of all the fragments in all the four tubes the nucleotide at any given position can be identified.

During the 1990s improvements were made to the original Sanger method described above resulting in automated sequencing which involves tagging each ddNTP base with a fluorescent dye of a different colour. After DNA replication the fragments are separated by size within thin glass capillaries and the ddNTP is detected by laser excitation. For example, ddATP could be tagged by a red dye, ddCTP by a blue dye, dideoxyguanosine triphosphate (ddGTP) by a green dye and dideoxythymidine triphosphate (ddTTP) by a magenta dye. A fragment of DNA that is 51 base pairs long ending in a blue dye means that the sequence has a **C** at position 51. Among the advantages of the Sanger sequencing technique is that relatively long sequences of DNA (up to 1000 nucleotides) can be sequenced. One of the disadvantages is that a lot of space is required for the reactions that determine the length of the DNA to occur (in capillary tubes). This limits the number of reactions that can be conducted at a time.

## Second-generation

These sequencing techniques overcome some of the limitation encountered by the Sanger technique by not moving the DNA during sequencing. In general, second-generation technologies employ strategies that involve fragmenting genomic DNA into small pieces which are then attached at separate locations on a solid surface. Each DNA segment is then amplified to form clusters or colonies, sometimes called polonies. In this way thousands to millions of polonies can generate templates that are sequenced in parallel (simultaneously) in one run.

The second-generation sequencing technologies are based on two main sequencing chemistry processes: *sequencing-by-synthesis* and *sequencing-by-ligation* (Liu et al., 2012). Sequencing-by-synthesis utilises DNA polymerase or ligase enzymes to synthesize (extend) a DNA

strand. This can be done in two different ways, the first method involves extending the DNA strand one nucleotide (or short oligonucleotide) at a time. The second method involves tagging a nucleotide (or short oligonucleotide) and identifying it as the DNA extension occurs. Sequencing-by-synthesis may also be categorised by the number of DNA strands sequenced. The first category is where a single input DNA strand is sequenced while in the second category, the input DNA strand is replicated into multiple identical copies which are then sequenced. The sequencing may be real-time, in which case the process is not interrupted and incorporated nucleotides are identified on the fly, or synchronous-controlled. Alternatively, the process is interrupted in order to identify the latest included nucleotide. The synchronous-controlled process can be achieved by using a reversible sequence terminator or by adding only one type of dNTP at a time (Fuller et al., 2009).

Reversible termination sequencing technology is a sequencing-by-synthesis approach that infers the sequence of a template by stepwise elongation. It was popularised as a second generation sequencing technology on the Illumina platform. The general reversible termination sequencing process involves (i) immobilising the sequencing templates and primers on a solid support; (ii) primer extension by one base and termination; (iii) recognising the color of the fluorophore carried by the extended base to identify the incorporated nucleotide after washing away the unincorporated nucleotides; (iv) removal of the fluorescent tag and the 3'-O-blocking group; (v) washing again and repeating the aforementioned steps (ii-iv). The whole process can be summarised as an extension-termination-cleavage-extension cycle (Mardis, 2008; Metzker, 2010).

Sequencing-by-ligation starts by hybridising an anchor primer to one of the regions flanking a genomic region to be sequenced. Then a degenerate fluorescently labeled oligonucleotide (e.g. an octamer i.e. 8-nucleotide DNA molecule) is ligated in each cycle of sequencing. The oligonucleotides are labelled with a fluorescent dye according to the identity of specific position(s) within them, for instance the nucleotide at position 5 (Ho et al., 2011; Shendure and Ji, 2008). The oligonucleotides are degenerate for all positions except a single position that is being sequenced. This allows the sequencing of a single position based on the design of the query primer (Fuller et al., 2009). Once a position is sequenced, the anchor primer and oligonucleotide are cleaved off from the DNA. The process restarts, sequencing a different position (e.g. at the  $n-1$  position) by using a different oligonucleotide (Ho et al., 2011).

The second-generation sequencing platforms include 454 (Hoffmann-La Roche; Basel, Switzerland), HiSeq and MiSeq (Illumina Inc.; California, USA), SOLiD (sequencing by oligonucleotide ligation and detection) (Applied Biosystems Inc.; California, USA) and Ion Torrent (Life Technologies; California, USA) (Anderson and Schrijver, 2010; Siqueira Jr. et al., 2012; Thompson and Steinmann, 2010).

The 454 platform is sequence-by-synthesis based and uses emulsion polymerase chain reaction (emPCR) to clonally amplify the fragments and then pyrosequence them. Pyrosequencing works by detecting pyrophosphate that is released when a nucleotide is incorporated to a growing strand of DNA. This process involves a template DNA strand, a dNTP, adenosine triphosphate (ATP) sulphurylase, luciferase, luciferin, DNA polymerase and adenosine 5' phosphosulphate. When a dNTP is incorporated pyrophosphate is released and transformed into ATP by sulphurylase, in the presence of adenosine 5' phosphosulphate. The ATP is a substrate in the reaction in which luciferase converts luciferin into oxyluciferin thereby releasing light which is captured by a camera. The excess bases are removed by pyrase and the process is repeated by the addition of another dNTP.

The SOLiD platform employs the sequence-by-ligation chemistry with a two-base sequencing system (Mardis, 2008). It uses an octamer (8 base DNA strand) that has the first base as the ligation site, the fifth base as the cleavage site while the eighth base is linked to a fluorescent dye. Once the octamer is ligated to the template strand the fluorescent signal is recorded before it is removed through cleavage at the cleavage site. Five rounds of sequencing are conducted in order to determine the sequence of a strand with each successive round placing a primer at a  $n - 1$  position.

Ion Torrent is similar to other second-generation sequencing platforms in that it uses emPCR to amplify template DNA and sequence-by-synthesis to determine the sequence (Rothberg et al., 2011). However, it does not use fluorescence or chemiluminescence to detect the incorporated bases instead it adopts an electrochemical detection system called ion-sensitive field-effect transistors (ISFET). This system detects a hydrogen ion ( $H^+$ ) released each time a nucleotide is added by DNA polymerase during sequencing. Because this does not require detection of light using a camera, it makes sequencing cheaper than using the optic-based technologies. In addition, the camera-free approach results in higher speed of sequencing and smaller instrument size (Liu et al., 2012).

The HiSeq and MiSeq platforms use bridge amplification to clonally amplify the fragments that are then sequenced using the sequence-by-synthesis chemistry. In this method, DNA molecules and primers are first attached on a slide and amplified with polymerase so that local clonal DNA colonies (DNA clusters) are formed. Four types of reversible terminator bases are added among which one is incorporated into the elongating DNA strand while non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labeled nucleotides, then the dye, along with the terminal 3' blocker, is chemically removed from the DNA, allowing for the next cycle to begin. Additional washing is performed before starting the next incorporation step (Bentley et al., 2008; Mardis, 2008). This technology was used in this thesis (metagenomic study — Chapter 6).

### Third-generation

The main feature distinguishing third-generation sequencing technologies from the previous sequencing generations is that they are PCR-free, i.e. template DNA is not amplified prior to sequencing (Rothberg et al., 2011; Shin et al., 2013), leading to overall shortened sequencing time. Some platforms in this generation detect the sequencing signal in real-time. This means that once the sequencing process starts it is not interrupted and the signal is monitored as each nucleotide is added to the growing DNA strand (Liu et al., 2012), hence the name single molecule real-time (SMRT) sequencing technologies (Eid et al., 2009; Metzker, 2009). The third-generation sequencing platforms include HeliScope (Helicos BioSciences Corporation; Massachusetts, USA), GridION (and MinION) (Oxford Nanopore Technologies; Oxford, UK) and PacBio RS (Pacific Biosciences; California, USA).

In the HeliScope technology, template DNA molecules are hybridised to a primer immobilised on disposable glass flow cells and the sequencing-by-synthesis chemistry is employed as described for the second-generation sequencing technologies but on a single strand of template DNA. This platform is said to have the simplest sample requirements of the available technologies: input DNA sample can be in sub-nanogram quantity and of very poor quality, including degraded or modified DNA (Thompson and Milos, 2011; Thompson and Steinmann, 2010).

The PacBio RS platform also uses the sequencing-by-synthesis approach with fluorescently labeled nucleotides. In this system, template DNA is constrained to nanophotonic structures called zero-mode waveguide (ZMW)s and the presence of a fluorescently labeled cognate nucleotide near the DNA polymerase is measured (Ferrarini et al., 2013; Thompson and Milos, 2011). The recording of the sequencing process is in real-time i.e. the activity of polymerase are optically recorded as it incorporates fluorescent nucleotides without interruption (Shin et al., 2013).

The GridION<sup>3</sup> platform uses nanopore sequencing or *strand sequencing*, an optical-free and DNA label-free approach. In this method, nanopores (biopores at the nanoscale) are formed in an electronically resistant membrane surrounded by a physiological fluid enabling ion exchange. A ssDNA molecule is threaded through a protein nanopore (haemolysin —  $\alpha$ HL, isolated from *Staphylococcus aureus*) which contains an enzyme that ratchets the DNA across. As the DNA molecule is threaded through the nanopore it causes a disturbance to a continuous ionic current which occurs because of a voltage applied across the membrane. Electrophysiological techniques are used to detect signature disturbances caused by each type of nucleotide (Liu et al., 2012).

<sup>3</sup><https://nanoporetech.com/about-us/summary>

In summary, currently the second-generation sequencing technologies are in common use. The second-generation sequencing technologies offer much shorter sequencing times (Mardis, 2011; Turner, 2011) and sequence a massive amount of template DNA in a single run at a much reduced cost compared to first-generation sequencing technologies (Wetterstrand, 2014). These factors have allowed more and more researchers to apply sequencing techniques and genomically profile diverse habitats. Such research will accumulate information on microbial community profiles and functional factors that could lead to better understanding of microbial evolution and subsequently lead to development of enhanced pathogen control measures.

### 1.3 The structure of this thesis

The objective of this thesis was to estimate the microbial-related public health risk associated with drinking water in New Zealand and to determine where in the environment such risk originated. A combination of molecular, epidemiological, geospatial and statistical modelling tools were utilised to achieve this objective. Three studies were commissioned for this purpose: *Catchment*, *River flow* and *Campground* studies. The catchment and campground studies were field studies while the river flow study was based on ten-year (1997–2006) routine national surveillance data of human disease cases caused by pathogens that are associated, at least in part, with water (drinking or recreational contact). The campground study is presented in two separate chapters: one describing the culture-dependent microbiological aspect of drinking water and the other describing the metagenomic aspect. Together the three studies provide a holistic perspective of the public health impacts attributed to drinking water supply systems in New Zealand, from the catchment to the community.

Chapter 2 is a review of the published literature regarding drinking water supply and the related public health safety. Chapter 3 (catchment study) is an application of statistical modelling tools to investigate factors within the catchment associated with microbial water quality at the source of the drinking water supply system. Three and a half year's worth of field study data were used to perform the risk assessment. In Chapter 4 (river flow study) statistical modelling techniques not commonly used in veterinary public health were used to gain insight in the relationship between factors at the drinking water source level and the disease burden in the community. The study utilised ten-year (1997–2006) routine national disease surveillance data and river flow recordings for the same period. The culture-dependent microbiological aspect of the campground study is presented in Chapter 5. In this study the potential public health risk at the point of water consumption (i.e. tap) was estimated using conventional culture-dependent microbiology tools. This was in addition to the microbial risk assessment at the point of water abstraction (intake) and within the catchment. The metagenomic aspects of the campground study are presented

in Chapter 6. Here NGS techniques with related emerging analysis tools were used to both estimate public health hazard in drinking water and perform microbial source tracking.

A discussion of the findings of these studies and their implications in the delivery of microbiologically safe drinking water concludes this thesis in Chapter 7. This chapter also considers the study limitations and discusses the challenges encountered during the course of conducting the research presented in this thesis.

The research question addressed in Chapter 3 was ‘What factors in the catchments supplying drinking water to the New Zealand public are associated with the presence of pathogens in raw water?’. In Chapter 4 the null hypothesis was that ‘River flow on drinking water source rivers is not associated with gastrointestinal illness reports in the local communities.’ The research question addressed in Chapter 5 was ‘What are the microbiological public health risk factors associated with drinking water at campgrounds in New Zealand?’ In Chapter 6 the null hypothesis was that microbial communities do not vary with varying environment.





# Two

## Literature review

### 2.1 Background

Water is essential for virtually all forms of life and is highly abundant on Earth, occupying almost three quarters of the globe's area. Although this vital resource exists in abundance, it is not uniformly available for human consumption. For example, only about 3 % is freshwater and the other 97 % is saline. Freshwater, particularly surface water, is frequently contaminated with debris, chemicals and microbes (Calderon, 2000; Smith Jr. and Perdek, 2004) thus requiring some form of treatment before it is safe for human consumption. Estimates by the World Health Organization (WHO) indicate that around 770 million people worldwide did not have access to safe drinking water in 2012 (World Health Organization, 2013). Inaccessibility to safe drinking water is an acute problem in developing countries, but less so in developed countries, including New Zealand, where more than 90 % of the 4.2 million population<sup>1</sup> is supplied with water whose quality is regularly monitored by authorities (New Zealand Ministry of Health, 2014).

Drinking water availability and quality is greatly affected by climatic changes. For instance, heavy runoff<sup>2</sup> leads to increased surface water contamination with organic material including human or animal excreta (Smith Jr. and Perdek, 2004). Previously, studies have evaluated the effect of climate change on drinking water availability and quality (Arnell, 2004; Delpla et al., 2009; Milly et al., 2005; Shen et al., 2008). Milly et al. (2005) predicted a 10–40 % increase in runoff by the year 2050 in some regions that include Eastern parts of Equatorial Africa, the Platine basin and the northern parts of North America and Eurasia. However, southern parts of Africa and Europe, the Middle East and mid-Western areas of North America are likely to experience a 10–30 % decrease in runoff over the same period. Recently, the Intergovernmental Panel on Climate Change (2013) projected that both near-term (2016–2035) and long-term (to the end of the 21st century) global climatic changes are likely to result in more pronounced extreme weather conditions compared to the reference period, 1986–2005. Some regions are likely to experience an increase in mean precipitation, with decreases in some regions while other regions might not experience any change. In general, wet regions are likely to get wetter while dry regions are likely to get

---

<sup>1</sup>Population information obtained from Statistics New Zealand (2013)

<sup>2</sup>Rainfall that is not absorbed by soil

drier (Intergovernmental Panel on Climate Change, 2013, chp. 11–12). In New Zealand, a recent study by McBride et al. (2014) projected that the average annual rates of campylobacteriosis could increase by about 20 % and 36 % for cryptosporidiosis by the year 2090 as a result of climate change.

Contamination of water with human and/or animal faeces represents a continued major threat to public health (Prüss et al., 2002). Table 2.1 lists some of the bacterial pathogens likely to be transmitted through drinking water, indicates whether or not they are likely to be of animal origin (World Health Organization, 2011, pp. 117-119) and shows their taxonomic Class. Minor sources of microbial contamination to drinking water include biofilms, e.g. *Legionella* has been reported to grow within biofilms on the inside of water pipes (Schmeisser et al., 2003), and intermediate hosts, e.g. schistosoma parasites, multiply in aquatic snails (Snel et al., 2009) while some bacteria are carried by free-living amoebae (Thomas et al., 2004). Drinking water is commonly supplied from either surface or ground sources. Microbial contamination of surface water sources occurs in many different ways, for instance, through effluent from animal production units, meat processing plants, sewage treatment plants or directly from animal faeces. Microbial contamination of groundwater sources originates from sources such as wastewater storage facilities (e.g. septic tanks and pit latrines) and various types of land usage including the application of manure or sewage sludge on cropland (Medema et al., 2002).

**Table 2.1:** Bacterial pathogens associated with drinking water; adapted from World Health Organization (2011)

Pathogen	Persistence in water	Animal source	Class
<i>Burkholderia pseudomallei</i>	May multiply	No	<i>Betaproteobacteria</i>
<i>Campylobacter jejuni</i>	Moderate	Yes	<i>Epsilonproteobacteria</i>
<i>Campylobacter coli</i>	Moderate	Yes	<i>Epsilonproteobacteria</i>
<i>Escherichia coli</i> - pathogenic	Moderate	Yes	<i>Gammaproteobacteria</i>
<i>Francisella tularensis</i>	Long	Yes	<i>Gammaproteobacteria</i>
<i>Legionella</i> spp.	May multiply	No	<i>Gammaproteobacteria</i>
<i>Leptospira</i> spp.	Long	Yes	<i>Gammaproteobacteria</i>
<i>Mycobacteria</i> spp.	May multiply	No	<i>Gammaproteobacteria</i>
<i>Salmonella</i> Typhi	Moderate	No	<i>Gammaproteobacteria</i>
Other salmonellae	May multiply	Yes	<i>Gammaproteobacteria</i>
<i>Shigella</i> spp.	Short	No	<i>Gammaproteobacteria</i>
<i>Vibrio cholerae</i>	Short to long	No	<i>Gammaproteobacteria</i>

The microbial burden of water is often diverse and can change rapidly (Brown et al., 1992) depending on factors such as increased inflow of runoff after a storm or increased effluent discharge. During such occasions the sudden increase in microbial load can overwhelm a set treatment regime rendering the entire treatment process ineffective, allowing pathogens to appear in drinking water thereby causing enteric disease to consumers (Auld et al., 2004;

Mackenzie et al., 1994). Many of the current methods used for detecting contamination in water used for human consumption require a number of days between sample collection and production of screening results, which means that a large number of consumers could be exposed before contamination is detected and measures are applied to render water safe for drinking. Since waterborne infections tend to spread very quickly within exposed populations, it is recommended that frequent pre- and post-treatment water tests are conducted (World Health Organization, 2011). It is worth pointing out that although drinking water is an important vehicle for the spread of organisms classified as waterborne pathogens, other routes of transmission do exist such as through consumption of food, and direct contact (person-to-person or animal-to-person). Apportionment of the origin of infection by waterborne pathogens, commonly manifested as enteric disease, among these different routes is usually not a simple task (Craun and Calderon, 2006; Müllner et al., 2009). This is partly due to the fact that many pathogens have multiple hosts.

In order to supply microbiologically safe drinking water, one or more pathogen barriers are required along the water supply chain. These barriers may be designed to prevent contamination or remove pathogens from water. Creation of riparian buffer zones around water sources in which domestic and feral animals are excluded (Hughes and Quinn, 2014) is an example of a preventive barrier. Removal barriers are those employed at water treatment plants, they include filtration, irradiation and chemical treatment e.g. chlorination or chloramination. Water treatment plants commonly use a combination of these methods depending on the anticipated level of source water contamination (Betancourt and Rose, 2004). Although absolute prevention of source water contamination is a desired goal, mere reduction of the contamination may be more practical. Pathogen removal is cheaper and more efficient when the microbial load is reduced in source water (Bouwer and Crowe, 1988; World Health Organization, 2004). Knowledge of activities occurring in the catchment and how they impact on microbial water contamination is important for designing and implementing efficient preventive barriers.

Drinking water management and quality have improved greatly since the days of John Snow and the 1854 cholera outbreak in London, particularly in developed countries. However, waterborne infections have not been completely eliminated. It is estimated that 4% of all deaths and 6% of all illness worldwide are caused by water and hygiene related infections (Prüss et al., 2002). According to the fact sheets on the [WHO website](http://www.who.int/mediacentre/factsheets/fs310/en/index2.html)<sup>3</sup> around 55 million people died from various causes in 2011 and 88% of the diarrhoeal deaths were caused by consumption of unsafe water and poor hygiene. Based on a retrospective longitudinal study of notified enteric disease cases in New Zealand, Ball (2007) estimated that 17 000 cases of waterborne gastroenteritis are notified in New Zealand per year. This is thought to be an underestimate of the actual number of cases, believed to be between 18 000 and 34 000 per

<sup>3</sup><http://www.who.int/mediacentre/factsheets/fs310/en/index2.html>

year (Moore et al., 2010). It has been suggested that the difference between the number of notified and actual cases is due to filtration of cases at various levels of the health system. For example, it is estimated that after adjusting for age, gender and ethnicity only 20 % of acute gastrointestinal cases from all causes consult a medical practitioner and of these only around 23 % submit a faecal sample for laboratory confirmation (Lake et al., 2009). The number of notified cases is based on laboratory confirmed cases. Hospital admissions due to infectious enteric disease in New Zealand increased steadily during the twenty-year period up to 2008 (Baker et al., 2012). Although the study of Baker and co-workers did not apportion the number of admissions due to waterborne infection specifically, it could be assumed that the portion transmitted through water had also been increasing as waterborne pathogens are among the causes of infectious enteric disease. Among the most reported causes of acute gastrointestinal illness in New Zealand are campylobacteriosis, cryptosporidiosis, gastroenteritis of unspecified cause, giardiasis, salmonellosis, shigellosis, verocytotoxin-producing *E. coli* (VTEC)/shiga toxin-producing *E. coli* (STEC) infection and yersiniosis (Environmental Science and Research, 2014; Lake et al., 2010). Elsewhere, similar causes have been associated with waterborne disease outbreaks. For example, in 2007–2008 waterborne bacterial disease outbreaks in the United States of America (USA) included campylobacteriosis, *E. coli* O157:H7 infection, legionellosis, *Providencia* infection and salmonellosis (Brunkard et al., 2011; Hlavsa et al., 2011). In England and Wales, outbreaks of waterborne infections were caused by *Campylobacter* spp., *Cryptosporidium* spp., *E. coli* O157, *Giardia* spp. and *Astrovirus* during the period 1992–2003 (Smith et al., 2006).

## 2.2 Drinking water sources and supply in New Zealand

### 2.2.1 Drinking water sources

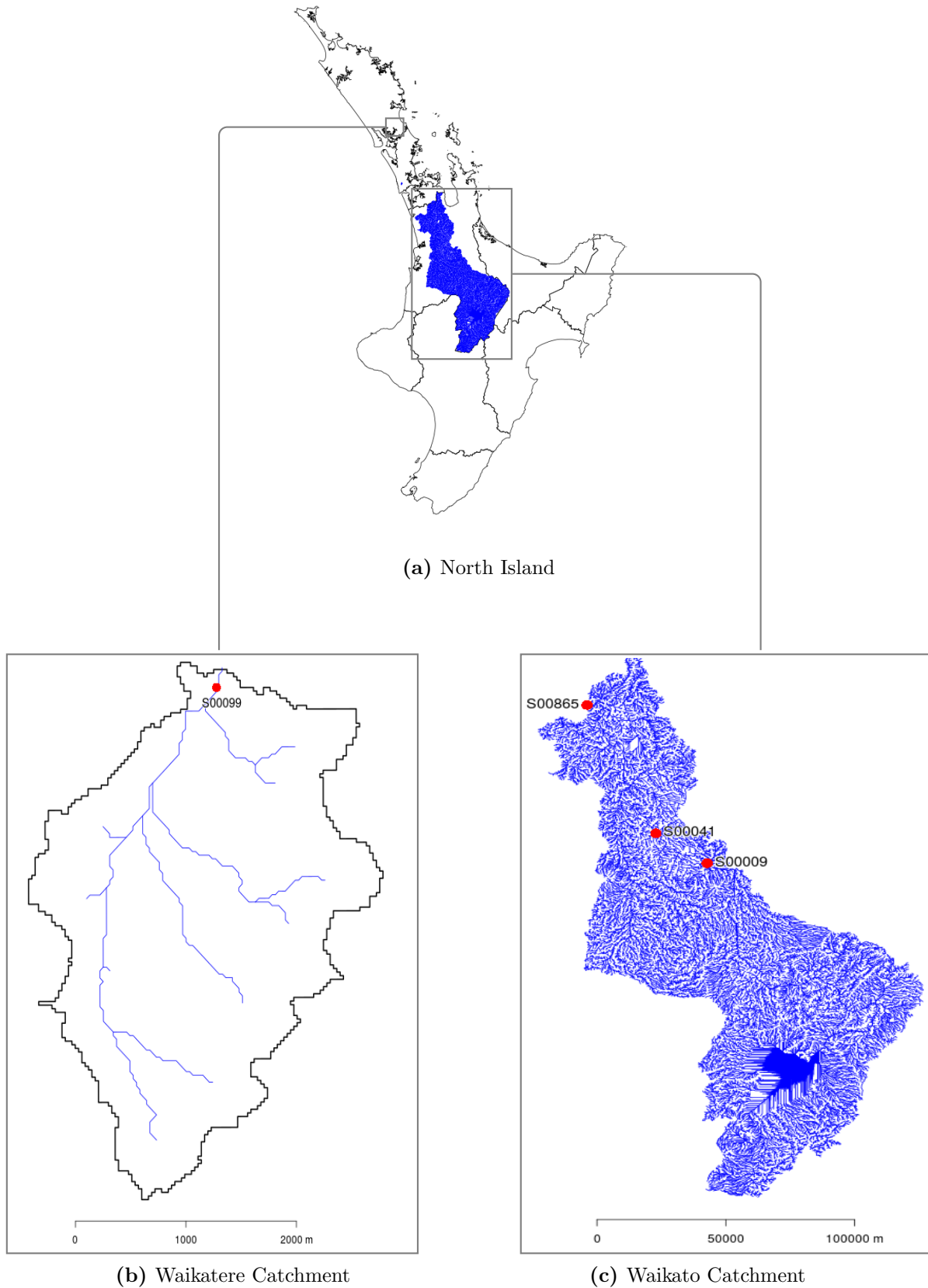
In New Zealand drinking water is mainly abstracted from surface and underground sources with rain water being harvested from household roofs on a smaller scale. Surface water sources include creeks, streams, rivers and lakes while ground sources include wells, boreholes, springs and aquifers. The drinking water source catchments vary in size, for example, the largest included in the current research, the Waikato river catchment, measured 14 000 square kilometres while the smallest, Waitakere Dam catchment, measured 8 square kilometres (Figure 2.1). The Waikato catchment is composed of a vast network of tributaries while the Waitakere catchment has only a few.

### 2.2.2 Drinking Water supply system

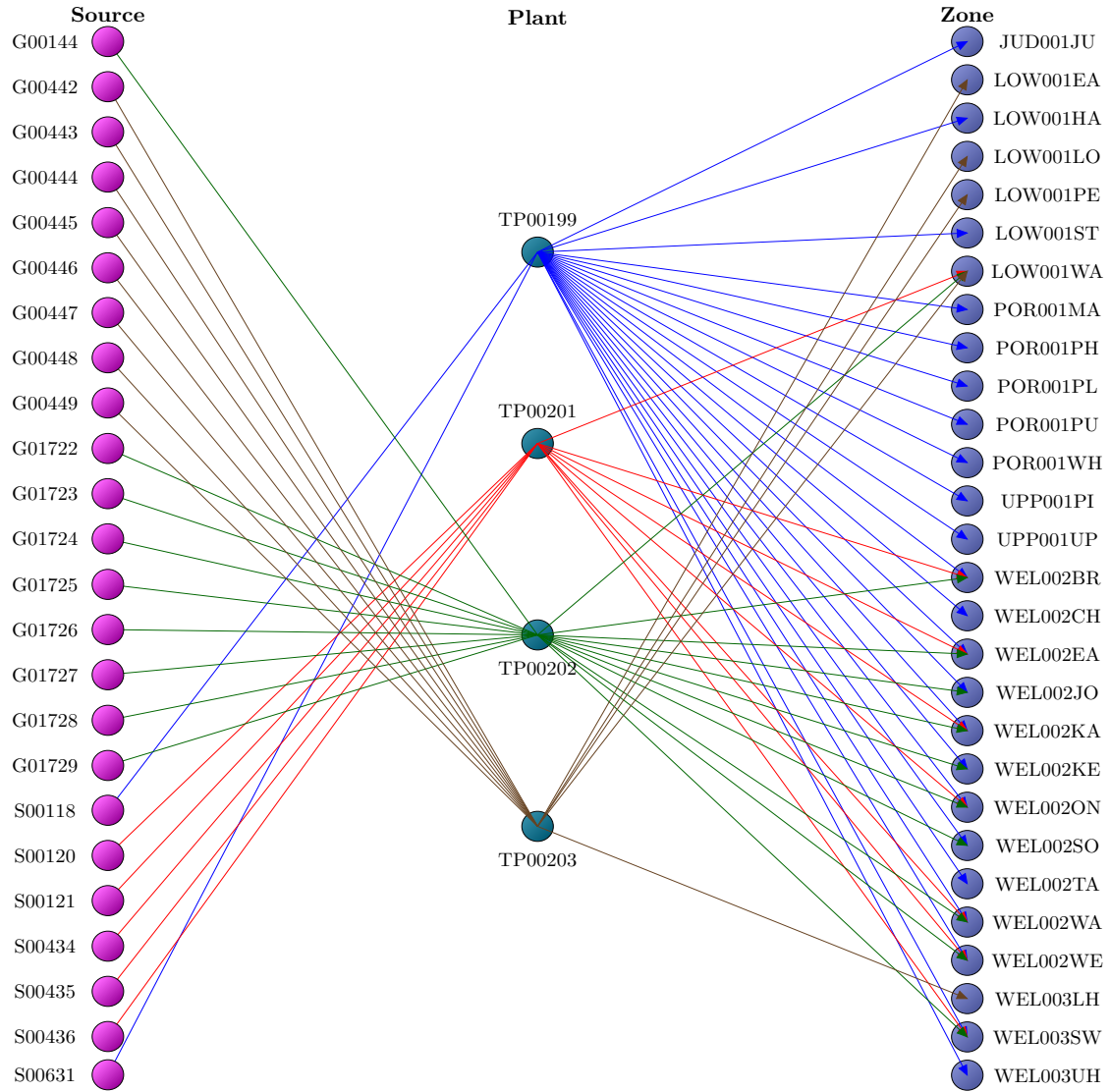
Like many other countries worldwide, the quality of drinking water in New Zealand is regulated by legislation. The New Zealand Ministry of Health (MoH) has the legislative mandate to regulate the quality of drinking water supplied to communities within New Zealand. Regulatory tools used by MoH include the Health (Drinking-water) Amendment

Act of 2007 (New Zealand Ministry of Health, 2007), the drinking water standards for New Zealand (DWSNZ) (New Zealand Ministry of Health, 2008) and the Register of Community Drinking-water Supplies in New Zealand (New Zealand Ministry of Health, 2011). It is a legal requirement for supplies servicing 25 or more people for 60 or more days per year to be registered with the New Zealand Government. In April 2012 there were 2258 registered supplies and 2329 distribution zones, according to Institute of Environmental Science and Research Limited (ESR) Limited, a state-owned Crown Research Institute responsible for providing scientific advice to both local and central government authorities in New Zealand. The registered supplies distribute water to about 91 % of the population. Nine percent of the population consumes self-supplied drinking water i.e. from unregistered sources, such as roof or borehole water, that are not monitored by MoH.

Water intended for public consumption is distributed to consumers by various water treatment plants (also known as ‘supplies’) through designated water supply zones. Generally, a drinking water supply network can be divided into three main parts: the source, treatment plant and distribution zone. In a simple standard network, water is abstracted at the source, treated at a treatment plant and delivered to the public via a distribution zone. However, in large communities complex water supply networks do exist which involve multiple sources, treatment plants and distribution zones. Figure 2.2 shows a schematic diagram of the drinking water supply network in the Wellington area based on the 2011 Register of Community Drinking Water Supplies in New Zealand. In this network, 24 sources supply raw water to four treatment plants that subsequently supply treated water to populations in Wellington City and surrounding areas through 28 distribution zones. In large supply networks the three different parts are physically in different locations while in very small networks all the three parts may be in the same location. An example in which all three parts of a network are located on the same premises would be a household or school with a roof water supply. Once the water is harvested from a roof (source) it is stored in a ground tank then treated using an inline filter or ultra violet (UV) (treatment plant) before being distributed to taps (distribution).



**Figure 2.1:** The Waikare and Waikato public drinking water catchments; the North Island map shows the location of the two catchments in New Zealand. The lines inside the catchment maps represent waterways.



**Figure 2.2:** Schematic representation of the drinking water distribution network in the Wellington area, based on the 2011 Register of Community Drinking Water Supplies in New Zealand. The circles represent water abstraction sources, treatment plants and distribution zones. The lines, colour coded according to treatment plants, represent the supply lines from water sources through treatment plants to distribution zones, from left to right.



## 2.3 Drinking water treatment processes

In order to reduce the concentration of pathogens in water intended for public consumption, conventional water treatment involves application of a combination of processes to raw water. Conventional drinking water treatment processes are not designed to sterilise water but to make it microbiologically and chemically safe for human consumption. The main processes include coagulation, flocculation, clarification, sedimentation, filtration and disinfection (Betancourt and Rose, 2004; United States Environmental Protection Agency, 2004).

During coagulation, a positively-charged coagulant is added to raw water in order to coagulate negatively-charged contaminants. Common coagulants include aluminium salts, iron salts and organic polymers. Coagulation is followed by flocculation which further agglomerates the coagulated particles into larger precipitates, also known as flocs. Flocculation achieves floc formation by gentle mixing and accelerating the rate of particle collision (Edzwald, 1993; Gao et al., 2002; Matilainen et al., 2010). Clarification, a process in which solids are separated from liquid (Volk et al., 2000), follows flocculation and flocs are removed by either sedimentation or skimming. Dense particles settle out of the water to the bottom of the treatment tank while light particles float to the surface. In order to enhance particle floatation fine air bubbles may be blown through the water in a process known as dissolved air floatation (DAF). The particles attach to the bubbles and hence float (Edzwald, 1995, 2007, 2010).

Physical removal of turbidity and microorganisms from water is ultimately accomplished by filtration. Particles that were not removed by sedimentation and floatation, such as clay or silt, are removed by filtration. The removal of particles in suspension occurs by straining through pores in a filter bed; by adsorption of the particles to the filter grains; by sedimentation of particles while in the media pores; and by coagulation while traveling through the pores. Commonly used filter types include dual-media filters composed of anthracite (a hard form of coal that contains relatively pure carbon) overlaying layers of sand (Betancourt and Rose, 2004). Filtration is considered a critical barrier for removal of protozoan (oo)cysts before water enters the distribution system (Cornwell et al., 2003). The majority of suspended particles are trapped in the top portion of the filter media. The trapped particles are dislodged by backwashing to keep the filter media clean and avoid clogging. Backwash water is usually returned to the start of the water treatment process, however, this practice has declined in recent years due to concerns about recycling of microorganisms (e.g. *Giardia* and *Cryptosporidium*), heavy metals or disinfection by-products (Curko et al., 2013; Reissmann and Uhl, 2006).

During disinfection organisms are inactivated or killed as a result of the use of physical or chemical disinfectants. Physical disinfection methods include boiling, UV irradiation, the use of electric discharges, cavitation, ultrasonic treatment, ultrafiltration, magnetic treatment and reverse osmosis (Biryukov et al., 2005; World Health Organization, 2004). Chemical disinfection, which is more commonly used than physical disinfection, includes chlorination, chlorination, and ozonification (OMOE, 2006). The efficacy of water disinfection is affected by many factors that include disinfectant concentration, contact time, temperature and pH. For instance, chlorine reacts with water to produce hypochlorous acid (HOCl) and hydrochloric acid (HCl); depending on the pH, hypochlorous acid may further break down into prohypochlorite ( $\text{OCl}^-$ ) and hydrogen ( $\text{H}^+$ ) ions. Hypochlorous acid is the main biocidal ingredient of a chlorine solution (Fair and Morris, 1949). Hypochlorous acid and prohypochlorite ions are often referred to as *free available chloride*. Further, hypochlorous acid reacts with many nitrogenous compounds that naturally occur in water. An important one among these compounds is ammonia, which reacts with hypochlorous acid to produce chloramines (monochloramine, dichloramine and trichloramine) in a pH dependent reaction. Chloramines are often referred to as *combined available chlorine* (ibid.). Increased pH results in decreased concentration of hypochlorous acid and increased concentration of prohypochlorite hence decreasing the disinfection effect. However, the best disinfection effect is obtained at pH 6–7 (Ward et al., 1984).

In general, water with a high debris content or high turbidity requires higher concentrations of free available chloride for effective disinfection compared to water with low debris content (World Health Organization, 2004). A potential side-effect of chemical disinfection is the existence of chemical residuals in drinking water, thus, monitoring of these products is recommended (New Zealand Ministry of Health, 2008; World Health Organization, 2004). Generally, protozoa are more resistant to disinfection than bacteria. *Cryptosporidium* is regarded as one of the most resistant organisms in water; previous studies have reported failure to achieve inactivation even after 18 hours of contact time with chlorine at very high levels and no inactivation has been observed with chloramines (Gyurek et al., 1997; Korich et al., 1990). This makes *Cryptosporidium* a benchmark organism for determining effective disinfection for protozoa (Betancourt and Rose, 2004; New Zealand Ministry of Health, 2008).

## 2.4 Common methods for detecting indicator organisms in drinking water

Currently it is not practical to monitor all known and potential human pathogens in drinking water due, in part, to the fact that many pathogens are difficult and costly to isolate in the laboratory. In addition, some pathogens exist in very low concentrations in water such that they are missed by many existing tests (Field and Samadpour, 2007). The solution to this problem has been the use of indicator organisms to show potential human and animal

excrement contamination (Harwood et al., 2014). Ideally, different indicator organisms are used for detecting the presence of different types of pathogens. Among the common indicator bacteria are the coliform group (total coliforms, faecal coliforms and *E. coli*), *Streptococci* (New Zealand Ministry of Health, 2008; World Health Organization, 2011) and spore formers such as *Clostridium perfringens*. An ideal indicator organism (Hoadley and Dutka, 1977; World Health Organization, 2011) should have the following qualities, although very few (if any) have all these qualities:

- Transmissible through water.
- Present in source waters.
- Able to survive as long as, or longer, than enteric pathogens.
- Persistent in the environment.
- Removable or inactivated by treatment processes.
- Unable to multiply in water.
- Should be present in faeces.
- Detectable using an easy and cheap test.
- Assessed for risk of exposure using available quantitative microbiological risk assessment (QMRA) data.

The specificity of indicator organisms to detect pathogens in water has generated debate over many years. For example, Borrego et al. (1987) suggested that coliphages were better indicators of faecal pollution than the conventional indicator systems used at the time. Payment et al. (1993) suggested that both somatic coliphages and *Clostridium perfringens* could be used as indicators for human enteric viruses and parasites (*Cryptosporidium* and *Giardia*) in treated drinking water. Conversely, Harwood et al. (2005) reported that no single indicator organism was correlated with pathogens in reclaimed water intended for non-potable use such as irrigation, cooling and industrial processing. Harwood and co-workers used total and faecal coliforms, enterococci, *Clostridium perfringens* and F-specific coliphages as indicators for pathogens that included enteric viruses and the protozoan parasites: *Cryptosporidium* and *Giardia*. Ferguson et al. (2012) reported that for the prediction of total bacterial pathogens using quantitative real-time polymerase chain reaction (qPCR)-based *E. coli* assay was the best, with monthly averages of culturable *E. coli* being better than daily measurements. Although F+RNA coliphages were found to predict bacterial pathogens well they predicted rotavirus poorly. Recently, Hewitt et al. (2013) proposed the use of human adenovirus (HAdV) and human polyomavirus (HPyV) as indicators of human faecal contamination. Hewitt and co-workers found that overall HAdV and HPyV correlated well with *Norovirus*, however, in wastewater impacted estuarine waters the two indicators tended to underestimate the concentrations of *Norovirus*.

Indicator organisms in water (drinking water, wastewater and recreational water) can be identified by several laboratory methods which may be categorised into three (Ashbolt et al., 2001): organism-isolation, immunology and gene-sequencing based methods.

### 2.4.1 Organism isolation-based methods

**Membrane filter technique (MFT):** This method consists of filtering a 10–100 mL water sample through a sterile filter with a 0.2 or 0.45 µm-pore size to trap organisms. The filter is incubated on a selective medium such as membrane filter agar medium (MI agar), in conditions favourable for indicator organism growth. Visible characteristic colonies are enumerated as *colony-forming units* or CFU per 100 mL (Ashbolt et al., 2001; Rompre et al., 2002). For example, on MI agar total coliforms form blue-white fluorescent colonies on exposure to long-wave ultraviolet light (366 nm) while *E. coli* form blue-green colonies (United States Environmental Protection Agency, 2002). Coliform bacteria form red colonies with a metallic sheen on an Endo-type medium containing lactose (Myers et al., 2007). These methods require a minimum of 24 hours of incubation before the colonies can be enumerated.

**Multiple-tube fermentation (MTF):** This method is also known as most probable number (MPN) and is used for enumerating total coliforms and *E. coli*. It has three stages (presumptive, confirmed and completed) and involves splitting a water sample into a series of dilutions. Each dilution is inoculated into a tube containing culture medium (World Health Organization, 1997, p. 60-62). During the first stage, the sample is inoculated into tubes containing lauryl tryptose broth. A positive presumptive test is obtained when gas is produced after 48 hours of incubation at 35 °C. In the second stage, presumptively positive tubes are used to inoculate tubes containing brilliant green lactose bile broth and incubated at 35 °C for a further 48 h. Any tube in which gas is produced during the incubation period is considered a confirmed positive test for coliforms. In the third stage, *E. coli* broth is inoculated with cultures from tubes that retained a confirmed positive test. Again, production of gas means a positive test, this time for *E. coli*. MPN is calculated by combining positive results in the second and third stages, using an approximation method:

$$\text{MPN}/100 \text{ mL} = \frac{100P}{\sqrt{V_n V_a}} \quad (2.1)$$

where  $P$  is the total number of positive results in the second or third stage;  $V_n$  is the combined volume of sample in the first stage tubes that produced negative results in the second or third stage;  $V_a$  is the combined volume of sample in all first stage tubes (Leboffe and Pierce, 2011). An example of how to calculate MPN using Equation 2.1 is provided in the Box 1 on page 24. Other methods for calculating MPN have published, for instance an exact method by McBride et al. (2003).

**Box 1:** Calculation of the most probable number (MPN)

Suppose a laboratory analysis is conducted on a water sample and the results are as presented in the Table below, the *E. coli* MPN for a *confirmed* test result can be obtained as follows:

Description	Dilution 1 (10 <sup>0</sup> )	Dilution 2 10 <sup>-1</sup>	Dilution 3 10 <sup>-2</sup>	Total (1+2+3)
Volume of original sample added to tube in stage 1	1.0 mL	0.1 mL	0.01 mL	
Number of tubes for each dilution	6	6	6	
Positive results in stage 2	6	4	3	13
Negative results in stage 2	0	2	3	
Total volume of original sample in all stage 1 tubes that produced negative results in stage 2	0.0 mL	0.4 mL	0.03 mL	0.43 mL
Volume of original sample in all stage 1 tubes inoculated	6.0 mL	0.6 mL	0.06 mL	6.66 mL

where stage 1 is the inoculation of the sample into tubes containing lauryl tryptose broth and stage 2 is the inoculation of tubes containing brilliant green lactose bile broth with inoculum from stage 1 positive tubes. MPN is then calculated as follows:

$$\begin{aligned}
 \text{MPN}/100 \text{ mL} &= \frac{100P}{\sqrt{V_n V_a}} \\
 &= \frac{100 \times 13}{\sqrt{0.43 \times 6.66}} \\
 &= 768
 \end{aligned}$$

### 2.4.2 Immunological methods

**Enzyme-linked immunosorbent assay (ELISA):** This is a biochemical assay based on the immunology concept of an antigen binding to its specific antibody. The basic principle is derived from the radioimmunoassay concept pioneered by Yalow and Berson (1960). Typically, an antigen (or antibody) in a given sample is immobilised to a solid surface (e.g. 96-well microtitre plate) and then complexed with a primary antibody. The complex thus formed is detected using a secondary antibody that is linked to an enzyme such as alkaline phosphatase or glucose oxidase. The activity of the conjugated enzyme in presence of a chromogenic substrate yields a measureable product, indicating the presence of antigen (Gan and Patel, 2013). In conventional ELISA the colour intensity of the solution often requires measuring using a plate reader. However, recently a technique that causes ELISA colour

changes to be readily visible to the naked eye was developed by De La Rica and Stevens (2012). Antigens that can be detected using ELISA include a wide variety of molecules such as proteins, peptides and hormones (Gan and Patel, 2013). Previous studies have reported the application of ELISA-based techniques for detecting microbes in drinking water: Hübner et al. (1992) reported a method for detecting members of the Family *Enterobacteriaceae* that produced results within 24 hours and Thiruppathiraja et al. (2011) developed an electrochemical immunosensor, based on ELISA, for detecting *Cryptosporidium parvum*.

### 2.4.3 Gene sequence-based methods

**Fluorescence *in situ* hybridisation (FISH):** This technique is based on principles pioneered by Gall and Pardue (1969) and involves the use of gene probes to target a specific portion of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) (Langer-Safer et al., 1982). The probe is a labeled oligonucleotide (i.e. DNA or RNA strand) complementary to the target, e.g. a target could be a region of the 16S or 23S ribosomal ribonucleic acid (rRNA) conserved regions (Manz et al., 1993; Speicher and Carter, 2005). The probe hybridises (binds) to the target sequence at elevated temperature. Incubation temperature and addition of chemicals can influence the stringency of the match between the gene probe and the target sequence. Examples of application of FISH-based techniques to enumerate bacteria in water intended for human consumption or recreational use include those enumerating *E. coli* (Baudart and Lebaron, 2010; Garcia Armisen and Servais, 2004), *Enterobacterium* (Baudart et al., 2002), *Mycobacterium* (Lehtola et al., 2006) and multiple bacterial species (Manz et al., 1993).

**Polymerase chain reaction (PCR):** This technique was developed in 1983 and is based on the natural processes a cell uses to replicate a new DNA strand. PCR makes numerous copies of a specific segment of DNA quickly and accurately (Mullis and Faloona, 1987). The process consists of five main components (Brunstein, 2013):

1. An aqueous buffer providing conditions suitable for the DNA polymerase to function, including  $Mg^{2+}$  ions, and a pH buffering agent (sometimes).
2. The basic building blocks of DNA, deoxyribonucleotide triphosphate (dNTP), used by the polymerase to form new DNA strands. dNTPs include the four nucleotides: deoxyadenosine triphosphate (dATP), deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP) and deoxythymidine triphosphate (dTTP). The chemical structures of the four dNTPs are shown in Figure 2.3.
3. *Taq*, a thermostable DNA polymerase.
4. Numerous copies of a pair of oligonucleotide primers. These are short, single-strand, DNA sequences that are complementary to the two sections flanking the target sequence to be amplified; one primer for either strand. Each of the primer's 3' end

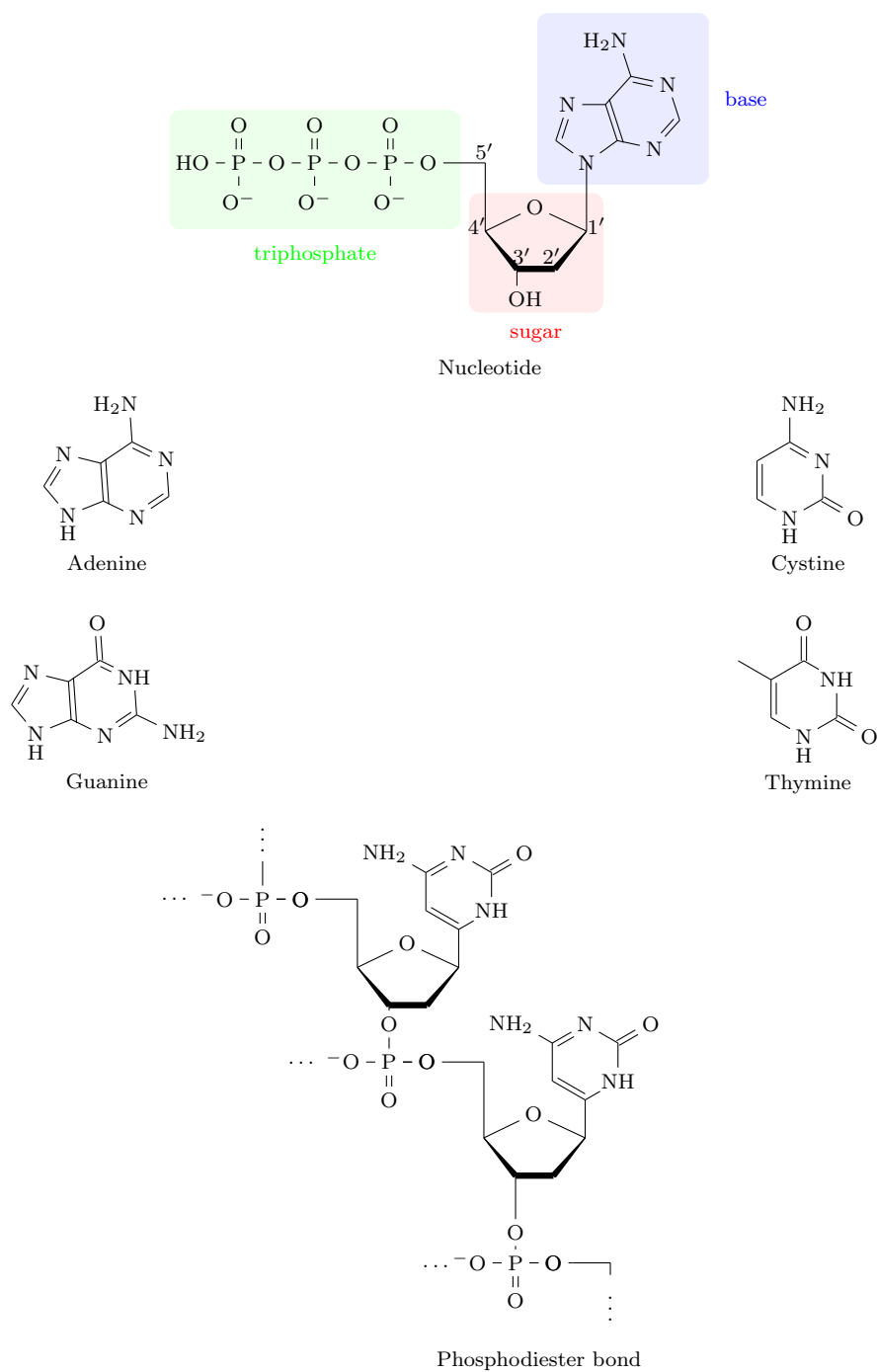
points towards the location of the opposite primer when annealed to their respective complementary sequences (Figure 2.4).

5. DNA fragments containing the target sequence.

PCR is a three-step process (Figure 2.4) that is repeated many times. During the first step the two strands of DNA are denatured (separated) by heating the sample to about 95 °C. In the second step the primers anneal to the templates after the temperature is reduced to about 55 °C. In the third step polymerase synthesizes DNA onto the ends of the annealed primers after the temperature is raised to about 72 °C.

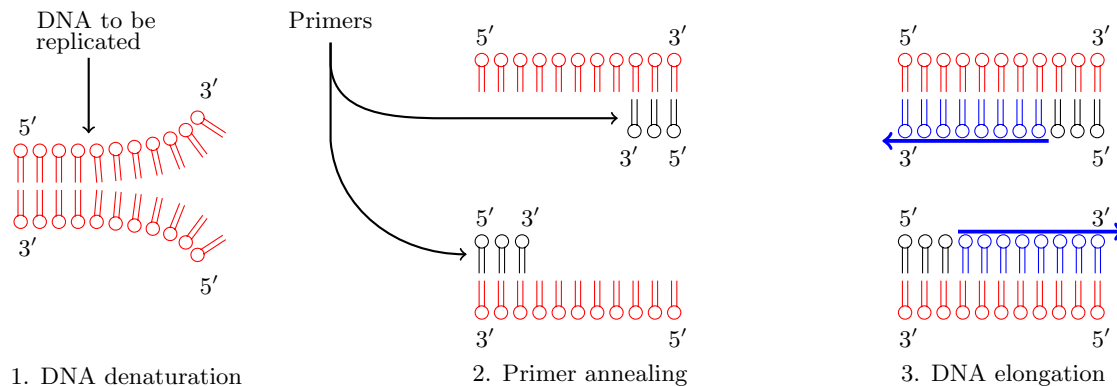
Bej et al. (1990) have demonstrated that PCR can be successfully used as an alternative to culture methods in detecting indicator pathogens in water with comparable specificity and sensitivity. In addition, PCR has the advantage of producing results within a shorter period of time compared to culture methods. Locas et al. (2008) used PCR-based methods to detect bacterial indicators as well as analyse the virological quality of water in Canadian municipal wells. qPCR is both quicker and more quantitative than culture methods and/or ordinary PCR. Reischer et al. (2007) developed a qPCR assay specific for detecting human faecal contamination in spring water from an alpine catchment in Eastern Austria. The assay was designed to detect human-specific markers (BacH) on the 16S rRNA gene from bacteria of the phylum *Bacteroidetes*. A PCR-ELISA technique developed by Kuo et al. (2010) was used to detect coliforms and produce results within four hours.

**Loop-mediated isothermal amplification (LAMP):** This technique was developed by Notomi et al. (2000) and amplifies DNA using *Bst* DNA polymerase under isothermal conditions and a set of four specifically designed primers that target six different parts of the template DNA sequence. The technique uses the fragment of the *Bst* DNA polymerase from *Geobacillus stearothermophilus* which has 5' → 3' polymerase activity but not 5' → 3' exonuclease activity (Niessen et al., 2013). In recent years, LAMP of DNA has been gaining popularity for use in examining the microbial safety of food and drinking water (Gallas-Lindemann et al., 2013; Karanis et al., 2007; Niessen et al., 2013; Plutzer et al., 2010; Wang et al., 2012a). Reasons for the technique's gain in popularity include its relatively shorter reaction time, cost effectiveness, high sensitivity, high specificity and ease of application (Niessen et al., 2013; Notomi et al., 2000). Karanis et al. (2007) developed a LAMP procedure for detection of *Cryptosporidium* oocysts in faecal and water samples by targeting the 60-kDa glycoprotein (gp60) gene. The study reported that the LAMP technique performed better than the PCR technique in detecting the oocysts. Plutzer et al. (2010) demonstrated that a combination of LAMP and ARAD<sup>®</sup> microfibre filtration can be used to continuously monitor *Cryptosporidium* and *Giardia* in drinking water supply systems. Gallas-Lindemann et al. (2013) used LAMP to detect *Toxoplasma gondii* in different water sources of the Lower Rhine area, Germany.



**Figure 2.3:** Schematic diagrams showing the three major parts of a nucleotide, the four bases and a phosphodiester bond of two adjacent nucleotides.





**Figure 2.4:** Schematic representation of the polymerase chain reaction process.

#### 2.4.4 Microbial compliance criteria for New Zealand

The use of *E. coli* as an indicator organism for the contamination of drinking water by faecal material is recommended by DWSNZ 2008. Other coliforms such as total and thermotolerant coliforms can be used but are less preferred. For a drinking water supply to be compliant with the bacteria safety standard, no *E. coli* should be detected in its treated water. *Cryptosporidium* is the organism used for assessing compliance with the protozoan safety standards. This is based on the fact that *Cryptosporidium* is the most difficult protozoan parasite to remove or inactivate in water hence its removal indicates that other protozoa have also been removed (New Zealand Ministry of Health, 2008).

The compliance requirements for protozoa under the DWSNZ 2008 are different from those for bacteria in that the water supplier is not specifically required to monitor the protozoan organisms directly. Instead, compliance is based on the ability of the treatment plant to remove protozoa, particularly *Cryptosporidium*. This requires an initial knowledge of the concentrations of *Cryptosporidium* in the source water as well as the efficiency of the treatment plant processes at removing or inactivating *Cryptosporidium*. The capacity of a treatment process to reduce the number of infectious *Cryptosporidium* oocysts in water is specified by the number of log credits it is assigned. The greater the number of log credits assigned to a treatment process, the larger the percentage of oocysts the process is able to remove or inactivate. The DWSNZ 2008 specify the number of log credits each treatment process can earn. Treatment plants often have more than one treatment process that can remove or inactivate *Cryptosporidium*. The overall effectiveness of the treatment plant, i.e. the total contribution made by all treatment processes, is calculated by adding the log credits of the individual processes together. The log credit for a treatment process is

related to the percentage of the protozoa the process can remove, it is determined by the following expression ([ibid.](#)):

$$\log \text{ credit} = \log_{10}(1/(1 - (\text{percentage removal}/100))) \quad (2.2)$$

If a treatment plant achieves, for instance 2-log credits, it means 99 % of the oocysts have been removed; 99.9 % oocyst removal for 3-log credits and 99.99 % oocyst removal for 4-log credits.

## 2.5 Microbial source tracking

Microbial source tracking (MST) is the process of identifying the origin of microbial contamination in water. The rationale behind MST is that microbiological, genotypic and/or phenotypic markers can be used to identify the original animal source of faecal contamination. Identification of the source is important because it helps in designing better strategies for preventing contamination. Methods used for conducting MST can be classified into three groups, depending on whether they require microbes to be cultured in the laboratory or not and whether microbial identification requires reference to a microbial reference database or not (Field and Samadpour, [2007](#)):

- Culture-dependent, microbial reference database-dependent methods — the microbes are grown in the laboratory and various tests performed on them and matched to a microbial reference database. The tests include phenotypic expression, genotyping, phage typing and bacterial ratios in water (Ames et al., [2013](#); Geldreich and Kenner, [1969](#); Meays et al., [2004](#); Scott et al., [2002](#)).
- Culture-independent, microbial reference database-dependent methods — the main methods in this category are those based on PCR and metagenomics, which involve microbial genome detection from environmental samples without culturing the microbes in the laboratory. For metagenomics, the microbial genomes are sequenced and matched to a microbial reference database (Field et al., [2003](#); Kildare et al., [2007](#); Lu et al., [2007](#)).
- Culture-independent, microbial reference database-independent methods — these methods involve detection of specific chemicals or biomarkers which can be traced to humans or other animals. For example, the presence of chemicals such as caffeine and laundry brighteners can indicate contamination of human origin. Biomarkers such as faecal sterols and stanols can be used to discriminate between human and non-human contamination (Leeming et al., [1996](#); Scott et al., [2002](#); Shah et al., [2007](#)).

Increasingly, culture-independent methods are being applied in MST processes within aquatic environments as alternatives to culture-dependent methods. For this purpose, more and more host-specific genomic markers are being identified, for instance, chicken (Lu et al., 2007), cattle (Shanks et al., 2006), duck (Devane et al., 2007) and human (Shanks et al., 2007) genomic markers have been proposed. Krentz et al. (2013) demonstrated that markers specific to cattle, chickens, geese, humans, pigs and seagulls, reported in previous studies, can be used to successfully identify sources of contamination in water. The target markers for the hosts investigated by Krentz and co-workers were located in host-specific *Bacteroides* 16S rRNA genes except for geese and seagulls. In these two hosts, the markers were located in the *Prevotella* 16S rRNA and *Catellibacillus marimammalius* 16S rRNA genes, respectively.

Biomarkers are chemical equivalents of indicator organisms, as they do not directly indicate presence of pathogens but suggest potential faecal contamination from a specific group of animals or humans. For instance, the presence of laundry chemical such as brighteners suggest a possible contamination of human origin (Moriarty and Gilpin, 2009). Digestion by-products such as faecal sterols can provide useful source signatures in both animals and humans. Chemically, sterols are a group of steroids belonging to 3-hydroxysteroids, that include C27-C30 crystalline alcohols. A combination of factors including diet, gastrointestinal microflora and the body's ability to synthesize its own sterols influences their composition in faeces (Jardé et al., 2007; Moriarty and Gilpin, 2009). Derrien et al. (2012) used stanols (saturated sterols) to discriminate human from bovine and porcine fecal contamination sources in surface water.

## 2.6 Indicator organism detection in recreational water

Most of the principles and methods for detecting contamination and for performing MST in drinking water are also applicable for recreational water. In addition, the main sources of microbial contamination in the two types of water are similar i.e. human and animal faeces (World Health Organization, 2003, 2011). Therefore, the purpose of this Section is not to repeat the discussion of the principles and methods that are common for detecting faecal contamination and for performing MST in the two different types of water (discussed in the previous two Sections) but to highlight differences where they occur. The differences in the methods partly arise from the fact that while drinking water sources are largely freshwater-based, recreational water include both freshwater and saltwater. Similar faecal contamination indicators can be used in both freshwater and saltwater intended for recreational use, e.g. *E. coli* and enterococci (World Health Organization, 2003). However, previous studies have shown that these two types of organisms survive at different rates in freshwater and saltwater (Anderson et al., 1979; Anderson et al., 2005). These studies reported that survival of *E. coli* was inversely proportional to levels of salinity

while enterococci survived high salinity levels better. In general, *E. coli* is recommended as a freshwater faecal contamination indicator (New Zealand Ministry for the Environment, 2003; New Zealand Ministry of Health, 2008) while enterococci is recommended in saltwater (New Zealand Ministry for the Environment, 2003; United States Environmental Protection Agency, 2000).

## 2.7 Pathogens in drinking water — New Zealand

A systematic search for published peer-reviewed research articles related to pathogens associated with drinking water in New Zealand was conducted in December 2013. The search was conducted using the search engines Scopus and Web of Knowledge, available through the Massey University library. Pathogens recognised by WHO as waterborne (Table 2.1) were used as keywords for the construction of a search term. After several iterations with different combinations of keywords redundant keywords were removed resulting in the following final search term being used: ((campylobact\* OR cryptosp\* OR ‘E. coli’ OR ‘escherichia’ OR enterovir\* OR giard\* OR norovir\* OR salmonell\*) AND (new zealand) AND (‘drinking water’ OR drinking-water)). The same search term was used in both search engines. The Scopus engine retrieved 40 articles while Web of Knowledge returned 67; these were combined and duplicates removed resulting in 74 unique retrievals. Each retrieved article was scanned to determine if it was relevant to the current literature review or not by examining the title and abstract. In total 24 retrieved articles were considered relevant and were included in the current review. The references of the relevant articles were scanned in order to identify more relevant articles not retrieved by the electronic search. A retrieved article was regarded as relevant if it reported a research study regarding waterborne disease, waterborne pathogen(s) or indicator organism(s) in relation to drinking water in New Zealand.

The purpose of performing this review was to summarise findings of published peer reviewed research related to waterborne pathogens and drinking water in New Zealand, hence identify research trends and knowledge gaps. The relevant articles were classified into two groups; the first category was composed of articles that focused primarily on human disease cases and the second category on waterborne organisms. Disease case studies were further subdivided, based on their study design, into outbreak investigations, case-control and retrospective cohort studies. All of the organism-focused studies used a cross-sectional study design.

### Outbreak investigations

Six studies described disease outbreak investigations; three involving campylobacteriosis, two salmonellosis and one *Norovirus* infection. Campylobacteriosis (Bohmer, 1997; Briese-man, 1987; Stehr-Green et al., 1991) and *Norovirus* (Hewitt et al., 2007) studies involved

campgrounds and a ski resort, respectively. Outbreaks at the campgrounds were due to consumption of water from untreated or inadequately treated supplies while at the ski resort an unusual contamination event of the water supply had occurred. The two studies investigating salmonellosis outbreaks analysed outbreak surveillance data; one over a 4 year period, 1998–2001 (Thornley et al., 2002), and another for a ten year period, 2000–2009 (King et al., 2011). Thornley et al. (2002) investigated 137 outbreaks while King et al. (2011) investigated 204 outbreaks. The largest proportion of salmonellosis outbreaks were attributed to food sources followed by person-to-person contact, water consumption and animal contact. Of the disease case studies 53 % used disease notification data, indicating that this is a very important resource for researchers. None of the retrieved articles reported use of metagenomic techniques exposing a new research area to be explored as far as microbial water quality is concerned in New Zealand.

### **Case-control studies**

Of the studies that had a case-control study design, two investigated the risk of acquiring giardiasis (Hoque et al., 2002; Mitchell et al., 1993), another two campylobacteriosis (Eberhart-Phillips et al., 1997; Ikram et al., 1994) and one *E. coli* infection (Jaros et al., 2013). Evidence from the two giardiasis studies showed that consuming drinking water other than that from regulated city supplies (probably a proxy for consumption of inadequately treated water) resulted in an elevated risk of acquiring giardiasis. Other risk factors for acquiring giardiasis included exposure to human waste, swimming and travelling outside New Zealand. Ikram et al. (1994) reported that there was an elevated risk of acquiring campylobacteriosis, although statistically non-significant (marginally), associated with consumption of water other than from the city mains. Risk factors that were statistically significant for acquiring campylobacteriosis were mainly food-related, including consuming poultry at a friend's home, consumption of undercooked poultry or barbecued chicken. In a study by Eberhart-Phillips et al. (1997), raw or undercooked chicken and consumption of chicken in restaurants were found to be strongly associated with campylobacteriosis. The likelihood of acquiring campylobacteriosis also increased with recent overseas travel, roof water at household level, consumption of dairy products, contact with puppies and contact with calves. Consumption of baked or roasted chicken appeared to protect against acquiring campylobacteriosis. Jaros et al. (2013) found that travel to areas in New Zealand with interrupted or no main water supply, contact with recreational water together with contact with animal manure and presence of cattle in a meshblock (smallest geographical unit in New Zealand) were risk factors for shiga toxin-producing *E. coli* (STEC) O157:H7. Food sources were not associated with acquiring STEC infection.

### **Retrospective cohort studies**

Seven studies used a retrospective cohort study design. Five of these studies used New Zealand's national disease notification data for periods ranging from 2 to 50 years: 1996–

1998 (Duncanson et al., 2000), 1997–2006 (Britton et al., 2010b), 1996–2000 (Hoque et al., 2004), 1998–2002 (Khan et al., 2007) and 1952–2001 (Thornley et al., 2002). Cowie and Bell (2013) used disease notification data for the Waikato region (2004–2011) while Schousboe et al. (2013) used routine blood stream infection monitoring laboratory data belonging to the Canterbury District Health Board (2009–2011). The evidence from these studies suggest that consuming water from supplies that were ungraded, graded as unsatisfactory, not complying with DWSNZ, inadequately treated water or had a poor quality water source resulted in an increased risk of acquiring waterborne infection (Britton et al., 2010b; Duncanson et al., 2000; Khan et al., 2007). Such water supplies were likely to be those supplying small communities and probably located in rural areas. Other risk factors for gastrointestinal illness included being younger than 5 years, a history of travel outside of New Zealand, contact with persons with gastrointestinal symptoms and contact with farm animals. The evidence also showed that susceptibility to waterborne infections varied among ethnic groups, for example persons of Asian descent were more likely to acquire giardiasis than others (Hoque et al., 2004) while persons of European descent were more likely to acquire cryptosporidiosis (Cowie and Bell, 2013) or salmonellosis (Thornley et al., 2002). Cases of salmonellosis were reported to be more prevalent in late summer/early autumn (*ibid.*) while cryptosporidiosis was more prevalent in spring (Cowie and Bell, 2013). Britton et al. (2010b) reported that consumption of water from supplies with the best grade had an elevated risk for acquiring giardiasis. There is no immediate explanation for this seemingly counter-intuitive finding. Schousboe et al. (2013) reported that drinking water was not a factor in the increased incidence of *E. coli* bacteraemia after the 2011 Christchurch earthquake despite land liquefaction, widespread sewer system damage and possible drinking water source contamination.

### Cross-sectional studies (organism-focused)

Five studies (Close et al., 2008; Donnison et al., 2004; Savill et al., 2001; Simmons et al., 2001; Till et al., 2008) focused on indicator organisms. Indicator organisms were regularly detected in source waters, e.g. *E. coli* was detected in 99 % or more of surface water samples (Close et al., 2008; Donnison et al., 2004; Savill et al., 2001; Till et al., 2008) and 50 % of roof water samples (Simmons et al., 2001). The concentration of indicator organisms was measured as MPN of organism per 100 mL of water sample and a sample was declared positive if it had one or more MPN/100 mL. Pathogens were also detected i.e. *Campylobacter* was detected in 66 % of shallow wells, 60 % of surface water samples, 37 % of roof water samples and 29 % of reticulated water samples (Savill et al., 2001). The species of *Campylobacter* detected included *C. jejuni*, *C. coli* and *C. lari*.

Factors associated with the presence of *E. coli* in surface water sources included animal grazing activities in the catchment and increased river flow. *E. coli* concentrations were highest during the spring and summer months and lowest in autumn and/or winter (Donnison et

al., 2004). Various factors, including type of roof, type of storage tank, water turbidity and season, were investigated for their association with the presence of indicator organisms in roof harvested water. Indicator organisms investigated included heterotrophs, total coliform (TC), faecal coliform (FC) and *Enterococci* spp.. The presence of heterotrophs in roof water was associated with galvanised roofs and storage tanks; TC with water turbidity and galvanised roofs; FC with plastic storage tanks and *Enterococci* spp. with tiled roofs (Savill et al., 2001).

## 2.8 Metagenomics

Currently, only about 1 % of all existing microbial species have been cultured in the laboratory (Amann et al., 1995), posing a huge limitation on the use of culture-based methods in studying microbial community profiles and diversity. However, recent studies have reported techniques that greatly improve the number of cultured bacterial species (Ling et al., 2015; Nichols et al., 2010). In contrast, metagenomics does not require prior culturing of microbes and offers an opportunity to overcome some of the limitations of culture-based methods. In the recent past, studies have used metagenomics to describe microbial community metagenomes in a wide variety of ecosystems, for example, in human and animal gastrointestinal tracts (Ellis et al., 2013; Qin et al., 2010), freshwater (Ghai et al., 2011), drinking water distribution system biofilms (Schmeisser et al., 2003) and comparison of the phylogeny of microbial communities in diverse habitats (Von Mering et al., 2007). These studies demonstrate how diverse and complex the microbial ecology and function are in different habitats. Accumulating evidence on factors that influence the composition of microbial communities will inevitably lead to an improved ability to predict the presence of pathogens. Equipped with such knowledge, better methods of controlling pathogenic microbes in drinking water systems are likely to be developed.

Metagenomic studies have widely used sequencing of specific genes, particularly the 16S rRNA gene, to describe microbial profiles and phylogenetics while whole genome shotgun (WGS) sequencing has been used to describe microbial ecology and function. rRNA is used because it is essential for protein synthesis in all living organisms and comprises the predominant material within the ribosome, which is around 60 % rRNA and 40 % protein by weight. The rRNAs form two subunits, the large subunit (LSU) and small subunit (SSU). Both 16S rRNA in prokaryotes and 18S rRNA in most eukaryotes are contained within the SSU (Matheson, 1992).

The use of the 16S rRNA gene, also known as 16S ribosomal deoxyribonucleic acid (rDNA), for microbial species discrimination is based on the principle pioneered by Woese and Fox (1977). Among the characteristics that make the 16S rRNA gene optimal for use in microbial profiling studies are that firstly, the gene is ubiquitous among prokaryotic life. Secondly,



its size and high degree of functional conservation result in highly predictable mutation rates throughout prokaryotic evolution. Thirdly, the 16S rRNA gene includes both conserved and hypervariable regions. The conserved regions flank hypervariable regions and can be used for designing universal PCR amplification primers across taxa. Bacterial 16S rRNA genes generally contain nine hypervariable regions (V1-V9) that demonstrate considerable sequence diversity among different bacterial species and can be effectively used to distinguish between taxa (Kolbert and Persing, 1999; Pereira et al., 2010).

It is worth noting that hypervariable region sequences exhibit some degree of heterogeneity and that no single region can discriminate all bacteria (Chakravorty et al., 2007). This means that it is possible to misclassify bacteria belonging to the same species as belonging to different species and *vice versa*. Discriminating between *Campylobacter* species is a good example in which even sequencing multiple hypervariable regions may not be enough to correctly identify the different species (Gorkiewicz et al., 2003; Gunther et al., 2011). This clearly demonstrates that the choice of hypervariable region(s) to be sequenced depends on the aim of the particular investigation at hand. For instance, Mizrahi-Man et al. (2013) recommended V3 or V4 for profiling bacteria in general while Guo et al. (2013) recommended V1 and V2 for bacteria classification in wastewater.

Alternative schemes of species identification have recently been proposed in order to overcome some of the challenges posed by the 16S rRNA gene method. For instance, Pereira et al. (2010) proposed the use of species identification by insertions/deletions (SPInDel), a method that involves the sequencing of multiple conserved and hypervariable regions. Gunther et al. (2011) proposed the use of the *gyrB* gene, instead of the 16S rRNA, for discrimination among *Campylobacter* species while Peeters and Willems (2011) reported that the *gyrB* gene had a higher discriminatory power than the 16S rRNA gene among *Flavobacterium* strains. The *gyrB* gene encodes the subunit B protein of DNA gyrase, an essential bacterial enzyme involved in the control of topological transitions of DNA (Reece and Maxwell, 1991). The gene that encodes heat-shock protein (HSP)60 (also known as *cpn60* or *groEL*) has also been used for profiling microbial communities. HSP is involved in the prevention of damage to proteins in response to high levels of heat as well as in importation of proteins into the mitochondria and macromolecular assembly (Cheng et al., 1989). Hill et al. (2006) designed universal *cpn60* PCR primers that improved representation of high G+C content organisms in microbial community sequence data. High G+C content organisms such as *Actinobacteria* are known to be under-represented in 16S rRNA gene microbial community studies.

Among the disadvantages of PCR amplicon sequencing, such as the 16S rRNA gene, is the problem of primer and amplification bias (Patin et al., 2013; Schloss and Westcott, 2011). This has been reported to affect microbial profiling using as templates that do not perfectly



match primers are inefficiently amplified hence under-represented (Lee et al., 2012; Sipos et al., 2007). Conversely, templates that perfectly match primers are preferentially amplified and over-represented. To mitigate this problem the use of degenerate primers or a mixture of non-degenerate primers has been recommended (Lee et al., 2012; Schloss and Westcott, 2011).

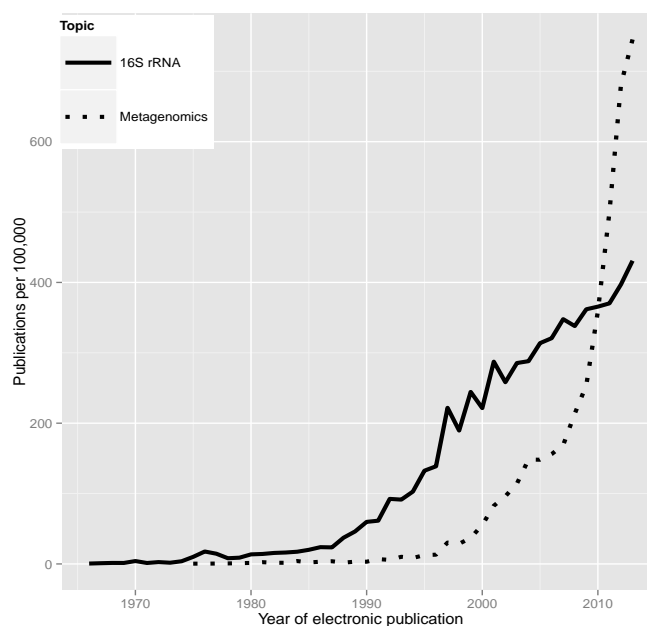
### 2.8.1 Metagenomics in drinking water

A systematic search for peer-reviewed research literature related to the use of metagenomics in drinking water was conducted using an approach similar to that described for waterborne pathogens and drinking water in New Zealand. The search term used was: ((metagenom\* OR metabiom\*) AND ('drinking water' OR drinking-water OR freshwater OR groundwater)). The Scopus engine retrieved 166 articles while Web of Knowledge retrieved 222. After combining the two sets of retrieved articles and removing duplicates, 11 were considered relevant to the current review. Relevant articles were those that conducted their investigations along the drinking water supply chain (from the water source to the tap) and applied metagenomic techniques for sample processing and analysis. Of the relevant articles, five used 454 sequencing technology (Delafont et al., 2013; Gomez-Alvarez et al., 2012; Kwon et al., 2011; Oh et al., 2011; Pinto et al., 2012), another five Illumina (Bai et al., 2013; Chao et al., 2013; Oh et al., 2011; Shi et al., 2013; Wang et al., 2012b) and one ABI (Schmeisser et al., 2003). All studies in the relevant articles used the 16S rRNA gene to describe microbial community profiles at various stages of the drinking water supply system while five studies additionally analysed functional genes.

### 2.8.2 Metagenomic research trends

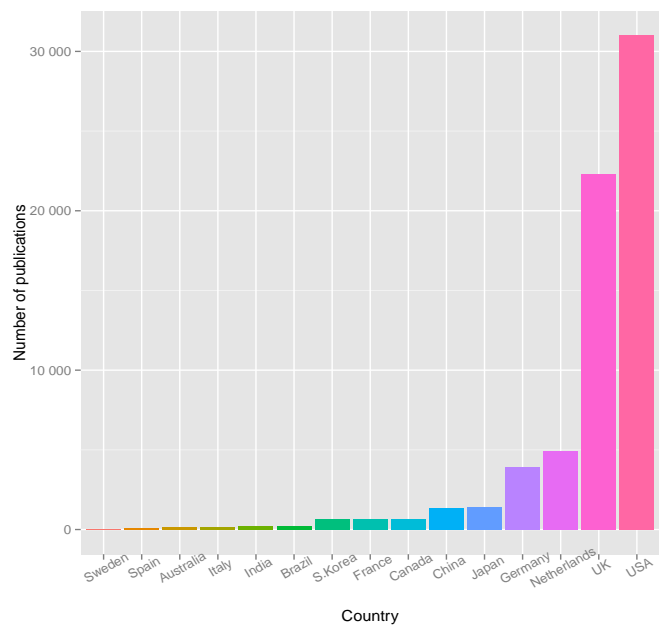
The advent of next-generation sequencing (NGS) technology, leading to substantial reduction in the cost (Wetterstrand, 2014) and time required to sequence samples (Mardis, 2011; Turner, 2011), has allowed metagenome sequencing to be applied more widely and more habitats being taxonomically profiled. Publication trends corroborate this fact as research employing 16S rRNA and metagenomics techniques is gaining popularity in recent years. An electronic PubMed<sup>®</sup> search for the available years (1950–2013) was conducted in January 2014 using the **R** package **RISmed** (Kovalchik, 2014). Two search terms, ('16S rrna' OR '16S rdna' OR (hypervariable AND region\*)) and ((high-throughput OR (next AND generation) OR shotgun) AND sequenc\* OR metagenom\* OR pyrosequenc\*), were constructed to retrieve research publications related to 16S rRNA and metagenomics, respectively. The two search terms were combined to obtain the number of publications per journal and country. Searches without any specified search term were conducted to obtain the total number of publications for each year, journal and country. The **R** code used for the systematic literature search is provided in Appendix A.1 on page 169.

A total of 41 214 16S rRNA and 31 713 metagenomic-related publications were retrieved. Figure 2.5 shows the number of 16S rRNA and metagenomic-related research articles per 100 000 publications for each year from 1950 to 2013. There was a steady increase in the proportion of publications related to 16S rRNA from the mid-1980s to 2013. The proportion of metagenomic-related publications increased steadily from the mid-1990s to late 2000s followed by a sharp increase to 2013. Figure 2.6 shows 15 countries with the highest number of 16S rRNA and metagenomic-related publications during the search period, totalling 67 767. Of these 45.7 % originated from USA and 33.9 % from the United Kingdom (UK). The leading 20 peer-reviewed scientific journals in publishing of 16S rRNA and metagenomic-related research over the search period are shown in Figure 2.7. Figure 2.7a displays the number of 16S rRNA and metagenomic-related publications retrieved per journal while in Figure 2.7b the percentage of the journal's publications comprising metagenomic-related articles is shown. For instance, the International Journal of Systematic and Evolutionary Microbiology published a total of 5800 16S rRNA and/or metagenomic-related articles which comprised about 90% of its total publications.

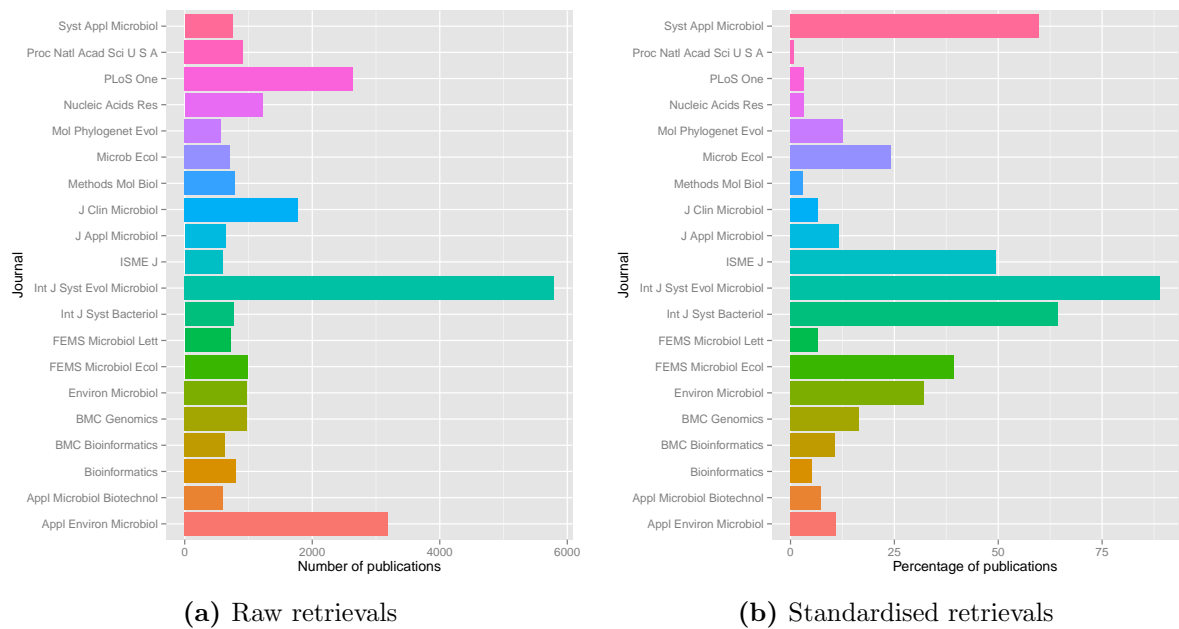


**Figure 2.5:** Line plot showing the number of articles on 16S rRNA and metagenomics published per 12-month period as a function of calendar year, 1950–2013.

The purpose of conducting a literature review on metagenomic methods was to gain an in-depth understanding of the principles on which these methods are based and how they have been applied in the drinking water industry. Metagenomic methods were adopted in the current research project to investigate the drinking water quality on campgrounds in New Zealand. Caution should be exercised when interpreting results of an automated electronic article retrieval such as the one used here through the the **R** package **RISmed**. This is because such retrievals often include articles not relevant to the search topic hence



**Figure 2.6:** Top fifteen countries with publications on 16S rRNA and metagenomics for the period 1950–2013



**Figure 2.7:** Top twenty peer-reviewed journals publishing articles on 16S rRNA and metagenomics publication, 1950–2013

tends to overstate the number of retrievals. For example, one of articles retrieved under the search term ((high-throughput OR (next AND generation) OR shotgun) AND sequenc\* OR metagenom\* OR pyrosequenc\*) was that of surgical management of injuries inflicted by a shotgun in human patients.

### 2.8.3 Microbial community profiles

Overall, *Proteobacteria* was frequently reported as the most abundant phylum in drinking water, with a median abundance of 60 % (range: 35–91 %). This is not surprising as *Proteobacteria* is the most abundant phylum among bacteria. Other bacterial phyla frequently recovered from drinking water systems include (from most to least): *Actinobacteria*, *Verrucomicrobia*, *Bacteroidetes*, *Planctomycetes*, *Cyanobacteria*, *Firmicutes*, *Acidobacteria* and *OD1* (Bai et al., 2013; Chao et al., 2013; Kwon et al., 2011). Among the *Proteobacteria* classes *Alphaproteobacteria* was the most common followed by *Betaproteobacteria*, *Gammaproteobacteria*, *Deltaproteobacteria*, *Epsilonproteobacteria* and *Zetaproteobacteria*. In contrast, the faecal microbiomes tend to be dominated by *Firmicutes*, *Bacteroidetes* and *Fusobacteria* as opposed to *Proteobacteria* (Hand et al., 2013; Oikonomou et al., 2013).

The evidence accumulated to date indicates that different microbial communities tend to favour colonising different parts of the water supply system. For instance, Kwon et al. (2011) reported that the proportion of *Betaproteobacteria* was higher than that of *Alphaproteobacteria* in raw water but the opposite was true in treated water. The most abundant *Proteobacteria* genera in source water were *Betaproteobacteria* followed by *Alphaproteobacteria*, *Gammaproteobacteria*, *Epsilonproteobacteria* and *Deltaproteobacteria* (Chao et al., 2013; Delafont et al., 2013; Kwon et al., 2011; Oh et al., 2011; Pinto et al., 2012). In water from treatment plants the order of abundance was *Alphaproteobacteria*, *Betaproteobacteria*, *Deltaproteobacteria* and *Gammaproteobacteria* (Bai et al., 2013; Kwon et al., 2011; Pinto et al., 2012) while in distribution networks it was *Betaproteobacteria*, *Alphaproteobacteria*, *Gammaproteobacteria* and *Deltaproteobacteria* (Gomez-Alvarez et al., 2012; Pinto et al., 2012; Schmeisser et al., 2003). While a great deal of research work has been published in this area, the number of studies available is too small to draw a definitive conclusion on the pattern of microbial abundance at different stages of water supply systems. However, the evidence show the potential usage of metagenomic techniques in the water industry, for example, the identification of groups of bacteria likely to survive the water treatment processes. Such information is crucial for developing improved water treatment methods.

### 2.8.4 Microbial community functional genes

Chao et al. (2013) reported that among the top level 1 functional genes (as classified by the SEED<sup>4</sup>) present in raw and treated water included those related to protein metabolism, carbohydrates, amino acids and amino acid derivatives. Genes related to glutathione syn-

<sup>4</sup>A software that uses a subsystem approach to organise functional gene categories into five hierarchical levels (Overbeek et al., 2005)

thesis largely increased after water treatment. Glutathione plays a key role in the protection mechanism of cells against the action of low pH, chlorine compounds, and oxidative and osmotic stresses (Masip et al., 2006). This indicates that organisms resistant to disinfection survive water treatment hence positive selection of the resistance genes. This implies that eventually more effort and resources will be required to achieve desired levels of water treatment and microbial drinking water safety.

The five most abundant functional genes in biofilms reported by Schmeisser et al. (2003) include those related to metabolism and catabolism, cell processes and structure, DNA/RNA-modifying enzymes, regulatory function and transport proteins. About a quarter were hypothetical proteins with unknown function. Genes associated with catabolism and metabolism of lipids, aromatic compounds, proteins, amino acids and polysaccharides were identified. This indicates that the microbes in biofilms are able to utilise a wide variety of carbon and energy sources. Other genes identified were those related to antibiotic resistance and metal detoxification, including resistance against tetracycline and  $\beta$ -lactam antibiotics. This implies that antibiotic resistance could be spread through inadequately treated water, emphasizing the need for strict regulatory processes.

Gomez-Alvarez et al. (2012) reported that *Mycobacterium* spp. was more abundant in water treated with chloramine compared to that treated with chlorine. In chloramine treated water, virulence factors involved in *Mycobacterium* intracellular parasitism were identified. Among these virulence factors were mammalian cell entry and phospholipid ABC transporter (*yrbE*) proteins that enable *Mycobacterium* invade host cells. Genes related to production of resistance against  $\beta$ -lactam antibiotics were also detected. Shi et al. (2013) investigated the effects of drinking water treatment through chlorination on microbial antibiotic resistance. *Proteobacteria* were found to be the main antibiotic resistant bacteria and chlorination significantly altered the microbial community profile. Bacteria surviving chlorination were likely to be resistant to chloramphenicol, trimethoprim and cephalothin. The most common antibiotic resistant gene was *sulI* followed by *tetA* and *tetG*. The evidence suggested that chlorination could concentrate various antibiotic resistant genes and also plasmids, insertion sequences and integrons involved in horizontal transfer of the antibiotic resistant genes.

## 2.9 Summary

Access to safe drinking water remains a challenge in many countries worldwide, particularly in developing countries. In developed countries, including New Zealand, the majority of the population has access to drinking water, the quality of which is regularly monitored by authorities. Despite the great amount of effort and resources directed towards making drinking water microbiologically safe, waterborne diseases continue to be reported among

consumers. Factors such as climatic change are likely to further compound the waterborne disease burden as they affect the availability and microbial burden of drinking water globally. In dry regions, projected to become drier, communities may resort to abstracting drinking water from more contaminated sources e.g. recycled sewerage water. While in wet regions, projected to become wetter, water sources are likely to become more polluted due to increased runoff and/or flooding. The net effect of climatic change can be interpreted as increased potential public health risk due to waterborne illness in both dry and wet regions. The cost of water treatment can be expected to increase in order to deal with the anticipated increase in water source contamination and the associated increase in public health risk.

Animal and human faecal matter remain important sources of waterborne pathogens that include bacteria, protozoa and viruses. Thus, strategies aimed at reducing water source pollution should be those intensifying the prevention of faeces from getting into waterways. For the pathogens already in waterways, characterising pathogenic organisms would help identify general features that could be exploited to develop better pathogen removal methods. For example, all the World Health Organization (WHO)-recognised waterborne bacterial pathogens taxonomically belong to one group, phylum *Proteobacteria*, the majority being from class *Gammaproteobacteria* (Table 2.1). Identifying general characteristics for this group of bacteria that could be exploited for developing target-specific removal strategies would greatly enhance the efficacy of the water treatment process. Thus an ideal test would be one that was able to indicate faecal contamination, identify the faecal source and identify specific pathogens. Such a test would provide a framework for improved estimation of the associated public health risk and development of enhanced control measures.

Infections associated with drinking water generally manifest themselves as gastrointestinal illness in humans. Among the most reported waterborne infections in developed countries, such as New Zealand, the USA and the United Kingdom, include those caused by *Campylobacter*, *Cryptosporidium*, *E. coli*, *Giardia* and *Salmonella*. In New Zealand the proportion of gastrointestinal infections attributed to food sources is greater than that attributed to drinking water. This is demonstrated by evidence provided by Muellner et al. (2013) which showed that control of contamination in the poultry supply chain resulted in a dramatic decrease in human campylobacteriosis cases. The proportion attributed to drinking water is likely to be due to consumption of untreated or insufficiently treated water. Such water is likely to come from supplies that are less compliant with drinking water standards for New Zealand (DWSNZ), possibly small supplies located in the rural areas rather than large supplies supplying drinking water to cities and towns.

Drinking water is rendered microbiologically safe through application of microbial preventive and eliminative measures to raw water. Post-treatment microbial safety is maintained

by allowing appropriate concentrations of treatment chemicals and/or their by-products in the distribution network. Factors such as intermittent water supply and cracked or old pipes can exacerbate the post-treatment microbial risk. In order for the treatment measures to be effective, regular pathogen monitoring in both raw and treated water is crucial. Monitoring the presence of such pathogens in drinking water is commonly through the use of surrogates or indicator organisms. Commonly used surrogates include coliforms, coliphages, *E. coli* and *Enterococcus*. These surrogates are detected using a variety of laboratory techniques such as membrane filter technique (MFT), multiple-tube fermentation (MTF), enzyme-linked immunosorbent assay (ELISA), fluorescence *in situ* hybridisation (FISH), polymerase chain reaction (PCR) and loop-mediated isothermal amplification (LAMP). Of these techniques, MFT and MTF are more widely used because they are simple to use and are relatively cheap. In general, the basic versions of these techniques require a minimum of 24–48 hours to produce results.

The correlation between surrogates and pathogens in water has been a source of debate for many years, with numerous studies reporting conflicting outcomes. In addition, the use of surrogates has for many years attracted criticism for its lack of consistency in terms of specificity and sensitivity in detecting all waterborne pathogens. Despite this inadequacy, the use of surrogates is still widely applied in the water industry. Further, the differential survival abilities of surrogates in different water conditions has lead to the use of different surrogates in freshwater and saltwater. The main reason for using indicator organisms is because the practice is more practical and cost effective than monitoring individual pathogens. The impracticality and prohibitive cost in monitoring individual waterborne pathogens is partly due to the fact that most detection techniques in current use are cultured-dependent and many microbes are very difficult to grow in the laboratory. Thus, developing an assay for every pathogen, known and emerging, would be time consuming and very costly. Another reason is that none of the existing techniques can effectively identify all the pathogens in a single test. Given these circumstances, it would seem the alternative would have to be a culture-independent test capable of simultaneously detecting multiple organisms. Beyond pathogen detection in water is the ability to trace contamination to its source. This practice is known as microbial source tracking (MST) and in recent years has received a lot of research attention leading to the identification of many host-specific biomarkers. Such biomarkers have raised the prospects of development of a test that is able to accurately pin-point the source of a given pathogen.

Metagenomic techniques are among the culture-independent methods that offer a practical alternative to the current-dependent, pathogen identification methods. The capabilities of metagenomics to identify multiple organisms in a single test have been demonstrated through numerous microbial community profiling studies. For example, microbial communities harbouring different stages of the water distribution networks have been studied and

show wide variations. Overall, water distribution systems are dominated by *Proteobacteria*. Some types of bacteria tend to dominate in niche environments such as biofilms or free living amoebae. The niche environments provide some kind of protection to microbial communities living within them such that biocides can not reach the microbes easily. This renders the water treatment process less effective. The possible consequence of microbes surviving a treatment process is the positive selection of biocide (including disinfectant and antibiotic) resistance genes.

Metagenomic techniques have also been used to identify functional genes associated with microbial communities in drinking water. For example, genes associated with resistance to commonly used antibiotics such as those belonging to the  $\beta$ -lactam group have been identified in treated drinking water (Gomez-Alvarez et al., 2012; Schmeisser et al., 2003). This means that although the microbes may not be pathogenic, once consumed through drinking water they may transfer the resistance genes (Martínez, 2008) to the microflora in the consumer, e.g. through plasmids or horizontal gene transfer (Nikaido, 2009), thereby interfering with antibiotic treatment when required.  $\beta$ -lactam antibiotics, including penicillin derivatives, cephalosporins, monobactams and carbapenems, are among the most widely used broad spectrum antibiotics (Babic et al., 2006; Drawz and Bonomo, 2010; Wilke et al., 2005). Thus resistance to these antibiotics means that many conditions can neither be treated easily nor cheaply.

The ability of metagenomic techniques to identify a large variety of microbes and reveal various functional genes in a single test could revolutionise the approach on how drinking water quality standards are set. Equally, metagenomic techniques could be used to identify host and pathogen-specific markers for tests that are easy to use and less time-consuming, such as quantitative real-time polymerase chain reaction (qPCR) and LAMP. This would greatly enhance specificity and sensitivity in water quality testing.





# Three

## Factors associated with the presence of pathogens in drinking water sources of New Zealand

### 3.1 Background

Supplying microbiologically safe drinking water remains a challenge in many countries worldwide. Prüss et al. (2002) estimated that 4% (2.2 million) mortality and 6% (82.2 million) morbidity globally are caused by water, sanitation and hygiene related infections. Payment et al. (1997) reported that 14–40% of gastrointestinal illness in the United States of America (USA) was attributed to tap water from supply networks that met existing regulatory standards. Subsequently, Craun et al. (2006) observed that in the three decades to 2002 there was a general decline in waterborne disease outbreaks in the USA, from 42% (1971–1980) to 50% (1981–1990), 34% (1991–2000) and 14% (2001–2002). This decline was attributed to increased regulation, improved water treatment and monitoring of surface water systems. In New Zealand rates of hospitalisations due to enteric disease were found to be increasing in the twenty years to 2008 (Baker et al., 2012). Since waterborne infections are among the causes of enteric disease, it may be assumed that their rates had also been increasing.

Reduction of microbial contamination is key to reducing the prevalence of drinking water-associated gastrointestinal illness. A multiple barrier approach is generally considered the best way to reduce microbial contamination and ensure drinking water safety. This approach involves implementation of preventive or eliminative processes at various stages of the drinking water supply chain, from the source through treatment to distribution. Before preventive or eliminative measures are implemented, risk assessment is necessary in order that appropriate measures are selected (World Health Organization, 2012, chp. 10). Often a proactive approach is desired, i.e. prevent microbes from getting into water sources in the first place. Thus, the current study focuses on identifying factors that are associated with the presence of pathogens in source water.

In general, there are two main types of water sources: surface water (e.g. rivers and lakes) and groundwater (e.g. boreholes and wells). Previous studies have used a variety of tools to identify factors that affect the quality of water in such sources. For example, Cinque and Jayasuriya (2010) identified factors associated with water pollution within an agricultural catchment that supplied drinking water to Melbourne, Australia. Cinque and Jayasuriya (2010) used two statistical approaches, factor analysis and event mean concentration, and

found that erosion and runoff were correlated with the presence of microbial indicators in source waters. Bengraïne and Marhaba (2003) used principal component analysis (PCA) to investigate factors associated with hydrochemical and biological variations in water quality with reference to spatial and temporal aspects on the Passaic River in New Jersey, USA. Among the findings by Bengraïne and Marhaba were that there was seasonal variation in the water quality parameters. Ferguson et al. (2007) developed a process-based mathematical model (called pathogen catchment budgets or PCB) for predicting *Cryptosporidium*, *Giardia* and *E. coli* loads within a catchment supplying drinking water to Sydney, Australia. The model developed by Ferguson and co-workers identified pathogen excretion rates from both animals and humans as well as manure mobilisation rates as factors that were key for predicting pathogen loads. In another study, a multivariate approach was applied by Wu and Kuo (2012) to investigate factors that affected the quality of water in a catchment for Taipei, Taiwan, with reference to eutrophication<sup>1</sup>. The analytical techniques applied by Wu and Kuo were factor analysis, cluster analysis and discriminant analysis. Tea growing activities and wastewater discharges in the catchment were implicated in reservoir pollution and exhibited seasonal variations. Another example of application of mathematical modelling in water catchment pollution is the study by Unwin (2014), who used random forest (RF) analysis with 28 independent variables to predict rivers exceeding a defined threshold concentration (0.8 mg/L) of dissolved inorganic nitrogen. Unwin observed that there was a strong association between high concentrations of dissolved inorganic nitrogen and heavy pastoral activities in the catchment.

These examples demonstrate a proactive approach to risk assessment and provide tools that can be applied in assessing factors associated with source water quality. The current study was part of the effort by the New Zealand Ministry of Health (MoH) in assessing the risk within drinking water catchments in order that strategies for further reducing the public health risk associated with drinking water could be developed. The objective of the current study was to investigate water catchment factors associated with the presence of waterborne pathogens (microbes related to drinking water quality) in raw water intended for public consumption in New Zealand. To accomplish this objective, 20 drinking water sources were monitored for four waterborne pathogens (*Campylobacter*, *E. coli*, *Cryptosporidium* and *Giardia*) between September 2009 and March 2014. The relationship between the presence/absence of pathogens in source (raw) water and the catchment attributes was analysed using both parametric and non-parametric statistical approaches. Although viruses that are pathogenic to humans have been isolated in New Zealand surface drinking water sources (Williamson et al., 2011), they were not included in the current study because of lack of laboratory capability to analyse viral samples at molecular epidemiology and public health laboratory (<sup>m</sup>EpiLab).

---

<sup>1</sup>Enrichment of a water body with nutrients that stimulate aquatic plant growth resulting in dissolved oxygen depletion; often related to land use in the catchment

## 3.2 Materials and methods

### 3.2.1 Study sites

This was a prospective longitudinal study comprising 20 water abstraction sites, with their associated catchments, located throughout New Zealand. The primary unit of study was water source (abstraction site or intake). The abstraction sites were selected to represent the different types of catchments supplying drinking water to the public in New Zealand. Sixteen of the sites were surface water sources while four were groundwater sources (Table 3.1 and Figure 3.1). The surface water sources were located on rivers, streams and creeks; fourteen in the North Island and two in the South Island. The ground sources were shallow boreholes with two located in each of the North and South Islands. Generally, these sources supplied raw water to distribution networks that received water from one or more other sources, except for Whakarewarewa Forest Spring (G00183), Seadon Well (G00197) and Waingawa River (S00383) which were the sole sources in their respective networks. In a simple standard drinking water distribution network water is abstracted at a source, treated at a treatment plant and distributed to the public through a distribution zone. Among the study sources, Lower Huia Dam (S00092) supplied the highest number of zones (Table 3.1). The sources supplied populations ranging from 750 (Seadon Well (G00197)) to over 1 million (Waikato River-Tuakau (S00865)), serving 2.3 million people in total, equivalent to 51.5 % of the New Zealand population<sup>2</sup> in 2011.

Sites on the Wainuiomata River (S00120), Orongorongo River (S00121) and Big Huia Creek (S00434) were located in relatively close proximity and all supplied water to Wellington, Lower/Upper Hutt and Porirua, including the surrounding areas. This implies that the population in these areas were supplied by multiple water sources. Lake Karapiro (S00009), Waikato River-Hamilton (S00041) and Waikato River-Tuakau (S00865) sites shared the same catchment, which was the largest in the current study, measuring  $1.4 \times 10^4 \text{ km}^2$ . The Waikato River-Tuakau site was furthest upstream, receiving water from the entire catchment, followed by Lake Karapiro ( $7.8 \times 10^3 \text{ km}^2$ ) and Waikato River-Hamilton ( $8.3 \times 10^3 \text{ km}^2$ ).

### 3.2.2 Sample collection

The study drinking water sources were monitored for four microbes associated with drinking water quality (*Campylobacter*, *E. coli*, *Cryptosporidium* and *Giardia*) from September 2009 to March 2014. Sample collection from these sources was carried out every three months. It is worth noting that no sample was collected during the month of December during the study period, due to the Christmas break. At each site three water samples were collected: 100 mL for *Campylobacter* isolation, another 100 mL for *E. coli* enumeration and 100 L filtered through a Filtamax<sup>®</sup> filter for the enumeration of *Cryptosporidium*

<sup>2</sup>[http://www.stats.govt.nz/browse\\_for\\_stats/snapshots-of-nz/nz-in-profile-2012/population.aspx](http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/nz-in-profile-2012/population.aspx)

oocysts and *Giardia* cysts. All samples were stored on ice soon after sampling and also during transportation to <sup>m</sup>EpiLab at the Hopkirk Research Institute, Massey University, for examination.

**Table 3.1:** Description of the twenty drinking water sources monitored for microbes associated with drinking water quality between September 2009 and March 2014, New Zealand. Also tabulated is the size of population (2011 estimates) supplied by each water source and the number of distribution zones through which the water is distributed to the public. The sources beginning in the letter **G** are groundwater sources while those beginning with the letter **S** are surface water sources.

Source	Location	Description	Population	Zones	City/Town
G00122	NW Christchurch Aquifer 1	Depth 20-40m	83000	1	Christchurch
G00183	Whakarewarewa Forest Spring	Natural spring in Whakarewarewa Forest, Rotorua. Depth <10m	10060	1	Rotorua
G00197	Seadown Well	Near the Opihi River, part of the Timaru water supply. Depth <10m	750	1	Timaru
G01679	Hicks Road Spring	Fed by two springs at Maungatautari to a balance tank, where it then gravitates to the Karapiro Reservoirs. Depth <10m at well head	13368	1	Waikato
S00009	Lake Karapiro	Supplies Karapiro water treatment plant	13500	2	Waipa
S00041	Waikato River - Hamilton	Waioara Terrace treatment plant, Hamilton	137840	11	Hamilton
S00082	Turitea Dam	Concrete-faced dam located in Turitea Reserve in Palmerston North	67653	4	Palmerston North
S00088	Oroua River	Abstracted upstream of the Almadale water treatment plant in Feilding	13000	1	Manawatu
S00092	Lower Huia Dam	Earth dam in Waitakere Ranges	732611	38	Auckland
S00099	Waitakere Dam	Concrete dam in Waitakere Ranges	604544	29	Auckland
S00118	Hutt River	Abstracted at Kaitoke, supplies Te Marua water treatment plant	536990	23	Wellington
S00120	Wainuiomata River	In Rimutaka Ranges, supplies Wainuiomata water treatment plant	296835	11	Wellington
S00121	Orongorongo River	In Rimutaka Ranges, supplies Wainuiomata water treatment plant	296835	11	Wellington
S00124	Ashley River	Infiltration gallery at River Road, supplies Rangiora water treatment plant	13346	1	Waimakariri
S00200	Pareora River	Abstracted at Upper Gorge, supplies the Claremont Reservoir in Timaru	27368	2	Timaru
S00298	Waiorohi Stream	Supplies Oropi Road water treatment plant in Tauranga	103783	2	Tauranga
S00299	Tautau Stream	Supplies Joyce Rd water treatment plant in Tauranga	103783	2	Tauranga
S00383	Waingawa River	Abstracted upstream of the Kaituna water treatment plant in Masterton	19000	1	Masterton
S00434	Big Huia Creek	In Orongorongo Valley, supplies Wainuiomata water treatment plant	296835	11	The Hutt
S00865	Waikato River - Tuakau	Tuakau treatment plant	921008	25	Auckland



**Figure 3.1:** Location of the twenty study drinking water sources monitored for microbes associated with drinking water quality between September 2009 and March 2014, New Zealand.

### 3.2.3 Laboratory procedures

#### *Campylobacter* and *E. coli*:

Samples for *Campylobacter* isolation were filtered through 0.45  $\mu\text{m}$ -pore, 47 mm-diameter disks (Millipore Corporation; Massachusetts, USA) soon after arrival at *m*EpiLab, Hopkirk Research Institute, Massey University. The filter disks were immediately incubated in Bolton's broth under microaerophilic conditions (85 %  $\text{N}_2$ , 5 %  $\text{O}_2$ , 0 %  $\text{H}_2$  and 10 %  $\text{CO}_2$ ) at 42 °C for 48 h using a Macs-VA500 microaerophilic workstation (Don Whitley Scientific Limited; Yorkshire, UK) in order to enrich *Campylobacter*. After 48 h of enrichment, the

broth was cultured onto selective medium, modified charcoal cefoperazone deoxycholate agar (mCCDA), for a further 48 h. From each mCCDA plate with *Campylobacter*-like growth two colonies were subcultured onto separate horse blood agar plates and incubated in the microaerophilic workstation for another 48 h. A presumptive *Campylobacter*-positive result was declared if the blood agar growth exhibited typical *Campylobacter* phenotypic characteristics. A confirmatory positive result was obtained after subjecting the isolates to *Campylobacter* genus (Stucki et al., 1995) and species-specific (e.g. *C. jejuni* and *C. coli*) polymerase chain reaction (PCR)s. The PCR protocols used are outlined in Section 5.2.5 on page 112. Samples for *E. coli* enumeration were submitted to the Central Environmental Laboratories, accredited regional laboratories for water quality testing located in Palmerston North. The Central Environmental Laboratories used a modified Colilert<sup>®</sup> (IDEXX Laboratories Inc.; Maine, USA) method for the enumeration of *E. coli*. A summary of this method is provided in Section 5.2.6 on page 114

### ***Cryptosporidium* and *Giardia*:**

Filta-Max<sup>®</sup> filters were used for screening *Cryptosporidium* and *Giardia* following a modified United States Environmental Protection Agency (USEPA) method 1623 (United States Environmental Protection Agency, 2012) (List 6 on page 199). In summary, the filter module was removed from a transportation bag and dismantled to recover foam disks, along with residual fluid, and placed in a Stomacher<sup>®</sup> 3500 bag (Seward Ltd; West Sussex, UK) containing 500 mL phosphate buffered saline (PBS). The mixture was homogenised using a Stomacher<sup>®</sup> 3500 (Seward Ltd; West Sussex, UK) for 10 min. After homogenisation, the eluent was transferred into a 500 mL conical centrifuge tube and centrifuged at 3000 *g* for 15 min at 10 °C using a Sorvall RT7 Benchtop centrifuge (GMI Inc.; Minnesota, USA). The top 450 mL supernatant was aspirated off and discarded. The remaining fluid was vortexed to resuspend the pellet that had collected at the bottom of the tube and then transferred into a 50 mL centrifuge tube and centrifuged as before. The supernatant was aspirated, as before, leaving 10 mL in which the pellet was resuspended. Immunomagnetic separation (IMS) using a Dynabeads GC-Combo kit (Invitrogen Corporation; California, USA) was applied to the mixture. The resultant 50 µL fluid was transferred onto a fluorescence microscopy slide with reaction wells (Marienfeld GmbH & Co. KG; Lauda-Königshofen, Germany) and placed in a humidity chamber. Then 50 µL of diluted Aqua-Glo<sup>®3</sup> stain (Waterborne Inc.; New Orleans, USA) was added and the slide was incubated at 37 °C for 30–60 min. The stain was washed off using 50 µL PBS and the slide was air dried for 2 min.

Once the slide had been prepared, a BX 60 fluorescence microscope (Olympus; Tokyo, Japan) was used to scan for *Cryptosporidium* oocysts and *Giardia* cysts. An initial scan was conducted at 200 X magnification followed by a detailed scan at 400 X magnification, focusing on areas with fluorescent particles. A presumptive *Cryptosporidium*-positive result

---

<sup>3</sup>Contains fluorescein-labeled monoclonal antibodies specific to *Cryptosporidium parvum* and *Giardia lamblia* as well as 4',6-diamidino-2-phenylindole (DAPI).

was declared if apple-green ovoid or spherical particle(s), measuring 4–6  $\mu\text{m}$ , were observed at 200 X magnification. A sample was confirmed *Cryptosporidium*-positive if one of the following was observed at 400 X magnification, under a DAPI filter: light blue internal staining with a green rim, intense blue internal staining or up to four distinct sky-blue nuclei. Equivalently, a presumptive *Giardia*-positive result was declared if apple-green round or ovoid particle(s) with bright edges, measuring 5–15  $\mu\text{m}$  in width and 8–18  $\mu\text{m}$  in length, were observed at 200 X magnification. A confirmed *Giardia*-positive test was declared if observations similar to those described for *Cryptosporidium* were made at 400 X magnification and under a DAPI filter. The (oo)cysts were enumerated using Equation 3.1 and an example is provided in Box 2.

$$\text{Count of (oo)cysts/100 L} = \frac{\text{Number of oocysts observed} \times 100}{F \times V} \quad (3.1)$$

**Box 2:** Enumeration of *Cryptosporidium* oocysts and *Giardia* cysts

Suppose a 120 L ( $V = 120$ ) water sample was collected and processed, yielding a 1 mL ( $P = 1$ ) pellet. If half the pellet ( $F = 0.5$ ) was purified using immunomagnetic separation and examined microscopically, and assuming that two *Cryptosporidium* oocysts and three *Giardia* cysts were observed, the number of (oo)cysts in the 100 L sample can be calculated as follows:

*Cryptosporidium*:

$$\begin{aligned} \text{Count of oocysts/100 L} &= \frac{\text{Number of oocysts observed} \times 100}{F \times V} \\ &= \frac{2 \times 100}{0.5 \times 120} \\ &= 3 \end{aligned}$$

*Giardia*:

$$\begin{aligned} \text{Count of cysts/100 L} &= \frac{\text{Number of cysts observed} \times 100}{F \times V} \\ &= \frac{3 \times 100}{0.5 \times 120} \\ &= 5 \end{aligned}$$

This calculation is based on USEPA ICR Protozoan Method for Detecting *Giardia* cysts and *Cryptosporidium* oocysts in water by a Fluorescent Antibody Procedure (EPA/814-B-95-003)



### 3.2.4 Data

#### Laboratory data

The data generated in the current study were presence/absence of *Campylobacter* per 100 mL, the most probable number (MPN) of *E. coli* organisms per 100 mL, the count of *Cryptosporidium* oocysts per 100 L and the count of *Giardia* cysts per 100 L.

#### Geospatial data

Geospatial attributes for the study sites were extracted from digital map files (Environmental Systems Research Institute (ESRI) shapefiles) created by government-supported institutions and were freely available on the internet (Table A.1 on page 182). An exception to this was the farmland shapefile which was purchased fromASUREQuality New Zealand Limited<sup>4</sup>. The geospatial attributes were those related to the features within the surface water catchments. A catchment was created by locating the river or stream segment on which the water abstraction point was located in the rivers and streams shapefile. Hereafter, *river* will be used to refer to river, stream or creek. Once the source river segment had been sited, all its tributaries upstream were identified using the river environment classification (REC) tracer tool in ArcView 3.2a to obtain a riverbed network. Using the **select by location** function in ArcMap, the riverbed network was used to select corresponding catchment polygons from a REC watershed shapefile. The catchment polygons were dissolved to obtain catchment outer boundaries, which were then intersected with various other shapefiles e.g. farmland and land cover shapefiles in order to extract animal population and land cover attributes, respectively. Animal densities for each catchment were then calculated by dividing the number of animals present in the catchment by the area (square kilometres) of the catchment. Similarly, areal proportions covered by each attribute were calculated.

### 3.2.5 Statistical techniques

The data used in the present study comprised numerous variables with possible complex interactions. The relationships among these variables were analysed using both parametric and non-parametric techniques. A non-parametric regression analysis approach using random forest (RF) (Breiman, 2001) analysis, was used for both variable selection and prediction purposes. Parameter estimations for the reduced number of variables identified by RF analysis as important for the prediction of the presence of microbes in source waters were performed using a parametric regression framework — generalised linear modelling.

#### Random forests

RF analysis is a non-parametric regression approach based on machine learning and also classification and regression trees (CART) (Breiman et al., 1984). By incorporating the

---

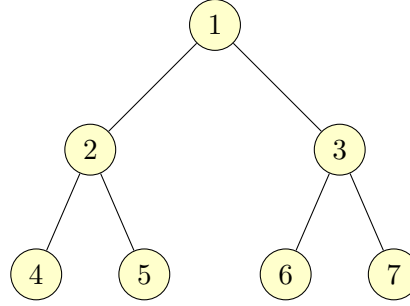
<sup>4</sup>A commercial company owned by the government of New Zealand

machine learning and CART principles, RF analysis may be used for both data mining and making predictions. Strobl et al. (2009) provides a detailed discussion on the performance of RF methods and their implementation using the package `party` (Hothorn et al., 2006a) in **R** (R Core Team, 2013).

The main objective of machine learning is to construct models that can learn from data and perform predictions. The data from which the models learn are called training or learning data. The CART algorithm finds patterns in the data by recursively splitting data into clusters containing observations with similar response values. The recursive splitting of observations into nodes (i.e. growing of a decision tree) continues until a stop condition is reached. For example, a stop condition may state that the splitting must end when a defined minimum number of elements (observations or data points) is left in a terminal node (see Figure 3.2 for terminology on the different parts of a basic decision tree) or when a given minimum change in the impurity (a *pure* node contains elements from a single response variable) measure is not succeeded any more by any variable. A stop condition may also state that the splitting should continue until all terminal nodes are pure. Other stopping criteria that incorporate the distribution of the splitting criterion have been suggested (Hothorn et al., 2006b).

In a basic decision tree model, the principle of impurity reduction is used for selecting the splitting variable and cut-point, in other words, an explanatory variable that is strongly associated with the response variable is used to split the data. This means that each split in the tree building process results in daughter nodes that proportionally contain more elements with similar response values than the parent node. RF is an ensemble method as it is a collection of classification or regression trees. The trees grown by RF are called classification trees if categorical explanatory variables are used and regression trees if continuous variables are used. In each node impurity is quantified using an entropy (uncertainty) measure such as the Gini Index or the Shannon Entropy. Maximum entropy is reached when a node has an equal number of elements from the response categories, conversely minimum entropy is reached when a node contains 100% elements from a single response category. The amount of impurity reduction attributed to a split is the difference between the impurity in the parent node and the average impurity in the two daughter nodes. The principle of impurity reduction is analogous to measuring the strength of association between the splitting and the response variables (Strobl et al., 2009).

The major disadvantage of simple tree models is that they are highly unstable, minor changes in the learning data can cause significant changes to the outcome (Hastie et al., 2009; Strobl et al., 2009). Ensemble methods are therefore designed to overcome this deficiency by constructing a large number of trees and averaging over them. In a RF each tree is constructed as described for CART but with two modifications: the first modifica-



**Figure 3.2:** Schematic representation of a basic decision tree. Node 1 is the root node, 2 and 3 are internal nodes while 4-7 are terminal nodes. All nodes, other than terminal nodes, are also called parent nodes as they give rise to two daughter nodes.

tion is that a random sample of the learning data, commonly drawn by bootstrap sampling (drawing with replacement), is used instead of the entire dataset. The second modification is that in a given split only a predefined number of predictor variables is used and these are randomly selected from the available variables. About one-third of the elements are not included in the bootstrap sample and not used in the construction of the resultant tree. The elements that are left out of the bootstrap sample are known as out-of-bag (OOB) elements and are used for estimating the unbiased classification error as each tree is added to the forest and also for estimating variable importance.

Since each tree is constructed using a different bootstrap sample and a different set of predictor variables, the forest is populated with trees that can be very different, usually grown very large without any stopping or pruning. The prediction power of RF is substantially increased as a result of averaging the predictions of these diverse collection of trees. The predefined number of randomly selected splitting variables (also known as `mtry`) as well as the overall number of trees (`ntree`) are parameters of RF that affect the stability of the results (Strobl et al., 2009, p. 16).

**Predictions:** In an ensemble of trees the predictions of individual trees are combined by weighted or unweighted averaging in regression (Equation 3.2) or voting in classification (Equation 3.3). *Voting* means that each element, with associated explanatory variable values, is classified by every tree in the ensemble. In this way, every tree predicts a category for a given element and the category that accrues the most *votes* is the prediction of the ensemble. For regression, the probabilities from individuals trees are averaged (Gatnar, 2008). Predictions can be performed using the entire dataset (ordinary prediction) or OOB data (OOB prediction). Ordinary prediction tends to underestimate the error rate while the OOB prediction is considered more realistic. This is because the OOB data were not involved in the construction of a given tree, hence, serve as a good test for the prediction ability of that tree (Hastie et al., 2009, p. 593).

Consider an ensemble of trees  $\{T_b\}_1^B$ , where  $T_b$  is a tree in a set of RF trees  $T_1, \dots, T_B$ , a prediction at a new point  $x$  is made by:

$$\text{Regression : } \hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (3.2)$$

and

$$\text{Classification : } \hat{C}_{rf}^B(x) = \text{majority vote}\{\hat{C}_b(x)\}_1^B \quad (3.3)$$

where  $\hat{C}_b(x)$  is the class prediction of the  $b$ th RF tree.

Merits of using RF analysis include its ability to handle ‘ $n < p$ ’ situations, where  $n$  is the number of elements (observations or data points) and  $p$  is the number of variables, situations that are not handled by parametric methods. RF is less likely to overfit the data yet has the ability to capture complex interactions in the data. In addition, if the trees are grown sufficiently deep, RF has relatively low bias ([ibid.](#)).

### Regression analysis

Regression methods are used for describing a relationship between an outcome (dependent or response) variable and explanatory (independent or predictor) variable(s). In using such methods the aim is to apply the best fitting and most parsimonious models which provide biologically plausible description of the outcome-explanatory variable relationship (Hosmer Jr and Lemeshow, [2000](#)). In the current study, generalised linear models, discussed in Section 5.3.1 (page 117), were used.

### 3.2.6 Data analysis

#### Random forests analysis

The data used in the RF analysis were those from surface water sources only, the data from groundwater sources were not analysed using this technique because all groundwater sources were negative for all four microbes on all sampling occasions except for the Hicks Road Spring (G01679) site which was positive for *Campylobacter* on a single occasion. RF analysis was conducted using the **R** package **party** (Hothorn et al., [2006a](#); Strobl et al., [2007](#), [2008](#)). Four separate RFs were constructed, one for each of the microbes under investigation as an outcome variable. The microbes were part of the explanatory variables in the models in which they were not outcome variables. Other explanatory variables included catchment size, season, soil temperature<sup>5</sup>, land cover, lithology types and domestic ruminant densities (Table 3.2). In total, there were 31 explanatory variables

<sup>5</sup>In soil taxonomy, mean annual temperatures measured at a depth of 50 cm from the surface are used to define soil regimes. Among these regimes are thermic (mean annual temperature  $\geq 15^\circ\text{C}$  but  $< 22^\circ\text{C}$ ) and mesic (mean annual temperature  $\geq 8^\circ\text{C}$  but  $< 15^\circ\text{C}$ ).

in each RF. *Campylobacter* (present/absent) and season were the only variables introduced in the RF models as categorical variables while all other variables were introduced as continuous variables. For each RF, variable importance was computed in order to identify variables that were essential for the prediction of the level of a given microbe in source water samples using the data from September 2009 to March 2014. In addition, the levels of the four microbes in samples collected during the January-March 2014 sampling round were predicted using RF analysis. In order to perform these predictions, data from September 2009 to November 2013 were used as the learning data. For the purpose of performing the predictions each outcome variable was categorised as follows:

- *Campylobacter* - level 1 (absent), level 2 (present)
- *E. coli* - level 1 ( $\leq 100$ ), level 2 (101-200), level 3 (201-500), level 4 ( $>500$ )
- *Cryptosporidium* - level 1 ( $\leq 3$ ), level 2 (4-7), level 3 (8-11), level 4 ( $>11$ )
- *Giardia* - level 1 ( $\leq 3$ ), level 2 (4-7), level 3 (8-11), level 4 ( $>11$ )

### Regression analysis

Five variables that had the highest RF variable importance scores for each microbe were selected for inclusion in the regression analysis. These variables are among those described in Table 3.2. *Campylobacter* was dichotomised as present or absent, the dichotomy levels for *E. coli* were  $< 200$  and  $\geq 200$  MPN/100 mL while those for *Cryptosporidium* and *Giardia* were  $< 1$  and  $\geq 1$  (oo)cysts/100 L. The five explanatory variables considered for inclusion in the *Campylobacter* model included dichotomised *E. coli* concentrations ( $< 200$  versus  $\geq 200$  MPN/100 mL), beef cattle densities (animal counts per km<sup>2</sup>), dairy cattle densities (animal counts per km<sup>2</sup>), tephra lapilli soil type (areal proportion) and grassland (areal proportion). For the *E. coli* model the explanatory variables considered were: *Campylobacter* (present or absent), beef cattle densities, dairy cattle densities, greywacke soil type (areal proportion) and tephra lapilli soil type. *Giardia* concentrations (cysts/100 L), catchment size (km<sup>2</sup>) at the logarithmic scale, dairy cattle densities, igneous volcanic soil type (areal proportion) and warm mesic soil temperature (areal proportion) were considered for inclusion in the *Cryptosporidium* model. Catchment size at the logarithmic scale, igneous volcanic soil type, dairy cattle densities, cold mesic soil temperature (areal proportion) and wetlands were considered for inclusion in the *Giardia* model. Catchment size was logarithmically transformed because it had a highly skewed distribution.

Initially, microbe-specific univariable generalised linear mixed model (GLMM)s belonging to the binomial family, in which drinking water source was the random effect were implemented for each of the five explanatory variables. The statistical significance in the univariable models was set at  $P$  value= 0.10, thus variables having equal or less than this value were considered for inclusion in a subsequent multivariable model. The microbe and ruminant variables were considered study factors, therefore, were included in multivariable

**Table 3.2:** Description of variables used in both random forests and regression analysis.

Variable label	Description
Catchment size	Size of catchment in km <sup>2</sup>
<b>Domestic ruminant densities<sup>§</sup></b>	
Beef density	Beef cattle
Dairy density	Dairy cattle
Deer density	Domestic deer
Sheep density	Sheep
<b>Land cover<sup>†</sup></b>	
Alpine prop	Alpine/sub-alpine vegetation or permanent snow/ice
Cropland prop	Orchard, vineyard or perennial/short-rotation crop
Forest prop	Both indigenous and exotic forest
Grassland prop	High/low producing, tall tussock, or depleted grassland
Gravel prop	Gravel, landslide, surface mine or dump
Settlement prop	Built-up area, urban parkland/open space or transport infrastructure
Shrubland prop	Gorse, broom, flax, fern, exotic shrub, matagouri or grey shrub
Wetland prop	River, lake, pond or herbaceous freshwater/saline vegetation
<b>Mean annual soil temperature<sup>†</sup></b>	
Thermic prop	15–22 °C
Warm Mesic prop	11–15 °C
Cool Mesic prop	8–11 °C for < 60days
Cold Mesic prop	8–11 °C for > 60days
<b>Lithology<sup>†</sup></b>	
Alluvium prop	Alluvial rock
Greywacke prop	Greywacke rock, including deep weathered or tuffaceous
Igneous Volcanics prop	Igneous intrusives, volcanics or volcanogenics
Loess prop	Loess rock (>1–2 thick)
Rhyolite prop	Rhyolite rock
Sedimentary Rock prop	Sedimentary rock
Tephra Ash Lapilli prop	Tephra, ash or lapilli
Weathered Mafic prop	Deep weathered mafic rock
Weathered Soft Rock prop	Deep weathered soft rock
<b>Season</b>	
Autumn	March – May
Winter	June – August
Spring	September – November
Summer	December – February

<sup>§</sup>Number of animals per km<sup>2</sup> of the catchment

<sup>†</sup>Proportion of catchment area covered by the variable

models regardless of their statistical significance in the univariable models. Thereafter, multivariable GLMMs similar to those described for univariable models were implemented for each of the four microbes using the **R** package **lme4** (Bates et al., 2014).

### 3.3 Results

A description of the twenty drinking water sources monitored for microbial contamination between September 2009 and March 2014 is presented in Table 3.1 on page 48. The population supplied by a single water source ranged from 750 to 1 064 876. Some communities were supplied by more than one study source e.g. Lower Huia Dam (S00092), Waitekere Dam (S00099) and Waikato River-Tuakau (S00865) sources supplied drinking water to communities living in Auckland. The total population supplied by the twenty water sources, excluding duplicate populations, was 2 270 008 based on the 2011 Register of Community Drinking-water Supplies in New Zealand estimates. This was equivalent to 51.1 % of the 2013 New Zealand population. Drinking water was supplied to this population through distribution zones ranging from 1 to 23 per water source, totalling 71 zones (excluding duplicates). The geospatial attributes for the surface water catchments are summarised in Table A.2 (page 183) while attributes for the groundwater sources are summarised in Table A.3 (page 184). In addition, topological maps for surface water catchments showing land cover and lithology (soil type) attributes are presented in Figures A.12–A.14 (pages 186–188) and Figures A.15–A.17 (pages A.15–A.17), respectively. Topological maps for groundwater sources are not shown because all samples (except one sample from the Hicks Road Spring site that was positive for *Campylobacter*) collected from these sources were negative for all four microbes and thus were excluded from attribute-related analyses.

#### 3.3.1 Descriptive statistics

Eighteen rounds of sampling were conducted between September 2009 and March 2014. During this period a total of 360 samples were collected for each study microbe i.e. eighteen samples for each of the four monitored microbes from each of the twenty drinking water sources (abstraction sites). The percentage of positive samples for each microbe at each abstraction site during the study period are presented in Table 3.3. Sixteen (80 %) of the twenty sites had one or more *Campylobacter*-positive samples, ten (50 %) had one or more samples with *E. coli* concentrations  $\geq 200$  MPN/100 mL and six (30 %) had one or more samples positive for either *Cryptosporidium* or *Giardia*. Overall, the most consistently contaminated sites were the two Waikato River sites (S00041 and S00865) followed by the Waiorohi Stream (S00298) and Oroua River (S00088) sites.

Of the 360 samples, 67 (18.6 %) were positive for *Campylobacter*. The only *Campylobacter*-positive sample over the study period among the groundwater sites was collected from the Hicks Road Spring (G01679) site. The surface sources with the highest number of *Campylobacter*-positive samples included the Waiorohi Stream and the two Waikato River sites. For *E. coli*, 264 out of 360 (73.3 %) samples had a concentration  $\geq 1$  MPN/100 mL (not shown) while 28 out of 360 (7.8 %) had concentrations  $\geq 200$  MPN/100 mL. For the protozoan organisms, 31 (8.6 %) and 37 (10.3 %) of the 360 samples were positive

( $\geq 1$  (oo)cysts/100 mL) for *Cryptosporidium* and *Giardia*, respectively. Four sites (Oroua River, Waiorohi Stream and the two Waikato River sites) had samples positive for both *Cryptosporidium* and *Giardia* with the two Waikato River sites having the highest percentage of protozoa-positive samples.

Figure 3.3 shows the ten sites that had one or more samples containing *E. coli* at a concentration  $\geq 200$  MPN/100 mL and when the samples were collected. All the ten sites were surface water sources. In contrast, the four groundwater sites had low concentrations of *E. coli*, the maximum MPN ranging from 1 to 15 per 100 mL. Over the study period, samples with *E. coli* concentrations  $\geq 200$  MPN/100 mL were most frequently collected at the Waiorohi Stream (S00298) site, which also had the highest median *E. coli* concentration of 144 MPN/100 mL (95 % CI<sup>6</sup>: 60; 2745), followed by Waikato River-Tuakau (S00865) 71 MPN/100 mL (95 % CI: 18.5; 2,450.8), Oroua River 61 MPN/100 mL (95 % CI: 22.7; 1,225.8) and Waikato River-Hamilton (S00041) 52 MPN/100 mL (95 % CI: 27.3; 1,524.5).

**Table 3.3:** Percentage of positive samples from the twenty study drinking water abstraction sites located throughout New Zealand. The samples were collected in eighteen rounds of sampling (i.e.  $n = 18$  for each pathogen and site) between September 2009 and March 2014.

Site	Location	Type	<i>Campylobacter</i>	<i>E. coli</i> *	<i>Cryptosporidium</i>	<i>Giardia</i>
G00122	NW Christchurch Aquifer 1	Ground water	0.0	0.0	0.0	0.0
G00183	Whakarewarewa Forest Spring	Ground water	0.0	0.0	0.0	0.0
G00197	Seadown Well	Ground water	0.0	0.0	0.0	0.0
G01679	Hicks Road Spring	Ground water	5.6	0.0	0.0	0.0
S00009	Lake Karapiro	Surface water	11.1	5.6	0.0	0.0
S00041	Waikato River - Hamilton	Surface water	55.6	22.2	55.6	77.8
S00082	Turitea Dam	Surface water	16.7	0.0	0.0	0.0
S00088	Oroua River	Surface water	44.4	22.2	16.7	16.7
S00092	Lower Huia Dam	Surface water	16.7	5.6	11.1	0.0
S00099	Waitakere Dam	Surface water	11.1	5.6	0.0	5.6
S00118	Hutt River	Surface water	5.6	0.0	0.0	0.0
S00120	Wainuiomata River	Surface water	11.1	0.0	0.0	0.0
S00121	Orongorongo River	Surface water	11.1	0.0	0.0	0.0
S00124	Ashley River	Surface water	0.0	5.6	5.6	0.0
S00200	Pareora River	Surface water	33.3	5.6	0.0	0.0
S00298	Waiorohi Stream	Surface water	77.8	44.4	22.2	16.7
S00299	Tautau Stream	Surface water	11.1	16.7	0.0	5.6
S00383	Waingawa River	Surface water	5.6	0.0	0.0	0.0
S00434	Big Huia Creek	Surface water	5.6	0.0	0.0	0.0
S00865	Waikato River - Tuakau	Surface water	50.0	22.2	61.1	83.3

\*Percentage of samples that had a concentration  $\geq 200$  MPN per 100mL

<sup>6</sup>Confidence interval



Eight sites that had one or more samples with  $\geq 1$  (oo)cysts/100 L and month of collection are shown in Figure 3.4. None of the groundwater sources were positive for either *Cryptosporidium* or *Giardia*. The highest concentration of *Cryptosporidium* was 12 oocysts/100 L while that of *Giardia* was 18 cysts/100 L, both samples were from the Waikato River-Tuakau (S00865) site. In general, the highest frequencies of *Cryptosporidium* and *Giardia* contamination were recorded at the Waikato River sites.

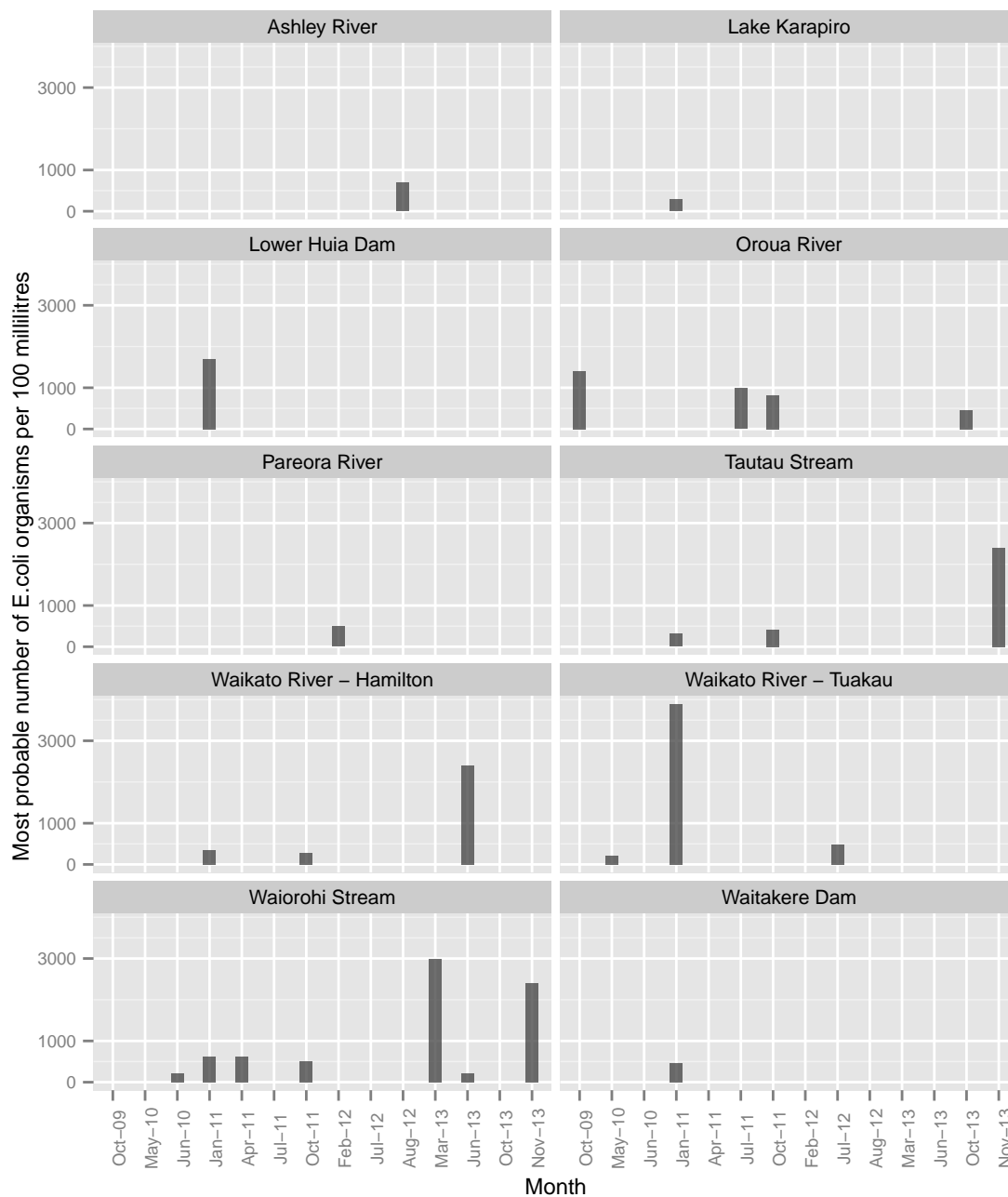
The number of sampling occasions at each drinking water source (abstraction site) and the number of positive samples for each monitored microbe within each calendar month and season, over the study period, are shown in Table 3.4. No samples were collected in the month of December while October was the month in which sampling was most conducted. In general, sample collection was variable by calendar month and across the sites. A similar pattern was observed for season. Sample collection was mostly conducted in spring followed by autumn, winter and summer.

**Table 3.4:** Number of sampling occasions and positive samples for each drinking water abstraction site per calendar month and per season between September 2009 and March 2014, New Zealand.

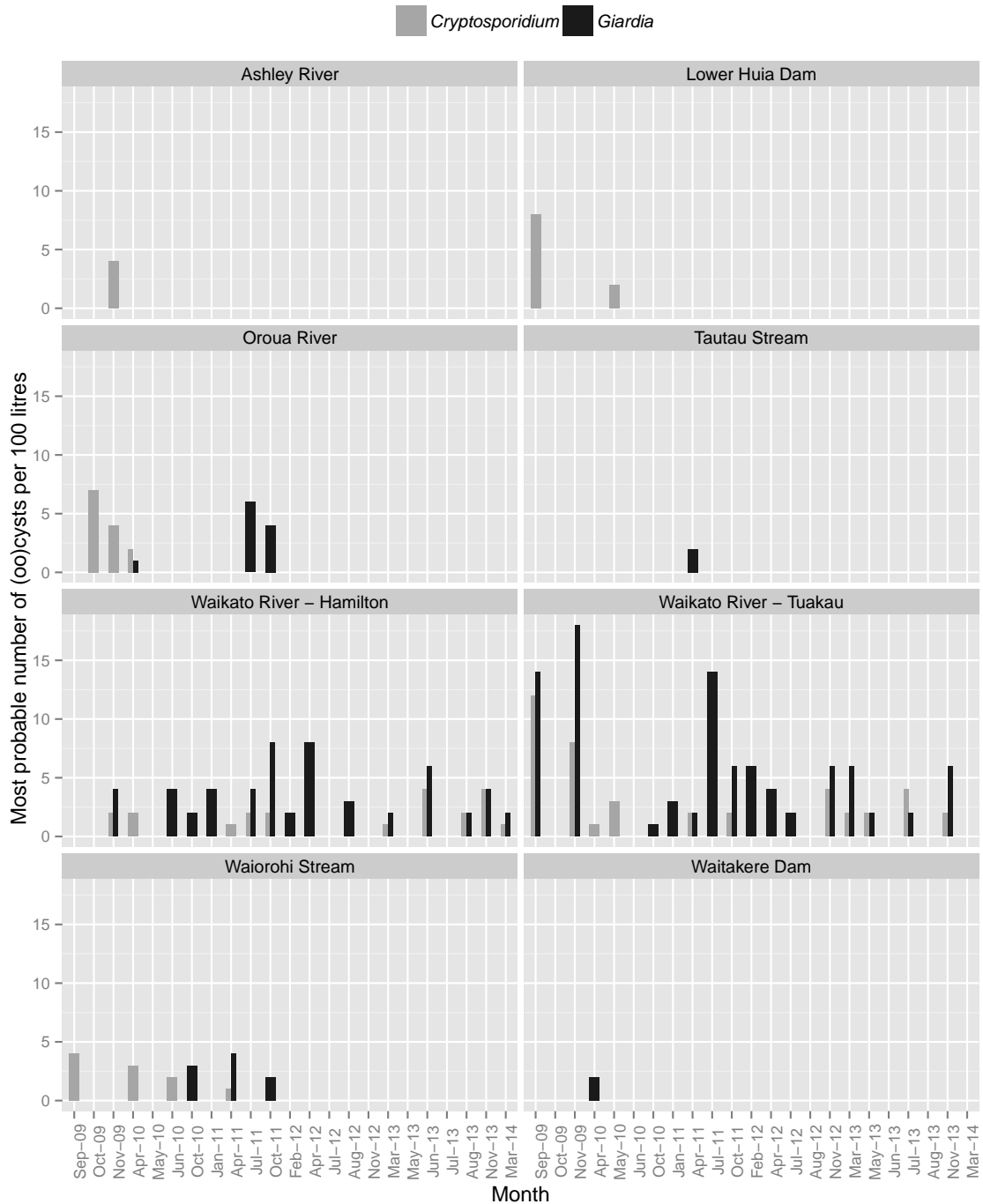
Site	Calendar month												Season			
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Total	Aut <sup>§</sup>	Win <sup>§</sup>	Spr <sup>§</sup>	Sum <sup>§</sup>
<b>Sampling occasions</b>																
G00122	1	2	1	2	1	2	0	3	0	4	2	18	4	5	6	3
G00183	1	1	2	2	1	2	1	2	1	3	2	18	5	5	6	2
G00197	1	2	1	2	1	2	0	3	0	4	2	18	4	5	6	3
G01679	1	1	2	3	0	2	1	2	1	3	2	18	5	5	6	2
S00009	1	1	2	3	0	2	1	2	1	3	2	18	5	5	6	2
S00041	1	1	2	3	0	2	1	2	1	3	2	18	5	5	6	2
S00082	1	1	2	2	1	2	3	0	0	4	2	18	5	5	6	2
S00088	1	1	2	2	2	1	2	1	0	5	1	18	6	4	6	2
S00092	1	2	1	3	2	0	3	0	1	2	3	18	6	3	6	3
S00099	1	2	1	3	2	0	3	0	1	2	3	18	6	3	6	3
S00118	1	2	1	2	3	0	3	0	0	5	1	18	6	3	6	3
S00120	1	2	1	2	3	0	3	0	0	5	1	18	6	3	6	3
S00121	1	2	1	2	3	0	3	0	0	5	1	18	6	3	6	3
S00124	1	2	1	2	1	2	0	3	0	3	3	18	4	5	6	3
S00200	1	2	1	2	1	2	0	3	0	4	2	18	4	5	6	3
S00298	1	1	2	2	1	2	1	2	1	3	2	18	5	5	6	2
S00299	1	1	2	2	1	2	1	2	1	3	2	18	5	5	6	2
S00383	1	1	2	2	1	2	2	1	0	5	1	18	5	5	6	2
S00434	1	2	1	2	3	0	3	0	0	5	1	18	6	3	6	3
S00865	1	2	1	3	2	0	3	0	1	2	3	18	6	3	6	3
Total	20	31	29	46	29	25	34	26	9	73	38	360	104	85	120	51
<b>Positives samples</b>																
Campy <sup>†</sup>	4	3	6	14	10	4	5	1	0	13	7	67	30	10	20	7
Ecoli <sup>†</sup>	7	1	1	1	1	3	2	1	0	6	2	25	3	6	8	8
Crypto <sup>†</sup>	0	0	3	7	3	2	2	1	3	3	7	31	13	5	13	0
Giard <sup>†</sup>	2	2	3	7	1	2	5	2	1	7	5	37	11	9	13	4
Total	13	6	13	29	15	11	14	5	4	29	21	160	57	30	54	19

<sup>§</sup>Season: Aut = Autumn; Win = Winter; Spr = Spring; Sum = Summer

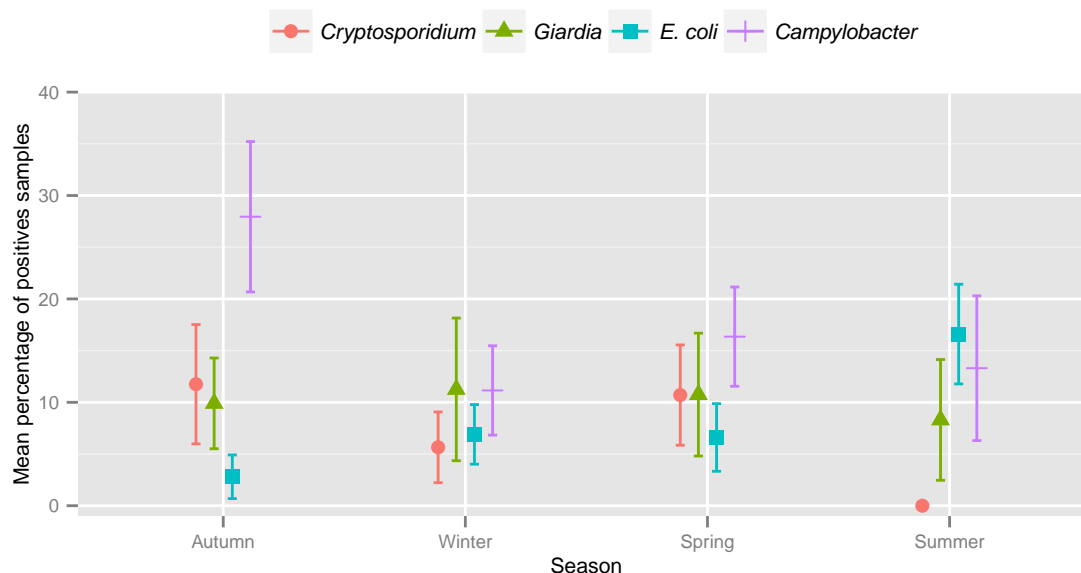
<sup>†</sup>Microbe: Campy = *Campylobacter*; Ecoli = *E. coli*; Crypto = *Cryptosporidium*; Giard = *Giardia*



**Figure 3.3:** Drinking water abstraction sites at which one or more samples with *E. coli* concentrations of 200 MPN or more per 100 mL were collected between September 2009 and March 2014, New Zealand.



**Figure 3.4:** Concentrations of *Cryptosporidium* and *Giardia* in samples collected from drinking water abstraction sites between September 2009 and March 2014, New Zealand. Sites from and months in which no sample was positive for either *Cryptosporidium* or *Giardia* are not shown here.



**Figure 3.5:** The site-adjusted mean percentage, with standard errors, of positive samples for the four study microbes in each season between September 2009 and March 2014, New Zealand. For *E. coli* the percentage was for samples that had a concentration greater or equal to 200 MPN per 100mL

Presented in Figure A.11 (page 185) are the percentages of positive samples by calendar month. April and May were the most likely time to find samples contaminated with *Campylobacter* while January was the most likely month to find samples with *E. coli* concentrations  $\geq 200$  MPN/100 mL. *Cryptosporidium*-positive ( $\geq 1$  oocysts/100 L) samples appeared to be bimodal with a higher peak occurring during September–November months and a lower peak around April. *Giardia*-positive ( $\geq 1$  cysts/100 L) samples appeared relatively constant throughout the year. Figure 3.5 shows the site-adjusted mean percentage of positive samples in each season. Autumn was the most likely season in which *Campylobacter* could be isolated from source water while isolation was less likely in winter. *E. coli* was most likely to be in concentrations  $\geq 200$  MPN/100 mL in source water during summer and less so in autumn. *Cryptosporidium* was least likely to be detected in summer while the detection of *Giardia* appeared to be relatively constant in all four seasons.

### 3.3.2 Random forest analysis

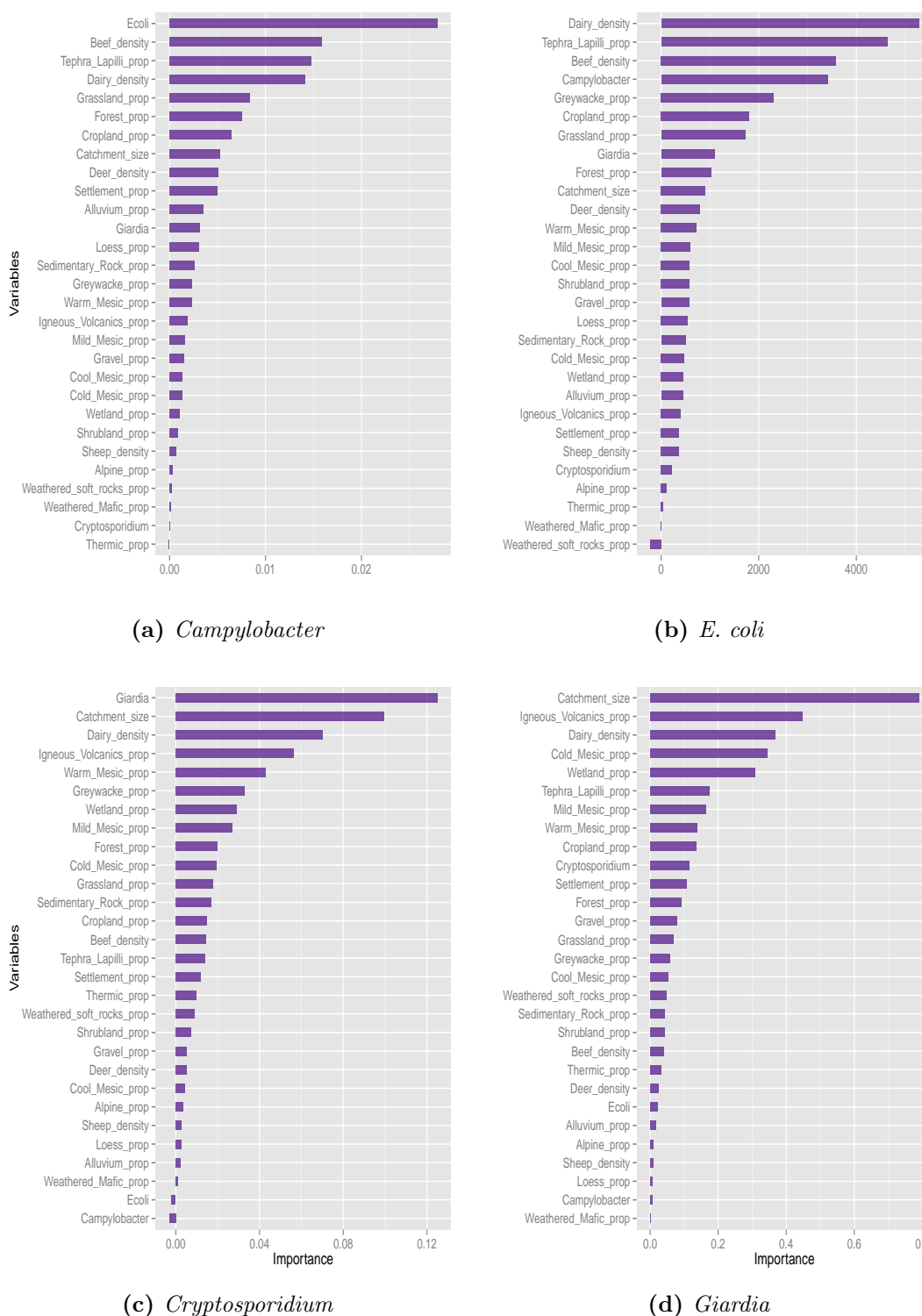
#### Variable importance

The five most influential variables for predicting the presence or absence of *Campylobacter* in source water included *E. coli* concentrations in source water, dairy cattle densities, beef cattle densities, tephra lapilli soil type and grassland (Figure 3.6a). Similarly, the five most influential variables for predicting *E. coli* concentrations in source water included dairy cattle densities, beef cattle densities, tephra lapilli soil type, presence/absence of *Campy-*

*lobacter* in source water and greywacke soil type (Figure 3.6b). Weathered mafic soil type and season, which had a negative importance scores, were irrelevant for the prediction of the concentrations of *E. coli* in source water. Concentrations of *Giardia*, catchment size, dairy cattle densities, igneous volcanic soil type and warm mesic soil temperature were among the variables with the five highest importance scores for predicting *Cryptosporidium* concentrations (Figure 3.6c). The two bacterial microbes (*Campylobacter* and *E. coli*) were irrelevant variables for predicting *Cryptosporidium* concentrations in source water. The five predictor variables with the highest importance scores for predicting *Giardia* concentrations included catchment size, igneous volcanic soil type, dairy cattle densities, cold mesic soil temperature and wetlands (Figure 3.6d). In general, the bacterial microbes shared similar predictor variables and were important predictors for each other. A similar pattern was observed for the protozoan microbes. The bacterial microbes were poor predictors of protozoan microbes and vice versa for the protozoan microbes. Season was irrelevant for the prediction of *E. coli* and *Giardia*.

### Predictions

Table 3.5 shows the RF predictions made for the levels of the four microbes (refer to Section 3.2.6 on page 56 for the description of the levels) in source water sampled during the January-March 2014 round. The actual microbe level recorded during the January-March 2014 round is denoted  $y$  while the level predicted by RF analysis is denoted  $\hat{y}$ . In addition, the estimated probability of a given sample having a microbe level  $1, \dots, 4$  is denoted  $\hat{p}_1, \dots, \hat{p}_4$ , respectively. These probabilities were the basis on which the predicted level of the microbe in a given sample was determined, i.e. the level with the highest estimated probability was deemed the predicted level. For example, Table 3.5a shows that the estimated probability of *Campylobacter* being absent (level 1) in a water sample from the Lake Karapiro (S00009) site was 0.88 and a probability of 0.12 was estimated for being present (level 2). Based on these probabilities, level 1 was selected as the predicted level of *Campylobacter* in that particular sample. Equivalently, Table 3.5b shows that the estimated probability of *E. coli* concentration being level 1 ( $< 100$  MPN/100 mL) in a sample from the Waikato River-Hamilton (S00041) site was 0.71 while the probabilities of 0.14, 0.09 and 0.06 were estimated for levels 2, 3 and 4, respectively. Level 1 having the highest estimated probability was selected as the most likely level of *E. coli* in the sample. Three out of sixteen (19 %) samples were misclassified for *Campylobacter* levels and one of the sixteen (6 %) samples was misclassified for *E. coli*. No samples were misclassified for *Cryptosporidium* and *Giardia* levels. The misclassified *Campylobacter* samples were those from the Turitea Dam (S00082) (level 1 predicted versus level 2 actual), Pareora River (S00200) (level 1 predicted versus level 2 actual) and Waikato River-Tuakau (S00865) (level 2 predicted versus level 1 actual) abstraction sites. The misclassified *E. coli* sample was collected the Waiorohi Stream (S00298) site which was predicted to have level 2 *E. coli* but had level 1.



**Figure 3.6:** Variable importance scores for the geospatial attributes of the drinking water catchments sampled between September 2009 and March 2014, New Zealand.

**Table 3.5:** Random Forest predicted levels of *Campylobacter*, *E. coli*, *Cryptosporidium* and *Giardia* in raw water collected in January-March 2014, New Zealand.  $y$  is the level of microbe recorded during the January-March 2014 sampling round while  $\hat{y}$  is the predicted level. Probabilities  $\hat{p}_1, \dots, \hat{p}_4$  indicate the likelihood of a given sample returning microbe level 1,  $\dots$ , level 4, respectively.

(a) <i>Campylobacter</i>					(b) <i>E. coli</i>						
Site	$y$	$\hat{y}$	$\hat{p}_1$	$\hat{p}_2$	Site	$y$	$\hat{y}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
S00092	Level 1	Level 1	0.88	0.12	S00092	Level 1	Level 1	0.99	0.00	0.00	0.01
S00099	Level 1	Level 1	0.94	0.06	S00099	Level 1	Level 1	0.99	0.00	0.01	0.00
S00865	Level 1	Level 2	0.46	0.54	S00865	Level 1	Level 1	0.66	0.17	0.10	0.07
S00121	Level 1	Level 1	0.93	0.07	S00121	Level 1	Level 1	0.99	0.01	0.00	0.00
S00434	Level 1	Level 1	0.97	0.03	S00434	Level 1	Level 1	0.99	0.01	0.00	0.00
S00120	Level 1	Level 1	0.89	0.11	S00120	Level 1	Level 1	1.00	0.00	0.00	0.00
S00118	Level 1	Level 1	0.94	0.06	S00118	Level 1	Level 1	1.00	0.00	0.00	0.00
S00124	Level 1	Level 1	0.99	0.01	S00124	Level 1	Level 1	0.92	0.04	0.00	0.04
S00200	Level 2	Level 1	0.76	0.24	S00200	Level 1	Level 1	0.90	0.03	0.05	0.02
S00041	Level 2	Level 2	0.45	0.55	S00041	Level 1	Level 1	0.71	0.14	0.09	0.06
S00298	Level 2	Level 2	0.23	0.77	S00298	Level 1	Level 2	0.25	0.35	0.11	0.28
S00299	Level 1	Level 1	0.95	0.05	S00299	Level 1	Level 1	0.86	0.02	0.09	0.03
S00009	Level 1	Level 1	0.89	0.11	S00009	Level 1	Level 1	0.96	0.00	0.03	0.00
S00383	Level 1	Level 1	0.95	0.05	S00383	Level 1	Level 1	1.00	0.00	0.00	0.00
S00088	Level 1	Level 1	0.62	0.38	S00088	Level 1	Level 1	0.85	0.01	0.05	0.09
S00082	Level 2	Level 1	0.89	0.11	S00082	Level 1	Level 1	0.94	0.03	0.02	0.01

(c) <i>Cryptosporidium</i>							(d) <i>Giardia</i>						
Site	$y$	$\hat{y}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	Site	$y$	$\hat{y}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$
S00092	Level 1	Level 1	0.99	0.00	0.01	0.00	S00092	Level 1	Level 1	1.00	0.00	0.00	0.00
S00099	Level 1	Level 1	1.00	0.00	0.00	0.00	S00099	Level 1	Level 1	1.00	0.00	0.00	0.00
S00865	Level 1	Level 1	0.90	0.06	0.02	0.02	S00865	Level 1	Level 1	0.53	0.32	0.02	0.13
S00121	Level 1	Level 1	1.00	0.00	0.00	0.00	S00121	Level 1	Level 1	1.00	0.00	0.00	0.00
S00434	Level 1	Level 1	1.00	0.00	0.00	0.00	S00434	Level 1	Level 1	1.00	0.00	0.00	0.00
S00120	Level 1	Level 1	1.00	0.00	0.00	0.00	S00120	Level 1	Level 1	1.00	0.00	0.00	0.00
S00118	Level 1	Level 1	1.00	0.00	0.00	0.00	S00118	Level 1	Level 1	1.00	0.00	0.00	0.00
S00124	Level 1	Level 1	0.97	0.03	0.00	0.00	S00124	Level 1	Level 1	1.00	0.00	0.00	0.00
S00200	Level 1	Level 1	1.00	0.00	0.00	0.00	S00200	Level 1	Level 1	1.00	0.00	0.00	0.00
S00041	Level 1	Level 1	0.97	0.02	0.00	0.00	S00041	Level 1	Level 1	0.57	0.32	0.09	0.02
S00298	Level 1	Level 1	0.99	0.01	0.00	0.00	S00298	Level 1	Level 1	0.99	0.01	0.00	0.00
S00299	Level 1	Level 1	1.00	0.00	0.00	0.00	S00299	Level 1	Level 1	1.00	0.00	0.00	0.00
S00009	Level 1	Level 1	1.00	0.00	0.00	0.00	S00009	Level 1	Level 1	1.00	0.00	0.00	0.00
S00383	Level 1	Level 1	1.00	0.00	0.00	0.00	S00383	Level 1	Level 1	1.00	0.00	0.00	0.00
S00088	Level 1	Level 1	0.96	0.04	0.00	0.00	S00088	Level 1	Level 1	0.98	0.02	0.00	0.00
S00082	Level 1	Level 1	1.00	0.00	0.00	0.00	S00082	Level 1	Level 1	1.00	0.00	0.00	0.00

### 3.3.3 Regression analysis

The GLMMs for the four microbes are presented in Tables 3.6–3.9. All the explanatory variables for the *Campylobacter* model were statistically non-significant (Table 3.6). However, the odds of a sample being *Campylobacter*-positive appeared to increase by 53 % if that sample had an *E. coli* concentration  $\geq 200$  MPN/100 mL compared to having a concentration  $< 200$  MPN/100 mL with beef and dairy cattle densities (count of animals per km<sup>2</sup>) held constant. In the *E. coli* model (Table 3.7) the cattle (beef and dairy) densities were marginally non-significant. However, the odds of a water sample having *E. coli* concentrations above the 200 MPN/100 mL threshold appeared to increase by 1 % for every unit increase in beef cattle density after adjusting for dairy cattle density and *Campylobacter* test result. A similar result was observed for every unit increase in dairy cattle density. The odds of a source water sample returning a *Cryptosporidium*-positive result ( $\geq 1$  oocysts/100 L)

increased by 3 % for every unit in dairy cattle density increase with catchment size (at the logarithmic scale) and *Giardia* test result held constant (Table 3.8). In the *Giardia* model (Table 3.9), both catchment size and dairy cattle density were marginally non-significant. However, a unit increase of catchment size at the logarithmic scale appeared to increase the odds of returning a positive *Giardia* test result after accounting for dairy cattle density.

The random effects, and their 95 % confidence intervals, for the four GLMMs are presented as caterpillar plots in Figure 3.7. Caterpillar plots which have random effects that are significantly different from zero (i.e. 95 % confidence intervals not crossing zero) suggest that there is some variation in the outcome variable between the sites not explained by the existing explanatory variables. For example, the caterpillar plot for the *Campylobacter* model (3.7a) has random effects for sites S00041, S00298 and S00124 being significantly different from zero, hence explanatory variables used in that model did explain all the outcome variation between sites.

**Table 3.6:** Generalised linear mixed model estimating the presence or absence of *Campylobacter* in raw water samples collected from sources supplying drinking water to the public in New Zealand, September 2009–March 2014.

Variable	Odds ratio (95% CI*)	P value	Observations
Intercept	0.11 (0.05; 0.25)	< 0.01	
<b><i>E. coli</i></b>			
<200 MPN/100 mL	1.00		222
≥200 MPN/100 mL	1.53 (0.59; 3.97)	0.39	66
<b>Ruminants in catchment</b>			
Beef cattle density	1.01 (0.99; 1.04)	0.24	288
Dairy cattle density	1.01 (0.99; 1.03)	0.41	288

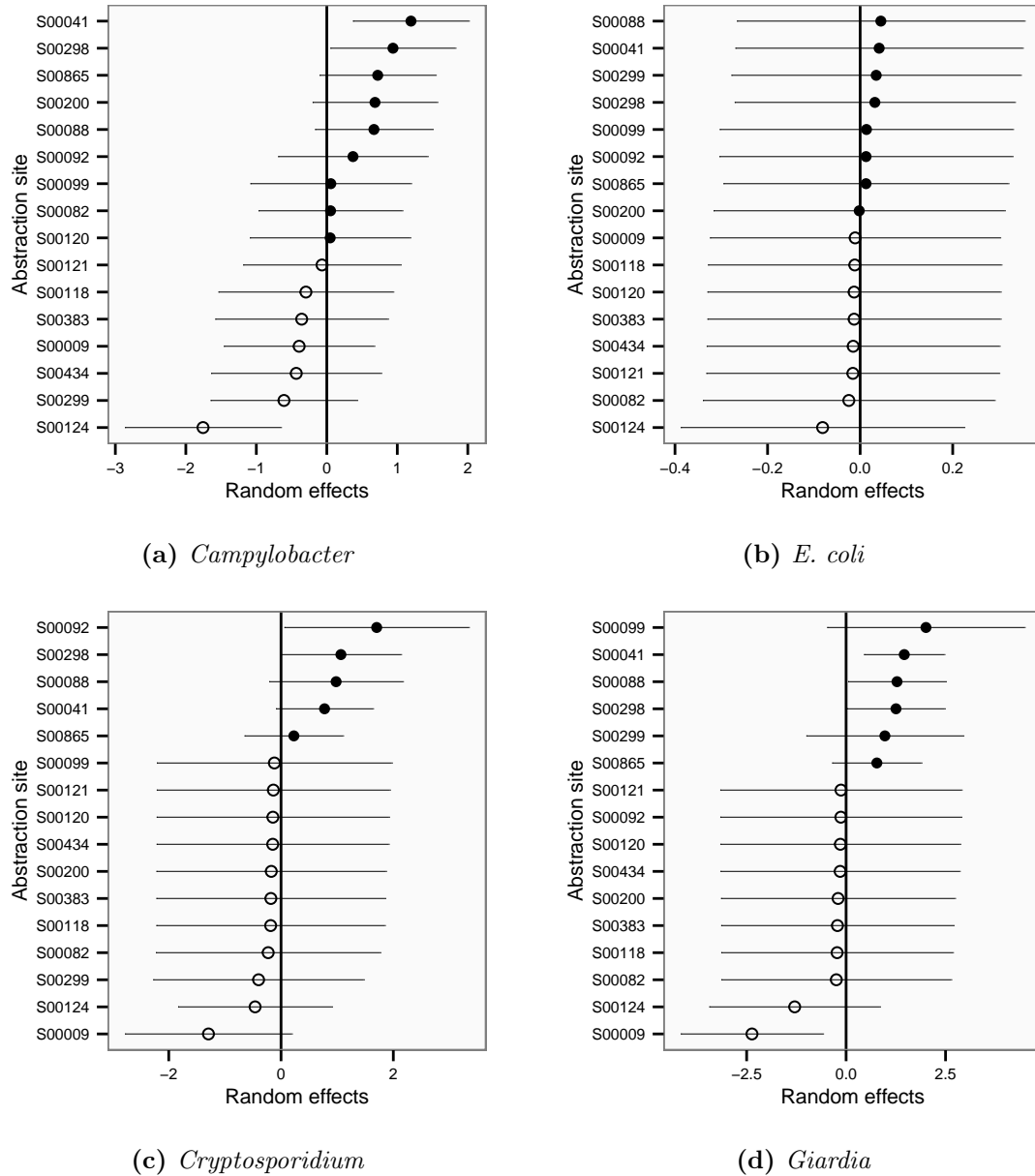
\*Confidence interval

**Table 3.7:** Generalised linear mixed model estimating the *E. coli* concentrations below or above 200 MPN per 100mL in raw water samples collected from sources supplying drinking water to the public in New Zealand, September 2009–March 2014.

Variable	Odds ratio (95% CI*)	P value	Observations
Intercept	0.02 (0.01; 0.06)	< 0.01	
<b><i>Campylobacter</i></b>			
Absent	1.00		260
Present	2.58 (0.79; 8.46)	0.12	28
<b>Ruminant</b>			
Beef cattle density	1.01 (1.00; 1.03)	0.06	288
Dairy cattle density	1.01 (1.00; 1.03)	0.08	288

\*Confidence interval





**Figure 3.7:** Random effects for the generalised linear mixed models estimating the concentrations of microbes in raw water samples collected from sources supplying drinking water to the public in New Zealand, September 2009–March 2014.

**Table 3.8:** Generalised linear mixed model estimating the presence or absence of *Cryptosporidium* in raw water samples collected from sources supplying drinking water to the public in New Zealand, September 2009–March 2014.

Variable	Odds ratio (95% CI*)	P value	Observations
Intercept	0.00 (0.00; 0.04)	< 0.01	
Catchment size (log scale)	1.28 (0.88; 1.85)	0.19	288
Dairy cattle density	1.03 (1.00; 1.06)	0.05	288
<b><i>Giardia</i></b>			
Absent	1.00		257
Present	2.32 (0.57; 9.35)	0.24	31

\*Confidence interval

**Table 3.9:** Generalised linear mixed model estimating the presence or absence of *Giardia* in raw water samples collected from sources supplying drinking water to the public in New Zealand, September 2009–March 2014.

Variable	Odds ratio (95% CI*)	P value	Observations
Intercept	0.00 (0.00; 0.01)	< 0.01	
Catchment size (log scale)	1.59 (0.98; 2.61)	0.06	288
Dairy cattle density	1.04 (1.00; 1.08)	0.07	288

\*Confidence interval

### 3.4 Discussion

The current study investigated risk factors associated with the presence of four microbes in raw water intended for treatment and supply to the public in New Zealand. The four microbes were those associated with microbial quality of drinking water that included two bacteria (*Campylobacter* and *E. coli*) and two protozoa (*Cryptosporidium* and *Giardia*). These microbes were monitored for  $4\frac{1}{2}$  years in samples collected from 20 public drinking water sources. The four microbes were chosen for use in the risk assessment for a number of reasons. Firstly, *E. coli* is an industry-standard indicator organism and is the recommended organism for water quality monitoring (Jose Figueras and Borrego, 2010; New Zealand Ministry of Health, 2008). Secondly, literature shows that these four microbes have previously been used for catchment risk assessment (Dechesne and Soyeux, 2007; Ferguson et al., 2007). Thirdly, the four microbes are among the leading cause of gastroenteritis in New Zealand (Environmental Science and Research, 2014).

The tools that were used to perform the risk assessment included an established non-parametric statistical method, random forest (RF) and a common statistical analysis framework — generalised linear modelling. RF analysis provided a framework for searching through numerous variables with possible complex interactions in order to identify a reduced number that had a positive relationship with levels of the four study microbes in raw water. Previous studies have used similar approaches for variable selection (Baca-Garcia et al., 2007; Dinsdale et al., 2013; Rossi et al., 2005). Dinsdale et al. (2013) used RF

analysis to identify functional hierarchies (genes and metabolic pathways), among a set of 27, that could be used to characterise 212 metagenomes in relation to their environments. Baca-Garcia et al. (2007) used RF and forward selection to search through 101 clinical variables in order to identify those that were associated with familial suicide attempts among 539 suicide attempters in Madrid, Spain. RF was among the tools used by Rossi et al. (2005) to identify variables characteristic of patients who had a once-only contact with the out-patient department of a Community Mental Health Service in South Verona, Italy. The study by Rossi and co-workers included nine categorical variables for 734 patients. In the present study, once a reduced number of variables had been identified using RF, the strength of association among them was estimated using generalised linear mixed model (GLMM)s. The latter were used in order to account for the hierarchical structure of the data i.e. repeat samples collected at each drinking water source.

Summarised microbial laboratory test results showed that surface water sources were more contaminated than groundwater sources. The surface water sources identified to be the most contaminated by all study pathogens were those on Oroua River (S00088), Waiorohi Stream (S00298) and Waikato River (S00041 and S00865). Given the heavy microbial load in raw water at these sites it implies that water treatment plants receiving water from these sources should be equipped with enhanced microbial removal capabilities in order to ensure drinking water microbial safety. However, the present study did not examine the microbial removal abilities of water treatment plants supplied by the study sources hence it is not possible to determine whether existing facilities are adequate or require improvements. Although surface water sources can be expected to be more contaminated than groundwater sources (Close et al., 2010), it has been shown that under certain conditions such as intensive dairy cattle farming and irrigation or use of unprotected shallow wells contamination can increase drastically (Close et al., 2008; Savill et al., 2001). Close et al. (2008) found that *E. coli* was detectable in three-quarters of groundwater samples in a border-strip irrigated dairy farm catchment in Canterbury, New Zealand. In the same catchment, about a tenth of the groundwater samples were positive for *Campylobacter*. In another study, Savill et al. (2001) reported detecting *Campylobacter* in more groundwater (shallow well) samples than surface water samples. These studies illustrate the importance of conducting a microbial risk assessment for each water supply as microbial loads vary from catchment to catchment. Similar observations have been previously been made by Dechesne and Soyeux (2007) who also advocated for implementation of local programmes when conducting risk assessment.

Over the study period, summarised monthly data and the site-adjusted seasonal mean percentages indicate that autumn (March to May) was the most likely period to find *Campylobacter* in raw water while summer, particularly during the month of January, was the most likely time to find *E. coli*. The RF variable importance analysis supported the seasonal effect for *Campylobacter* but not for *E. coli*. Previously, Till et al. (2008) reported similar

findings, i.e. *Campylobacter* was detected more frequently in summer-early autumn while the highest concentrations of *E. coli* were detected in summer and autumn at 25 freshwater recreational and water supply sites located throughout New Zealand. Donnison et al. (2004) reported finding higher concentrations of *E. coli* in surface water sources during summer and spring than during winter and autumn. The raw water contamination peak periods by *Campylobacter* and *E. coli* recorded in the present study do not coincide with the peak periods of campylobacteriosis and shiga toxin-producing *E. coli* (STEC)/verocytotoxin-producing *E. coli* (VTEC) infection in New Zealand. According to the annual surveillance report on notifiable diseases for the year 2013, cases of human campylobacteriosis in New Zealand tend to peak in autumn while reports of STEC/VTEC peak in autumn and spring (Institute of Environmental Science and Research Ltd, 2014). The difference between the temporal patterns of microbe detection in raw water and the temporal patterns of gastrointestinal illness in the population could be due to two main reasons. The first reason could be that water treatment acts as a modifier between the microbial concentrations in raw water and that of tap water (point of infection transmission). For example, the water treatment management could elect to deploy enhanced microbial removal measures during periods of anticipated high microbial load in raw water thereby distorting the temporal pattern of microbial concentrations between raw and tap water. The second reason could be that waterborne infection is not the main driver of gastrointestinal infection in the population as previous findings have shown that foodborne infection is strongly associated with reports of gastrointestinal illness in New Zealand (Eberhart-Phillips et al., 1997; Muellner et al., 2013).

*Cryptosporidium*-positive samples were likely to be detected in autumn (September in particular) and spring (March–May) while *Giardia*-positive samples did not exhibit a seasonal trend. The spring peak for *Cryptosporidium* contamination appears to coincide with the end of the calving season in New Zealand. Recent research has shown that *Cryptosporidium* in calves is prevalent in about a fifth of New Zealand dairy farms (Al Mawly et al., 2014). The temporal patterns of *Cryptosporidium* and *Giardia* contamination in raw drinking water coincide with the temporal patterns of cryptosporidiosis and giardiasis, respectively, in the New Zealand population. Cases of cryptosporidiosis tend to peak during spring and autumn while no seasonal patterns was recognisable for giardiasis (Institute of Environmental Science and Research Ltd, 2014). However, it is not immediately clear how much of the protozoan illness is attributable to waterborne infection.

RF analysis revealed that *Campylobacter* and *E. coli* shared similar predictor variables. In addition, the two bacterial microbes were good predictors of each other's levels in source water. A similar pattern was observed for the protozoan microbes. In general, bacterial microbes were poor predictors of protozoan microbes and vice versa. This suggests that as a microbial contamination indicator, *E. coli* can be expected to better indicate the presence of bacterial contamination than protozoan contamination. RF variable importance scores

also showed that cattle densities, especially dairy cattle density, were important predictors of all four microbes in raw water. Beef cattle density appeared to be a better predictor of levels of bacterial microbes than levels of protozoan microbes in raw water. The microbe contamination-cattle density relationship was supported by the regression analysis in the case of *Cryptosporidium* contamination-dairy cattle density relationship. This relationship was found to be statistically significant but of small magnitude. Although the other microbe contamination-cattle density relationships were statistically non-significant or marginally non-significant, their magnitudes were positive. This evidence implies that in order to reduce *Cryptosporidium* loads in raw drinking water, waterways in the catchment should be particularly protected from dairy cattle faecal contamination. Similar findings have previously shown that livestock farming is often a source of waterway contamination (Close et al., 2008).

The present study had both strengths and weaknesses. The strengths included the fact that a good number of public drinking water sources supplying a significant percentage of the New Zealand population were monitored for  $4\frac{1}{2}$  years. This length of time provided a framework for accounting temporal variations in raw water microbial contamination. Another strength was that the forecasting capabilities of RF analysis were demonstrated and this could be a useful tool for enhancing water treatment plant management and efficient use of resources. For example, adjustment to the water treatment regimes could incorporate forecasted microbial loads in raw water at any given time. Among the weaknesses of the study was the unsatisfactory performance of parametric models. Inclusion of all five variables identified by RF analysis as important predictors in a single GLMM resulted in either poor goodness of fit or model convergence failure. Part of the reason for this poor model performance could be that the complex interactions in the data could not be properly accounted for by the parametric regression modelling approach. Given these circumstances, it appears that the use of non-parametric methods such RF is a reasonable option although the magnitude of the outcome-explanatory variable relationship can then not be estimated. Another reason for poor parametric model performance could be that the explanatory variables used in the analyses were not good predictors hence addition of more variables only resulted in increased noise in the model estimation. This could mean that there are other unmeasured variables that are responsible for better explaining microbial contamination in raw water.

In summary, water sources that could require enhanced microbial removal capabilities were identified. It was also established that bacterial organisms were better predictors of each other just as protozoan organisms were. This implies that *E. coli* is not suitable for use as a universal microbial contamination indicator. Findings from RF analysis show that although season may be of some importance in predicting the presence/absence of *Campylobacter* in source water, it is not the case for predicting *E. coli* and *Giardia*. This implies that there

is no apparent benefit in targeting microbial removal from raw water at particular times of the year. RF and regression analyses indicate that it would be beneficial to reduce the densities of dairy cattle in drinking water catchments in order to reduce *Cryptosporidium* contamination in raw water.



# Four

## The relationship between river flow and notified cases of gastroenteritis in New Zealand

### 4.1 Background

The World Health Organization (2014b) ranked gastrointestinal illness fifth among the leading causes of years of life lost (YLL)<sup>1</sup> globally, causing around 1.5 million deaths<sup>2</sup> per year. Disease-specific morbidity incidence rates per 100 000 population for 2013 suggest that, in general, New Zealand has higher gastrointestinal illness rates compared to countries with a similar socioeconomic status (Table 4.1). Nevertheless, New Zealand reported the lowest rate for salmonellosis. The disease-specific statistics for New Zealand, Australia, England & Wales and United States of America (USA) were obtained from Environmental Science and Research (2014), Australia National Notifiable Disease Surveillance System (2014), United Kingdom Centre for Infectious Disease Surveillance and Control (2013) and Crim et al. (2014), respectively.

**Table 4.1:** Gastrointestinal illness annual incidence rates per 100 000 population for New Zealand and countries of similar socioeconomic status, 2013.

Country	Campylobacteriosis	Salmonellosis	Cryptosporidiosis	VTEC infection
New Zealand	157.1	13.1	23.4	4.5
Australia	93.5	55.3	16.6	0.8
England and Wales	114.6	14.7	10.0	1.4
United States of America	13.8	15.2	2.5	2.3

Drinking water is one of the routes through which gastroenteritis-causing pathogens are transmitted. Thus, accumulating evidence on mechanisms influencing the transmission of pathogens through this route is vital in an endeavour to finding better methods of controlling waterborne gastroenteritis. Numerous studies have previously investigated the relationship between the occurrence of gastrointestinal illness and climatic conditions, such as ambient temperature (Lopman et al., 2009; Onozuka and Hashizume, 2011; Onozuka et al., 2010) and rainfall (Auld et al., 2004; Curriero et al., 2001; Mackenzie et al., 1994; Nichols

<sup>1</sup>An indicator of premature mortality that is calculated by multiplying the number of deaths by the standard life expectancy. Greater weight is given to deaths at younger age than deaths at older age.

<sup>2</sup>[http://www.who.int/healthinfo/global\\_burden\\_disease/en/](http://www.who.int/healthinfo/global_burden_disease/en/) (accessed Jul 2014)



et al., 2009; Tornevi et al., 2014). However, few studies have examined such a relationship regarding river flow (Göransson et al., 2013; Jagai et al., 2012), a factor closely related to the climatic elements: rainfall, runoff, river flow and turbidity.

Online scientific databases, PubMed and Web of Science, were extensively queried for literature on the relationship between river flow and gastrointestinal illness. The queries were conducted in July 2014 using electronic resources available through the Massey University library. The final search term<sup>3</sup> included terms and word combinations related to waterborne gastroenteritis-causing microbes and river flow. A total of 118 articles were retrieved but only five (Beaudeau et al., 2014; Greer et al., 2009; Jagai et al., 2012; Lake et al., 2005; Thomas et al., 2006) reported investigating a river flow-gastroenteritis relationship. The majority of the retrieved articles reported associations between river flow and the presence or concentrations of microbes in rivers.

Thomas et al. (2006) conducted a study aimed at describing the incidence and distribution of waterborne disease outbreaks in Canada with regards to preceding weather conditions. The association between high impact weather events and waterborne disease outbreaks was also investigated. Warm weather and rainfall were found to be risk factors for waterborne disease outbreaks. No association between river flow and occurrence of the outbreaks was found and the significant amount of missing river flow data was thought to have adversely affected the study power in detecting any existing river flow-waterborne disease outbreak association. A study conducted by Jagai et al. (2012) examined the relationship between hospitalisations due to gastrointestinal illnesses and river flow. The study was conducted among the elderly living along the Ohio River, USA. A positive correlation between river flow and gastrointestinal illness among persons aged 65 years or older was found, however, the peak period for illness preceded that of river flow.

Greer et al. (2009) investigated environmental factors associated with *Norovirus* outbreaks in Toronto, Canada. Two hundred and fifty-three outbreaks of gastroenteritis linked to *Norovirus* were used to examine the relationship between acute changes in environmental factors and the risk of disease outbreaks. Low temperatures ( $\leq 4^{\circ}\text{C}$ ) on Lake Ontario and high flow in the Don River occurring 1–7 days prior were found to be risk factors. In another study, Lake et al. (2005) investigated the effect of precipitation, temperature and river flow on the rates of cryptosporidiosis in England and Wales. Cryptosporidiosis rates were positively correlated with current month's maximum river flow from April to July. Between August and November this positive relationship was only present after accounting for the previous month's temperature and precipitation. The effect of treated water turbidity and

---

<sup>3</sup>(gastroenter\* OR gastrointest\* OR enteri\* OR diarrh\* OR campylobact\* OR choler\* OR cryptosporid\* OR escherichia OR 'E. coli' OR enterovir\* OR giard\* OR legionell\* OR leptospir\* OR rotavir\* OR salmonell\* OR shigell\* OR norovirus) AND ('river flow' OR river-flow OR streamflow OR 'stream flow') AND ('drinking water' OR drinking-water OR water-borne OR 'water borne')

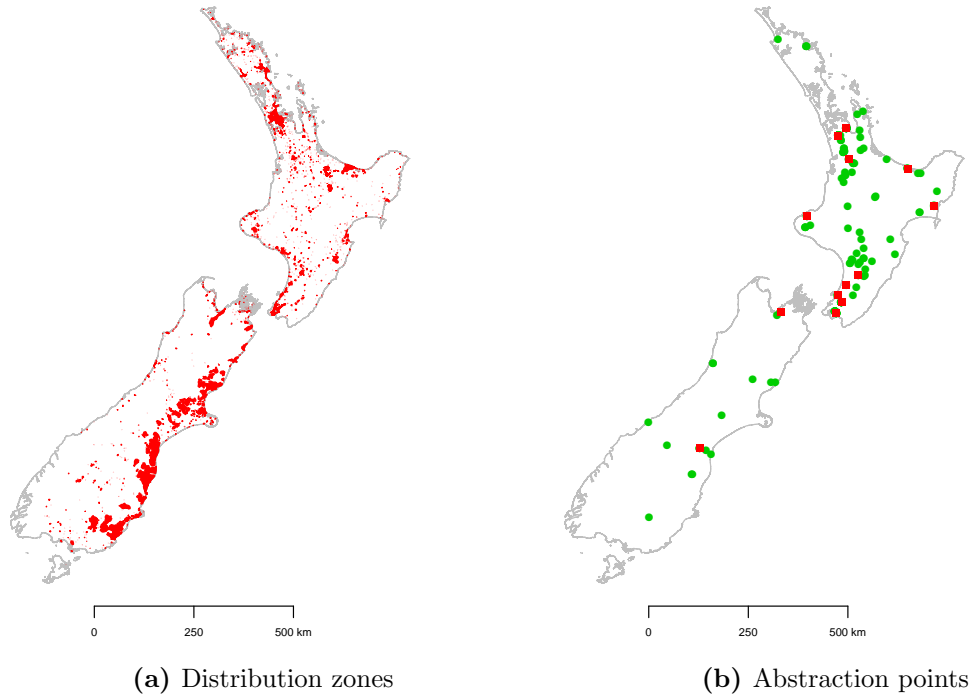
river flow on the incidence of acute gastroenteritis in the Nantes area, France, was examined by Beaudeau et al. (2014). High values of turbidity and river flow were associated with higher endemic levels of gastroenteritis.

These five studies show that increased river flow can be positively associated with the occurrence of gastrointestinal illnesses. Although increased river flow is usually a result of increased rainfall, other factors such as snowmelt can also influence the rate of flow. Further, Göransson et al. (2013) reported that the relationship between rainfall, turbidity and river flow can be complex especially in the presence of human activity on the affected river. This shows that although rainfall and river flow may be correlated, the rainfall-gastrointestinal illness relationship may not be equated to that of river flow-gastrointestinal illness. To the best of our knowledge this phenomenon has not previously been investigated in New Zealand. As part of an effort to fill this knowledge gap, the current study investigated how river flow rates varied with reports of gastrointestinal illness in New Zealand. The objective of the study was to quantify the association between river flow and reported cases of waterborne diseases in New Zealand for the ten-year period 1997-2006. The null hypothesis was that ‘River flow on drinking water source rivers is not associated with gastrointestinal illness reports in the local communities’ while the alternative hypothesis stated that ‘River flow on drinking water source rivers is associated with gastrointestinal illness reports in the local communities.’

## 4.2 Materials and methods

### 4.2.1 Study units

This was a retrospective longitudinal study and the primary unit of analysis was the drinking water distribution zone (Figure 4.1a). The drinking water supply system in New Zealand is described in Section 2.2.2 on page 16. Briefly, drinking water is supplied through a network that can be divided into three main parts: the source (intake), treatment plant and distribution zone. In a simple standard distribution network, water is abstracted at the source, treated at a treatment plant and delivered to consumers through a distribution zone. However, in cities and large communities the water supply networks are intricate and involve multiple sources, treatment plants and distribution zones. Conversely, in small private supplies, treatment may be lacking. In such a scenario water is abstracted from a source such as roof or borehole and delivered directly to the point of consumption e.g. a household. The water sources in the present study can be divided into three main groups: ground (borehole, spring and well), roof and surface (canal, creek, dam, gully, lake, river and stream). The location of drinking water sources sited on rivers which had river flow recordings are shown in Figure 4.1b.



**Figure 4.1:** Figure (a) shows the 2001 drinking water distribution zones while (b) shows point-location of abstraction points on rivers with flow recordings in New Zealand, 1997–2006. The square points in (b) had 300 or more gastrointestinal illness cases during the study period.

#### 4.2.2 Data

Four main datasets were used in the current study. The first dataset was that of daily river flow rates from recording sites on rivers within New Zealand, monitored by various recording agencies that included the National Institute of Water and Atmospheric Research (NIWA) and eighteen regional authorities. The second dataset comprised climatic data i.e. daily ambient temperature and daily rainfall recordings, freely available on a website<sup>4</sup> maintained by NIWA. The third dataset was composed of disease cases caused by pathogens associated with drinking water extracted from the New Zealand national notifiable disease surveillance (EpiSurv) database. The disease cases were those caused by *Campylobacter*, *Cryptosporidium*, *Giardia* and *Salmonella*. The fourth dataset was the 2001 Register of Community Drinking-water Supplies in New Zealand. The 2001 register was used as it represented the mid-point for the study period. The third and fourth datasets were provided by the Institute of Environmental Science and Research Limited (ESR) which maintains the two data systems on behalf of the New Zealand Ministry of Health (MoH). The four datasets were amalgamated into a relational database using the software MySQL (DuBois, 2008). A schematic representation of the relationships among tables in the database are shown in Figure A.1 on page 172.

---

<sup>4</sup><http://cliflo.niwa.co.nz/>

### 4.2.3 Multivariate data analysis

The relationship involving disease incidence, types of water sources and rurality of the water source were investigated using a multivariate approach. In this approach prior categorisation of variables as response (dependent) and explanatory (independent) is not necessary, instead, the relationships among all variables in the dataset are simultaneously assessed. There are many different types of multivariate analyses e.g. canonical correspondence analysis (CCA), factor analysis, linear discriminant analysis (LDA), multivariate analysis of variance (MANOVA), multivariate regression analysis and principal component analysis (PCA). The current study used PCA for data analysis, hence, it is discussed here.

PCA is a multivariate technique that reveals the internal structure of data in a way that best explains the variance in those data. Assuming that a given dataset is a cloud of data points, PCA rotates it such that the maximum variability is visible. In this way patterns in data can be identified i.e. their similarities and differences can be highlighted. This is achieved by transforming the dataset variables into uncorrelated variables called principal components. The first principal component is the line that passes through the data cloud centroid and minimises the square of the distance to each point. Thus, in some sense, the line is as close to all of the data as possible. Equivalently, the first principal component accounts for much of the variation in the data. Subsequent principal components progressively account for as much of the remaining variability as possible. This is in such a way that each subsequent principal component passes through the data centroid and is orthogonal (i.e. at right angles or uncorrelated) to the preceding component(s). For instance, the second principal component not only passes through the centroid but is also orthogonal to the first principal component having gone through the maximum of the remaining variation in the data (Jolliffe, 2002; Jolliffe and Morgan, 1992).

Mathematically, PCA utilises matrix algebra to transform a dataset of observations into eigenvalues and eigenvectors. The dataset is regarded as a matrix in which the variables are the columns and the observations (elements) are the rows. From such a matrix another matrix that contains the variance of each variable (as diagonal terms) and covariance of each pair of variables (as off-diagonal terms) is obtained. This new matrix is called the *variance-covariance matrix* or simply the *covariance matrix*. The eigenvalues and eigenvectors are then derived from the covariance matrix. The elements of the matrix of eigenvectors are called the *principal component loadings*. Since the diagonal terms of the covariance matrix are the variable variance values, their sum is equal to the total variance, which is also equal to the sum of eigenvalues. In other words, each eigenvalue is a fraction of the total variance. Multiplication of the original observation matrix with that of eigenvectors yields a matrix of *principal component scores* (Davis and Sampson, 2002). Further, the scores are mea-

sured along axes that are perpendicular to each other. This means that the covariance (and the correlation) between scores is zero, equivalent to saying that the scores are uncorrelated.

PCA is sensitive to the magnitudes of observation measurements, especially if they are not all on the same scale. This difficulty can be overcome by standardising all variables so that they have a mean of zero and a variance of one. The PCA output can be visualised on a graphical representation called a *biplot*. A biplot not only shows variances and correlations of the variables but can also reveal inter-observation distances and clustering of observations (Gabriel, 1971). The prefix *bi* in the name biplot is derived from the fact that it is capable of simultaneously displaying both rows and columns of the transformed data matrix. In a PCA biplot the principal component scores are used as coordinates for the original observations and points in close proximity are similar (Acevedo, 2013).

The descriptions of the variables used in the PCA are given in Table 4.2. Also shown in Table 4.2 is the summary of the raw data on which the PCA was based: the cumulative number of gastrointestinal illness cases over the study period and the underlying populations that were based on the 2001 water distribution zone estimates. The location of a community served by a given water source was classified depending on how urban or rural it was. This was termed *rurality* and was categorised into five groups, with *Rural1* being less rural and *Rural4* highly rural; *Urban1* was less urban while *Urban3* was highly urban. Water source were categorised into six types with *multisource* indicating that a given community was served by more than one type of water sources. Thus three variables were included in the PCA: gastrointestinal illness-specific incidence rates, rurality and water source type. To perform the PCA, annual incidence rates for each zone, stratified by gastrointestinal illness, were first calculated. The median incidence rates were then calculated for both the rurality and water source variables. The data were standardised using the `scale` function in **R** (R Core Team, 2013). The PCA was performed in **R** and results displayed using a biplot.

#### 4.2.4 Geostatistical exploration

The geographical spread of the gastroenteritis cases reported between 1997 and 2006 in New Zealand was examined using geostatistical tools. In geostatistics, a variable that is spatially distributed is said to be regionalised, e.g. rainfall and elevation data. The properties of a regionalised variable are in-between those of a truly random variable and those of a completely deterministic one. Variation of such a variable may occur both in time and space, for instance, rainfall recorded at different times at various weather stations. The variable is assumed to be continuous over a spatial domain but varies in a complex manner that cannot be described using a deterministic function (Davis and Sampson, 2002). A regionalised variable can be referred to as stationary, in which case two assumptions are made. The first assumption is that the mean of the variable is constant from location to

location. The second assumption is that the covariance of the variable between different locations depends on the separation distance (also known as *lag* in geostatistics), and not on their absolute location (Webster and Oliver, 2008).

One major task in geostatistics is to measure spatial dependency or the degree of relatedness between values at different locations. Similarity between values at locations separated by a specified lag is measured by covariance while semivariance measures their dissimilarity. Covariance decreases as lag increases while semivariance increases until it plateaus when no additional differences are gained with the increased lag (Acevedo, 2013). Semivariance is a basic utility in geostatistics for visualising, interpreting, modelling and exploiting dependency in regionalised variables (Şen, 2009). Semivariance informs spatial models such as kriging in order to estimate (or predict) values at locations where data are not available. The terms *spatial (geostatistical) prediction*<sup>5</sup> and *spatial (geostatistical) estimation* are used synonymously here although statisticians prefer using the term prediction while geostatisticians tend to use estimation (Webster and Oliver, 2008).

### Kriging technique overview

Kriging was used for interpolation in the current study. This technique has for many decades been synonymous with geostatistical interpolation. The origins of kriging can be traced to the mining industry in the early 1950s where it was used to improve the estimation of minerals in ore reserves. The original idea is credited to the mining engineer D. G. Krige and the statistician H. S. Sichel. The name of the technique is in acknowledgement of Krige who pioneered its use. However, a French mathematician Matheron and his co-workers formalised the technique, establishing the basis for linear geostatistics (Cressie, 1990). Kriging may be referred to as a technique used for predicting values of regionalised variables in non-sampled points of a spatial domain using a collection of sampled points (Acevedo, 2013). This set of sampled points forms a marked point pattern, which can be regular or irregular. This is helpful for making generalisations because variables are typically measured at limited number of locations due to resource constraints.

Over the decades many variants of kriging have been developed and the major types include:

- *Simple kriging*: A critical assumption under this type of kriging is that the mean of the regionalised variable is known. Unfortunately, the mean is often unknown, thus, the application of this technique is severely limited (Şen, 2009). Its usage is usually in other types of kriging that utilise transformed data and have known means such as indicator and disjunctive kriging (Webster and Oliver, 2008).

<sup>5</sup>Spatial prediction (estimation) can be either interpolation or extrapolation (Hengl, 2009). Interpolation predicts values of the dependent variable within the data range while values outside the data range are predicted by extrapolation. Interpolation is more likely to produce valid estimates than extrapolation.

- *Ordinary kriging*: This is similar to simple kriging except that it assumes that the mean is regionally constant but unknown (Şen, 2009). It is a very robust technique that models a single variable and is commonly used (Webster and Oliver, 2008). This method was used in the present study and is described further later.
- *Universal kriging*: This method is also known as *kriging with drift*. It is an extension of ordinary kriging in which the mean is allowed to vary regionally. It considers a first-order non-stationary regionalised variable to be composed of two components. The first component is *drift*, the average (expected value) of a regionalised variable within a neighbourhood. This component is the non-stationary, slowly changing part of the variable. The second component are *residuals*, the difference between the actual observations and the drift. If drift is removed from the regionalised variable, the remainder are residuals which are stationary and can be modelled using ordinary kriging. In principle, universal kriging is a three-step procedure: Drift is estimated and removed from the regionalised variable; ordinary kriging is applied to obtain residuals at non-sampled locations; and the residuals are added to the original drift values (Şen, 2009). However, this three-step process can be avoided by incorporating Lagrangian multipliers<sup>6</sup>(Davis and Sampson, 2002).
- *Factorial kriging*: This is also known as *kriging analysis* and is intended to explain the spatial variability among multilevel coregionalised variables. Factorial kriging allows the decomposition of the different components that may be mapped separately, but in a single analysis (Goovaerts, 1992; Webster and Oliver, 2008).
- *Ordinary cokriging*: It is similar to ordinary kriging except that it models two or more variables that are coregionalised. The technique is beneficial in cases where one variable can be measured cheaply (or easily) at many locations while (an)other spatially correlated variable(s) can only be measured at limited locations due to resource constraints or other reasons. Estimates of the more sparsely variable can be derived using the more intensely measured one (Webster and Oliver, 2008).
- *Indicator kriging*: This type of kriging employs non-linear, non-parametric methods. It involves conversion of continuous variables to binary ones (indicators), hence can deal with many kinds of distributions. The binary categories are user-defined thresholds for which probabilities can be calculated, indicating the likelihood of a variable value being above or below. *Soft* qualitative data can also be incorporated to improve predictions (Li and Heap, 2008; Webster and Oliver, 2008).
- *Disjunctive kriging*: This is a non-linear but parametric method of kriging. It is useful in cases where conventional transformations (such as logarithm or square-rooting) of a regionalised variable does not result in normal (or near-normal) distribution. In addition to kriging estimates, it is also used for decision-making because thresholds can

---

<sup>6</sup>A strategy for finding the local maxima and minima of a function subject to equality constraints



be defined by the user then probabilities of exceeding or not exceeding the threshold determined (Webster and Oliver, 2008).

- *Bayesian kriging*: This is intermediate between simple and universal kriging. It utilises simple kriging equations but with non-stationary covariance. It takes into account the uncertainty about the prediction parameters (Omre, 1987; Omre and Halvorsen, 1989).

### Ordinary kriging

This method is sometimes referred to as best linear unbiased estimator (BLUE). *Linear* because its estimates are weighted linear combinations of the available data; *unbiased* since it tries to equate the mean residual error to 0; *best* because it aims at minimising the variance of the errors (Isaaks and Srivastava 1989).

Using geostatistical notations, ordinary kriging is said to interpolate values of a regionalised variable  $Z$ . The elements of  $Z$  can be denoted as  $z(\mathbf{s}_1), z(\mathbf{s}_2), \dots, z(\mathbf{s}_n)$ , where  $\mathbf{s}_i = (x_i, y_i)$  is a location and  $x_i$  and  $y_i$  are the coordinates in a spatial domain while  $n$  is the number of observations.  $Z$  is assumed to be stationary, meaning that its mean and variance do not change with the location  $x, y$ . The formula for predicting values  $\hat{z}$  at a new location  $\mathbf{s}_0$  is described in Equation 4.1 (Hengl, 2009):

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^n w_i(\mathbf{s}_0) \cdot z(\mathbf{s}_i) \quad (4.1)$$

where  $w_i$  are the kriging (prediction) weights and  $z(\mathbf{s}_i)$  are the values of a regionalised variable (input point data). The semivariances i.e. the differences between the neighboring values are calculated using Equation 4.2:

$$\gamma(\mathbf{h}) = \frac{1}{2} \mathbb{E} \left[ (z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h}))^2 \right] \quad (4.2)$$

where  $z(\mathbf{s}_i)$  is the value of a regionalised variable (input point data) and  $z(\mathbf{s}_i + \mathbf{h})$  is the value of the neighbour at distance  $\mathbf{s}_i + \mathbf{h}$ . Note that the symbol  $\mathbb{E}$  in Equation 4.2 is called *expected value*.

*Kriging weights*: In ordinary kriging, the distance from the prediction point is an important factor for determining the prediction weight at a given point. In determining weights, values in a cluster are assigned reduced weights than isolated values, a process known as *de-clustering*. This is on the premise that closely located values provide little more information compared to one that is isolated. Another feature of kriging is called *masking*, where values at nearby locations are assigned larger weights compared to those further away (Diggle and Ribeiro, 2007).



In the current study, geostatistical exploration of the disease case data was conducted in order to determine the spatial trend of the reported cases of gastroenteric illness over the study period (1997–2006). The data used for this exploration are those described in Tables 4.2 and 4.3 but without stratifying by rurality and water source variables. Annual incidence rates for water distribution zones that had global positioning system (GPS) coordinates formed the regionalised variable and the zone centroids were the point-locations. The overall spatial variation for the study period was estimated using the median annual incidence rates while respective annual incidence rates were used for the yearly interpolations. The initial step was to divide the map of New Zealand into regular eight square kilometre grids. Ordinary kriging was then applied in order to interpolate the incidence rate for each grid given the regionalised variable (zonal incidence rates). The interpolation results were then visualised on a smoothed surface map. In addition, the raw regionalised variable was plotted on a bubbleplot which showed the magnitude of the incidence rates at a given location. The size of the circles in the bubbleplot correspond to the incidence rates. Spatial interpolation was implemented using the **R** package `automap` (Hiemstra et al., 2008).

#### 4.2.5 Statistical modelling

The main task in the present study was to quantify the relationship between flow on rivers used as sources for public drinking water and the number of gastrointestinal illness cases reported in the local community. To perform this task, river flow was considered as the risk factor (explanatory variable) while counts of gastrointestinal disease cases formed the outcome (response) variable. It was assumed that the outcome was not observed on the same day that exposure to the risk occurred, but at a later date. Models that account for such a time delay (lag-based models) were thus employed in the analysis.

##### The lag model concept

Lag models are designed to quantify the association between an exposure occurrence and an outcome characterised by a lapse in time. The period between the time of exposure occurrence and outcome observation is the *lag* ( $l$ ). In a prospective description of such an association, an exposure occurring at time  $t$  determines the outcome at a later time  $t + l$ . Conversely, in a retrospective sense, the outcome at time  $t$  is determined by an exposure that occurred at a previous time  $t - l$ .

Using the prospective description, in a simple lag model the assumption is that the effect of a unit increase of an exposure is observed only at a single time point. Implementation of such a model is limited because the lag is often not known. To overcome this constraint a distributed lag model (DLM), in which multiple lags of exposure (also known as *exposure history*) are included simultaneously, can be used. Concretely, in a DLM the effect of the exposure is assumed not to occur instantaneously but spread over time, or *distributed*.

Another assumption in a DLM is that the relationship between exposure and outcome is linear. However, when the exposure-outcome relationship is non-linear a DLM is deficient, the alternative is a distributed lag non-linear model (DLNM) — an extension of DLM.

DLMs have been described by many previous studies (Almon, 1965; Goodman et al., 2004; Lütkepohl, 1981; Zanobetti et al., 2000), however, the current study adopted the approach and notations published by Gasparrini in a series of articles (Gasparrini, 2014; Gasparrini and Armstrong, 2013; Gasparrini et al., 2010). Further, the **R** package `dlm` (Gasparrini, 2011) was used for the implementation of the models. For the sake of clarity, two terms that are fundamental to understanding this topic are defined here. The first term is *vector space*: This is a collection of vectors which is closed under the operation of addition and multiplication by a scalar. This means that if  $\mathbf{a}$  and  $\mathbf{b}$  are in the collection, the sum  $\mathbf{a} + \mathbf{b}$  is also in the collection, and if  $\mathbf{a}$  is in the collection,  $\lambda\mathbf{a}$  is in the collection for any scalar  $\lambda$  (Hadley, 1988; Rose, 2002). For example, a *real vector space* involves the field  $\mathbb{R}$  of real numbers (also called scalars) and a non-empty set of  $\mathbf{V}$  whose elements (called vectors) can be of very different nature (e.g. geometric vector-arrows, matrix-columns, other kinds of matrices, polynomials, all sorts of functions) (Vujicic and Sanderson, 2008). The second term is *basis*: This is a finite set of vectors which span a given linear space (e.g. vector space) and are linearly independent (Lax, 2007). A vector space is called linearly dependent if it contains a finite number of distinct vectors  $v_1, v_2, \dots, v_n$  and scalars  $a_1, a_2, \dots, a_n$ , not all zero, such that  $a_1v_1 + a_2v_2 + \dots + a_nv_n = 0$  where 0 is the zero vector not the number zero.

In algebraic terms, a basic lag model involves the selection of a basis from which a *basis function* is chosen to describe the relationship between an exposure and outcome. Such a function can be expressed as:

$$s(x_t; \boldsymbol{\beta}) = \mathbf{z}_t^T \boldsymbol{\beta} \quad (4.3)$$

where  $s(\cdot)$  is a basis function applied to the original vector of exposures  $\mathbf{x}$ , of which  $x_t$  is the  $t$ th row;  $\boldsymbol{\beta}$  is a vector of unknown parameters;  $\mathbf{z}_t$  is the  $t$ th row of the  $n \times v_x$  (i.e. the number of observations by the number of variables) basis matrix  $\mathbf{Z}$ . The latter is obtained once the basis function has been applied to  $\mathbf{x}$ .

In contrast, DLMs involve the transformation of the vector  $\mathbf{x}$  into a matrix of lags  $\mathbf{Q}$ , with dimensions  $n \times (L + 1)$ , such that  $\mathbf{q}_t = [x_t, \dots, x_{t-L}]^T$ , where  $L$  is the maximum lag. A basis function can be applied to  $\mathbf{Q}$  in a similar way as described for a basic lag model:

$$s(x_t; \boldsymbol{\eta}) = \mathbf{q}_t^T \mathbf{C} \boldsymbol{\eta} \quad (4.4)$$

where  $\mathbf{C}$  is a matrix of *basis variables*, of dimensions  $(L + 1) \times v_l$ , and  $\boldsymbol{\eta}$  is a vector of unknown parameters. Different basis functions can be applied to obtain different forms of DLM e.g if  $\mathbf{C} \equiv 1$  the DLM is that of a moving average and smoothed curve-DLMs are specified when  $\mathbf{C}$  is a polynomial or spline function.

To extend a DLM to a DLNM a *cross-basis* is used. Note that in a simple lag model a basis function is applied to a vector of explanatory variable values while in a DLM a basis function is applied to a vector of lagged exposure values. The concept of cross-basis involves simultaneous application of these two basis functions, collectively termed *cross-basis function* and can be represented as:

$$s(x_t; \boldsymbol{\eta}) = \sum_{j=1}^{v_x} \sum_{k=1}^{v_l} \mathbf{r}_{tj}^T \cdot \mathbf{c}_{\cdot k} \eta_{jk} = \mathbf{w}_t^T \boldsymbol{\eta} \quad (4.5)$$

where  $\mathbf{r}_{tj}$  is the vector of lagged exposures for the time  $t$  transformed through the basis function  $j$ . The vector  $\mathbf{w}_t$  is obtained by applying the  $v_x \cdot v_l$  cross-basis functions to  $x_t$ , similar to Equation 4.4.

The  $\mathbf{Z}$  matrix obtained in Equation 4.3, a matrix  $\mathbf{W} = \mathbf{QC}$  derived from Equation 4.4 and  $\mathbf{W}$  obtained from Equation 4.5 as a tensor product can be included in a design matrix of a model (e.g. generalised linear model) in order to estimate the related unknown parameters  $\boldsymbol{\beta}$  and  $\boldsymbol{\eta}$ .

### Distributed lag non-linear model implementation

In the present study, the exposure-response relationship was assumed to be a lagged non-linear one hence it was investigated using a DLNM. River flow was standardised by converting flow rate ( $\text{m}^3/\text{s}$ ) into percentile over the study period. This was done to make flow comparable between large and small rivers. A cross-basis function was used to produce a cross-basis matrix for river flow. In the cross-basis function a B-spline function centred at the 50th flow percentile was chosen for creating a basis for river flow while a 4th degree polynomial basis function was chosen for creating a basis for the lags. River flow was lagged up to 50 days. Centering river flow at the 50th percentile allowed computation of the predicted effects to be compared to that of the 50th flow percentile. The river flow cross-basis matrix was introduced in a quasi-Poisson generalised linear model for the estimation of coefficients. A quasi-Poisson model was used in order to adjust for overdispersion as Poisson models tend to be overdispersed (Gelman and Hill, 2007). Other variables considered for inclusion in the generalised linear model included ambient temperature, rainfall, rainfall deficit (drought) and their respective lags. Month and abstraction site were also considered. The Akaike information criterion (AIC) was used for model selection with river flow, month and sites regarded as study factors. Based on AIC, ambient temperature, rainfall, drought

and their lags were not retained in the model. The explanatory variables retained in the model were river flow cross-basis matrix, month and site. Month was included in order to account for seasonality in the data. The outcome variable was the number of reported cases of gastroenteritis on each day over the ten-year study period in the population supplied by a given water source.

A summary of the data from eleven drinking water abstraction sites (Figure A.10) on rivers where flow was monitored and were available for the lag analysis is presented in Table 4.4. A DLNM and subsequently a quasi-Poisson model were fitted to the data from the eleven abstraction sites as described above. In addition, separate models for each abstraction site were similarly implemented but without sites as a covariable in the generalised linear model.

## 4.3 Results

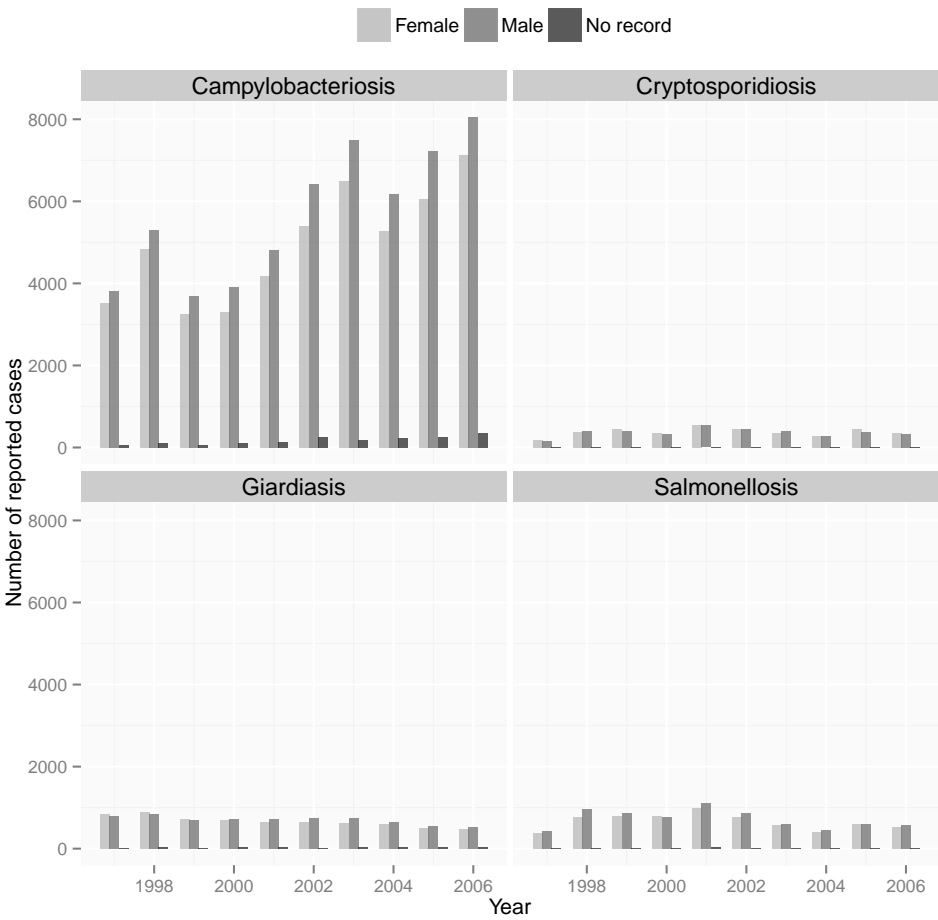
### 4.3.1 Descriptive statistics

In 2001 there were 1891 registered communities receiving drinking water through 2111 distribution zones. The zones were supplied with water from 2035 treatment plants which sourced water from 2338 abstraction points. A total population of 3.4 million was serviced, representing 88.1 % of the New Zealand population<sup>7</sup>.

During the study period, 1997–2006, a total of 143 455 cases of gastroenteritis caused by four pathogens associated with drinking water (*Campylobacter*, *Cryptosporidium*, *Giardia* and *Salmonella*) were recorded in the New Zealand national notifiable disease database, EpiSurv. Figure 4.2 shows the reported cases stratified by disease and gender. Campylobacteriosis was the most reported of the four gastrointestinal illnesses, comprising 75.4 % (108 103) of the reported cases, followed by salmonellosis 9.7 % (13 953), giardiasis 9.7 % (13 852) and cryptosporidiosis 5.3 % (7547). The number of reported cases of campylobacteriosis per year were steadily increasing in contrast to those of the other three diseases which remained relatively constant. In terms of gender, 52.1 % (74 788) of the reported cases were male, 46.3 % (66 400) were female and no gender status was recorded for 1.6 % (2267) of the cases.

Excluding zones that did not have population estimates, gastroenteritis cases were reported in 49.3 % (1041/2111) of the water distribution zones and these were available for analysis in the current study. There was a population of 2 389 055 in the 1041 zones, 42.4 % (1 014 014) of which was supplied with drinking water from a combination of source types, 29.6 % (706 187) was supplied with drinking water from surface water-only sources, 27.5 %

<sup>7</sup>In 2001 Statistics New Zealand estimated the New Zealand population to be 3 880 500: <http://nzdotstat.stats.govt.nz/wbos/Index.aspx?DataSetCode=TABLECODE7511>



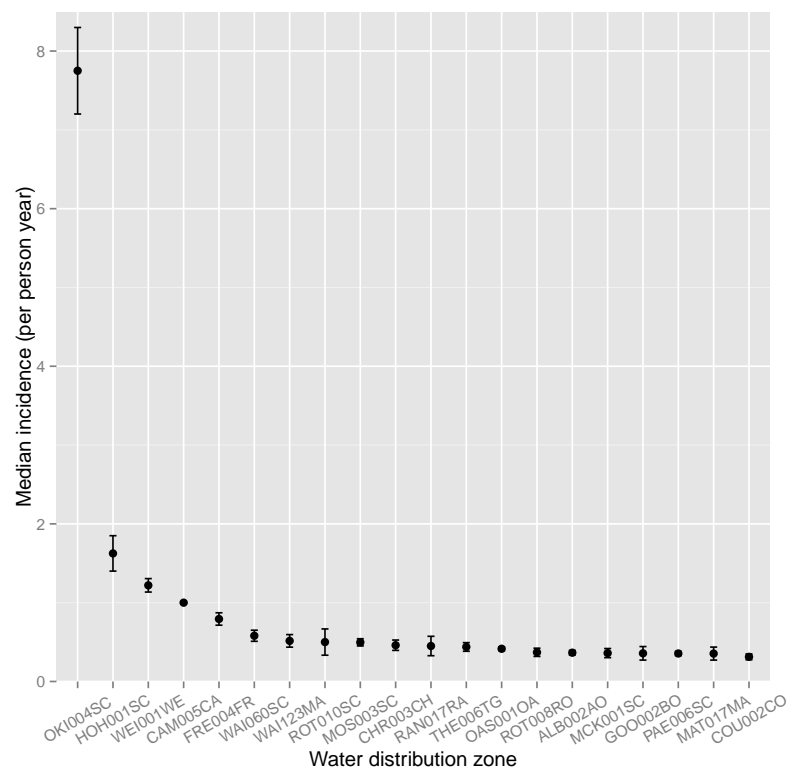
**Figure 4.2:** Number of gastrointestinal cases recorded in the national notifiable disease database (EpiSurv) between 1997 and 2006, stratified by gender and illness, New Zealand.

**Table 4.2:** Description of variables used in the principal correspondence analysis with accompanying cumulative numbers of cases for each gastrointestinal illness and the underlying populations during the study period (1997–2006), New Zealand. The drinking water distribution zone populations were based on the 2001 figures.

Variable	Description	Campy <sup>§</sup>	Crypto <sup>§</sup>	Giard <sup>§</sup>	Salmo <sup>§</sup>	Total cases	Population	Zones
<b>Community rurality</b>								
Rural1	Rural area with high urban influence	6998	345	869	691	8903	94420	116
Rural2	Rural area with moderate urban influence	3268	339	403	560	4570	102592	204
Rural3	Rural area with low urban influence	2770	415	430	517	4132	90069	314
Rural4	Highly rural-remote area	2016	221	187	293	2717	63970	167
Urban1	Independent Urban Area	5256	599	684	1052	7591	303826	88
Urban2	Satellite Urban Area	2000	149	265	324	2738	69262	25
Urban3	Main urban area	47543	2365	5819	5209	60936	1664916	127
Total		69851	4433	8657	8646	91587	2389055	1041
<b>Water source type</b>								
Bore	Ground water from boreholes	689	50	103	92	934	5523	23
Dam	Surface water from dams	9367	224	1144	858	11593	280974	15
Lake	Surface water from lakes	479	34	68	73	654	23614	17
Multisource	More than one type of water sources	26950	1381	3248	3016	34595	1014014	142
River	Surface water from rivers	7290	807	1017	1174	10288	401599	239
Roof	Roof water sources	3344	213	504	437	4498	10811	148
Spring	Ground water from springs	1226	123	210	145	1704	66561	62
Well	Ground water from wells	20506	1601	2363	2851	27321	585959	395
Total		69851	4433	8657	8646	91587	2389055	1041

<sup>§</sup>**Illness:** Campy = Campylobacteriosis; Crypto = Cryptosporidiosis; Giard = Giardiasis; Salmo = Salmonellosis

(658 043) was supplied with water from groundwater-only sources while roof water-only sources supplied 0.5 % (10 811) of the population. Table 4.2 shows the cumulative number of reported cases of gastroenteritis over the study period, the 2001 population estimates in the zones and the number of zones per variable. A total of 91 587 gastrointestinal illness cases were reported in the 1041 water distribution zones and of these 76.3 % were cases of campylobacteriosis, 9.5 % were giardiasis, 9.4 % were salmonellosis and 4.8 % were cryptosporidiosis. The 91 587 gastroenteritis cases represented 63.8 % of the cases recorded in the EpiSurv database between 1997 and 2006. Of the 91 587 cases 22.2 % (20 322) were located in rural areas (rural1–rural4) while 77.8 % (71 265) was located in urban areas (urban1–urban3). Figure 4.3 shows twenty zones with the highest median annual incidence rates over the study period while Figure 4.4 shows their locations within New Zealand. The median population for the zones shown in Figure 4.3 was 50 (Range: 1; 200).

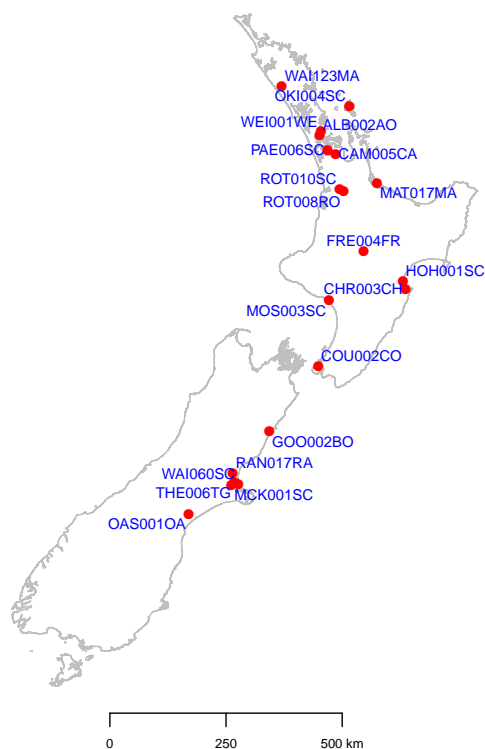


**Figure 4.3:** Twenty drinking water distribution zones with the highest incidence rates during the ten-year period 1997–2006, New Zealand.

4.3.2 Multivariate analysis

A summary of the median zone populations and the median number of cases reported per zone annually, with 95 % confidence intervals, are shown in Table 4.3. The median population for the drinking water distribution zones located in rural areas was smaller (~100) compared to that of distribution zones for urban communities (500–2000). In general, dams supplied water to communities with larger populations than other types of water sources, with bore and roof sources supplying the smallest communities. Also, ground and roof water sources supplied smaller communities compared to surface water sources. The rural communities were serviced by 801 zones while urban communities were serviced by 240 water distribution zones. Rural areas with high urban influence showed a wide variation in the number of reported cases of gastroenteritis compared to other rural areas although their population sizes were similar.

The relationships between gastroenteritis incidence rates, rurality of the serviced community and the types of water sources are shown on a PCA biplot in Figure 4.5. Campylobacteriosis incidence rates were closely correlated with zones having roof water sources located in rural areas with high urban influence. Water from wells and sources located in rural areas with low urban influence or main urban centres were minimally correlated with the incidence



**Figure 4.4:** Location of drinking water distribution zones with the highest gastroenteritis incidence rates in New Zealand, 1997–2006.

rates of gastrointestinal illness. The two protozoan diseases (cryptosporidiosis and giardiasis) and salmonellosis were highly correlated with peri-urban (urban1 and urban2) areas, rural areas with moderate urban influence, surface water sources and bore water sources.

### 4.3.3 Geostatistical analysis

The kriging and bubble plots (Figure 4.6) show that on average the north-western and the southern areas of the North Island were the places to likely report high incidence rates of gastrointestinal illness (all four diseases combined) during the study period. In the South Island, the mid-eastern areas were the predicted places of high gastrointestinal illness incidence rates. There were slight variations in this overall pattern from year to year but the general pattern remained relatively similar throughout the study period. The yearly patterns can be viewed in Figure 4.6 in the electronic version of this document by clicking the control buttons.

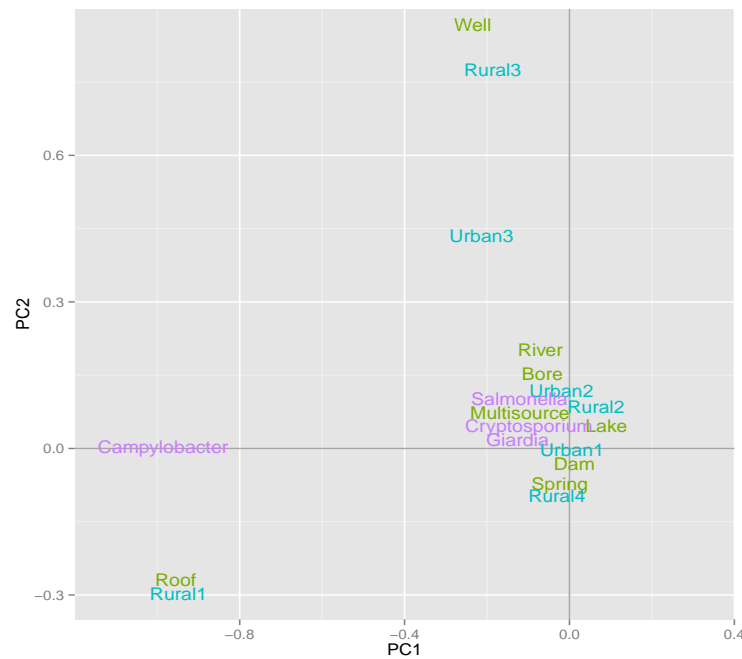
The disease-specific interpolated spatial patterns are shown in Figure 4.7 while bubble plots are shown in Figure A.2 on page 173. Campylobacteriosis patterns were very similar to those observed in the overall gastrointestinal patterns. This is probably because campylobacteriosis cases composed 76.3 % of all the cases used in the analysis. Over the study period, cryptosporidiosis was most likely to be reported in the central North Island and



**Table 4.3:** Median drinking water distribution zone populations and median annual cases, with accompanying 95% confidence intervals, reported per zone stratified by variables used in a principal correspondence analysis over the study period (1997–2006), New Zealand.

Variable	Description	Population			Cases		
		Median	Lower 95% CI *	Upper 95% CI *	Median	Lower 95% CI *	Upper 95% CI *
<b>Community rurality</b>							
Rural1	Rural area with high urban influence	100	21.5	1239.0	2	1.0	18.1
Rural2	Rural area with moderate urban influence	100	20.5	1000.0	1	1.0	8.0
Rural3	Rural area with low urban influence	88	21.3	1007.0	1	1.0	6.0
Rural4	Highly rural-remote area	95	24.3	900.0	1	1.0	7.0
Urban1	Independent Urban area	2000	52.7	13075.0	2	1.0	16.0
Urban2	Satellite urban area	1600	96.0	11100.0	3	1.0	17.0
Urban3	Main urban area	500	38.7	52850.0	4	1.0	85.3
<b>Water source type</b>							
Bore	Groundwater from bore-holes	50	20.4	721.5	2	1.0	12.8
Dam	Surface water from dams	2000	150.0	65920.0	7	1.0	154.0
Lake	Surface water from lakes	400	69.0	4300.0	2	1.0	12.0
Multisource	More than one type of water sources	500	26.2	26815.4	3	1.0	63.0
River	Surface water from rivers	250	30.0	5020.0	2	1.0	14.0
Roof	Roof water sources	58	16.8	200.0	1	1.0	14.0
Spring	Spring water sources	100	10.1	1509.0	1	1.0	12.0
Well	Groundwater from wells	120	25.0	2951.0	2	1.0	22.0

\*Confidence interval



**Figure 4.5:** Principal components analysis biplot of gastrointestinal illness annual incidence rates for the ten-year period 1997–2006, New Zealand. Rural1 is less rural while Rural4 is highly rural, similarly, Urban1 is less urban while Urban3 is highly urban.

to a lesser extent in the central South Island. There were large cryptosporidiosis pattern variations from year to year. The northern parts of central North Island were the likely areas, on average, to have reported cases of giardiasis over the study period. As with cryptosporidiosis, there were large variations in the yearly giardiasis spatial patterns. For salmonellosis, the central North Island and the mid-eastern parts of South Island were the likely areas to report high cases of illness. The yearly salmonellosis patterns showed that cases were more likely to be reported in the south-western areas of North Island and the mid-eastern parts of South Island.

#### 4.3.4 Distributed lag analysis

**Table 4.4:** Drinking water abstraction sites used in distributed lag non-linear modelling. For each source site tabulated is the number of drinking water distribution zones serviced, the number of cases per disease, the total number of cases over the study period and the population serviced. Also tabulated is the river on which the abstraction point was sited and the city or town in which the serviced population was located.

Site	River	City/Town	Zones	Campy <sup>§</sup>	Crypto <sup>§</sup>	Giard <sup>§</sup>	Salmo <sup>§</sup>	Total Cases	Population
S00041	Waikato	Waikato, Hamilton, Waipa	12	2138	160	361	203	2862	120812
S00079	Ohau	Horowhenua	1	180	27	27	46	280	20000
S00082	Turitea	Palmerston North	4	466	58	76	105	705	70800
S00106	Te Arai	Gisborne	1	221	20	61	29	331	30000
S00107	Waipaoa	Gisborne	1	221	20	61	29	331	30000
S00118	Hutt	Wellington, Lower Hutt, Upper Hutt, Porirua	24	9891	622	1163	944	12620	232556
S00120	Wainuiomata	Wellington, Lower Hutt, Upper Hutt	12	5990	354	699	516	7559	144940
S00121	Orongorongo	Wellington, Lower Hutt, Upper Hutt	12	5990	354	699	516	7559	144940
S00123	Waikanae	Kapiti Coast	5	383	16	28	40	467	28818
S00200	Pareora	Timaru	2	801	52	31	89	973	26832
S00217	Whakatane	Whakatane	2	227	9	37	47	320	15000
S00233	Waiwhakaiho	New Plymouth	7	1024	27	41	107	1199	48777
S00268	Maitai S. Branch	Nelson	3	301	10	33	87	431	40000
S00270	Roding	Nelson	2	301	10	33	87	431	20000
S00735	Mangatangi	Auckland, Manukau, North Shore, Papakura	35	9860	194	1344	880	12278	797818
S00865	Waikato	Auckland, Manukau, North Shore	25	7921	161	1146	721	9949	606601
Total			96	25492	1195	3202	2577	32466	1431413

<sup>§</sup>Illness: Campy = Campylobacteriosis; Crypto = Cryptosporidiosis; Giard = Giardiasis; Salmo = Salmonellosis

The eleven drinking water abstraction sites used in the lag analysis supplied drinking water through 96 distribution zones to a population of 1 431 413 (Table 4.4). The 96 distribution zones were 9.2 % of the 1041 zones available for analyses in the current study while the 1 431 413 population was equivalent to 36.9 % of the 2001 New Zealand population. A total of 32 466 cases of gastroenteritis, representing 22.6 % of the cases recorded in the EpiSurv database between 1997 and 2006 were included in the distributed lag analysis. Of these 78.5 % were cases of campylobacteriosis, 9.9 % were giardiasis, salmonellosis 7.9 % and 3.7 % were cryptosporidiosis.

A DLNM involving all the eleven sites (Figure 4.8) shows that the relationship between river flow and reports of gastrointestinal illness (campylobacteriosis, cryptosporidiosis, giardiasis and salmonellosis combined) was most positive at a lag of about 10 days and around 90th flow percentile. At these values there was an average increase of 3% in the number of gastrointestinal illness reports compared to the 50th flow percentile after controlling for month and drinking water abstraction site. Figures 4.9–4.12 (pages 101–104) show a subset of the single-site DLNMs with the rest of the single-site models shown in Figures A.3–A.9 (pages 174–180). Single-site DLNMs show variations in the river flow-gastrointestinal illness relationship with respect to both lag and river flow percentile. For example, flow on the Ohau River appeared to be most positively related to gastrointestinal illness reports at lag 20–30 days and around 70th percentile flow (Figure 4.9) while the river flow-gastrointestinal illness relationship was most positive at lag 10 days and 90th percentile flow for Hutt River (Figure 4.10). In general, there was a smaller increase in the number of gastrointestinal illness reports attributed to distributed lag river flow in distribution networks with large populations compared to those with smaller populations. For instance, the average relative risk of gastrointestinal illness reports peaked at about 3% for the network supplied by Mangatangi River (population: 797 818) compared to a peak of 15% in the Ohau River-supplied network (population: 20 000).

## 4.4 Discussion

Previous studies have used passive surveillance data to show how waterborne illness varies in space and time (Britton et al., 2010a,b; Khan et al., 2007). Other studies have reported a profound relationship between hydrology and surface water quality (Göransson et al., 2013; Jagai et al., 2012; Lawler et al., 2006). In turn, source water quality has been associated with waterborne illness (Beaudeau et al., 2014). Other research has reported evidence of a link between gastrointestinal illness and river flow (Beaudeau et al., 2014; Khan et al., 2007). Such studies motivate the present study to investigate factors related to gastrointestinal illness associated with drinking water in New Zealand with respect to space, time and hydrology. In order to achieve this, passive national disease surveillance data were used to examine both the spatial and temporal patterns of cases of gastrointestinal illness in New Zealand over a ten-year period, 1997–2006. Correlation in the data was exploited in order to reveal residential and water supply factors associated with the occurrence of gastrointestinal illness, although the magnitude of the association was not quantified. Geospatial exploration of gastrointestinal illness across New Zealand was conducted and identified areas where high incidence rates were prevalent. The relationship between reports of gastrointestinal illness and river flow on drinking water source rivers was quantified with an assumption that there was a lapse between flow activities and reports of illness.

Data analysis was performed using three main techniques: principal component analysis (PCA), kriging and distributed lag non-linear model (DLNM)s. PCA is a common multi-

variate technique that utilises correlation in a given dataset in order to expose similarities and/or differences among variables. Previous studies have used PCA as a tool for investigating factors associated with water quality and/or gastrointestinal illness (Cinque and Jayasuriya, 2010; Lyra et al., 2009; Pham-Duc et al., 2014). Kriging is a well established spatial interpolation technique that is used to identify spatial patterns in a given dataset. Pardhan-Ali et al. (2012) used the kriging technique to describe the spatial distribution of notifiable gastrointestinal illness in the Northwest Territories, Canada. In another study, Berke (2004) performed exploratory disease mapping to demonstrate the application of regional estimates in the kriging analysis. DLNMs are suitable for describing outcome-exposure relationships that are characterised by time lapse such as between exposure to infectious disease-causing organisms and manifestation of illness. Distributed lag analysis has previously been used to investigate the relationship between precipitation and microbial pollution of raw water (Tornevi et al., 2014); ambient temperature and mortality (Armstrong, 2006); air pollution and mortality (Schwartz, 2000; Zanobetti et al., 2000); capital appropriations and expenditure (Almon, 1965).

Summary data show that between 1997 and 2006 *Campylobacter* was the major cause of gastrointestinal illness in New Zealand and the rates of illness were increasing. Males were slightly more affected than females. Crude incidence rates showed that gastroenteritis was likely to be reported in small water distribution zones, among which were those not regulated by government. These were likely to be zones with inadequate or no water treatment facilities such as households harvesting rain water using the roof of the house. These findings were supported by PCA which showed that communities with roof water supplies located in rural areas with high urban influence were strongly associated with campylobacteriosis. Further, evidence from the distributed lag analysis showed that the relative risk of gastrointestinal illness was likely to be higher in small water distribution networks than large ones. A previous study by Eberhart-Phillips et al. (1997) also reported roof water as a risk factor for campylobacteriosis in New Zealand. However, Eberhart and co-workers did not find residence (urban or rural) as a risk factor. Elsewhere, Jose Figueras and Borrego (2010) observed that the European Commission had recognised that small drinking water distribution systems posed high public health risk and required special attention in order that universal delivery of microbiologically safe drinking water could be achieved. In a study by Ahmed et al. (2012) faecal indicator organisms were detected in more than two thirds of roof water samples in Southeast Queensland, Australia. In the same study, *Campylobacter*, *Salmonella* and *Giardia* were isolated in 21 %, 4 % and 13 %, respectively, of roof water samples.

Spatial interpolation identified areas where high rates of gastrointestinal illness (i.e. all four study gastrointestinal diseases combined) were likely to be reported in New Zealand. The northern, north-eastern and southern parts of the North Island of New Zealand were found

to be most likely to report high incidences of gastrointestinal illness. Also, high incidence rates of gastrointestinal illness were likely to occur in the north-western, mid-eastern and southern areas of the South Island. The findings of the spatial analysis suggest that there are local factors, such as common food sources or water sources, associated with high incidence rates of gastrointestinal illness. This means that in order to reduce the incidence of gastrointestinal illness in New Zealand the identified areas should be the focus of attention by the health authorities. Measures such as enhanced active surveillance and application of local-based interventions would be highly valuable in controlling the rates of illness.

The distributed lag analysis revealed that the highest increase in risk of gastrointestinal illness reports was around ten days after high flow ( $\sim 90$ th percentile). The ten-day duration between exposure and report of illness is consistent with the incubation period of *Campylobacter* organisms which averages about 2–4 days (range: 1; 10 days) (Horn and Lake, 2013). The 3% increase in cases could be an underestimate of the real increase for two main reasons. The first reason is that only about 20% of persons with gastrointestinal illness in New Zealand seek medical attention of which about a quarter submit a laboratory sample (Lake et al., 2010; Lake et al., 2009). The notified cases recorded in the the EpiSurv database are derived from laboratory reports. The second reason is that the largest percentage of the data used in the distributed lag analysis (Table 4.4) were from relatively large distribution networks that were likely to have water treatment plants capable of coping with large variations in the microbial load of the source water following high flow. During the study period river flow data were not available on rivers used as water sources by many small water distribution networks in New Zealand. If such networks were included in the analysis probably the observed increase in the number of gastrointestinal illness cases would have been higher as small networks were less likely to cope with heavier than normal occurrence of microbial load in raw water following high flow. Evidence of a positive relationship between reports of gastrointestinal illness in the community and the river flow rates on drinking water source rivers suggests that strategies aimed at enhancing drinking water safety should include hydrological factors. For example, river flow could be included as a factor in the calculation of a log credit for a given water treatment plant at any particular time. Thus by combining meteorological forecast data and anticipated river flow rates the pathogen removal measures could be adjusted accordingly at any given time.

Sources of bias in the current study include the fact that the number of cases among the four gastrointestinal diseases used in the different analyses was not even. For example, campylobacteriosis cases composed over three-quarters of all cases while cryptosporidiosis cases made up only 5% of the data. This implies that inferences based on analyses in which the disease data were combined would be greatly influenced by the variations in campylobacteriosis cases. In the distributed lag analysis, the multiple-site model was implemented with the assumption that the exposure-response (river flow-gastroenteritis)

relationship was constant across all sites. However, variations due to factors such as differences in catchment microbial load, water treatment plant's ability to remove pathogens and the capacity of treatment plants to withstand sudden increases in raw water microbial load due to increased river flow can be expected to influence the relationship differentially across sites. Therefore, a modelling approach such as the one proposed by Gasparrini et al. (2012) would be most appropriate. However, model implementation with an assumption of constant variance across space and time is not uncommon. Examples in which models are implemented with such kind of assumption include ordinary kriging (Webster and Oliver, 2008). Further, Gelman and Hill (2007) suggests that when model implementation is conducted without a group-level predictor, separate group-level models could be provided to guide the estimates of the varying coefficients. This approach was adopted in the current study.

Caution should be exercised when interpreting the magnitude of the relationship between gastrointestinal illness and drinking water-related factors such as types of water sources and river flow. This is because it has previously been shown that foodborne infection is the most significant contributor to the occurrence of gastrointestinal illness in New Zealand (Eberhart-Phillips et al., 1997; Muellner et al., 2013). Elsewhere foodborne infection has been shown to be strongly associated with campylobacteriosis in humans (Gormley et al., 2008; Sheppard et al., 2009; Wagenaar et al., 2006). Therefore, infection attributed to drinking water in the current study may be overstated as the data used did not distinguish whether the cases were of foodborne or waterborne origin. However, the fact that reports of gastrointestinal illness appear to occur within the *Campylobacter* (the largest proportion of the data) incubation period and is consistent across many sites suggests that a true positive gastrointestinal illness-river flow relationship does exist.

In conclusion, evidence in the current study show that reports of gastrointestinal illness are likely to be highest in small supplies such as those with roof water sources located in rural areas. Based on these findings it seems prudent to recommend that in order to reduce the incidence rate of gastrointestinal illness attributed to drinking water, measures such as installation of small-scale filtration and ultra violet (UV) treatment units should be promoted among small supplies that currently have no water treatment facilities. Further, enhanced active health surveillance and application of local-based intervention measures should be promoted in areas identified as likely to report high incidence rates of gastrointestinal illness. The gastrointestinal illness-river flow analysis provided evidence suggesting that river flow is an important factor that should be taken into account when determining a drinking water treatment plant's ability to remove pathogens from raw water. A proposed way in which to incorporate river flow into the treatment plant decision making process is to include river flow rates as a factor in the log credit removal calculations.

(a) Kriging map

(b) Bubble plot

**Figure 4.6:** Median annual gastrointestinal illness (campylobacteriosis, cryptosporidiosis, giardiasis and salmonellosis) case incidence rates for the period 1997–2006, New Zealand. Subfigure (a) shows the predicted incidence rates using the kriging technique, with dark colours showing high rates and light colours showing low rates. In subfigure (b) the median incidence rates are shown as bubbles, large bubbles represent high rates and small bubbles represent low rates. The electronic version of this document also shows the incidence rates for each year.

(a) Campylobacteriosis

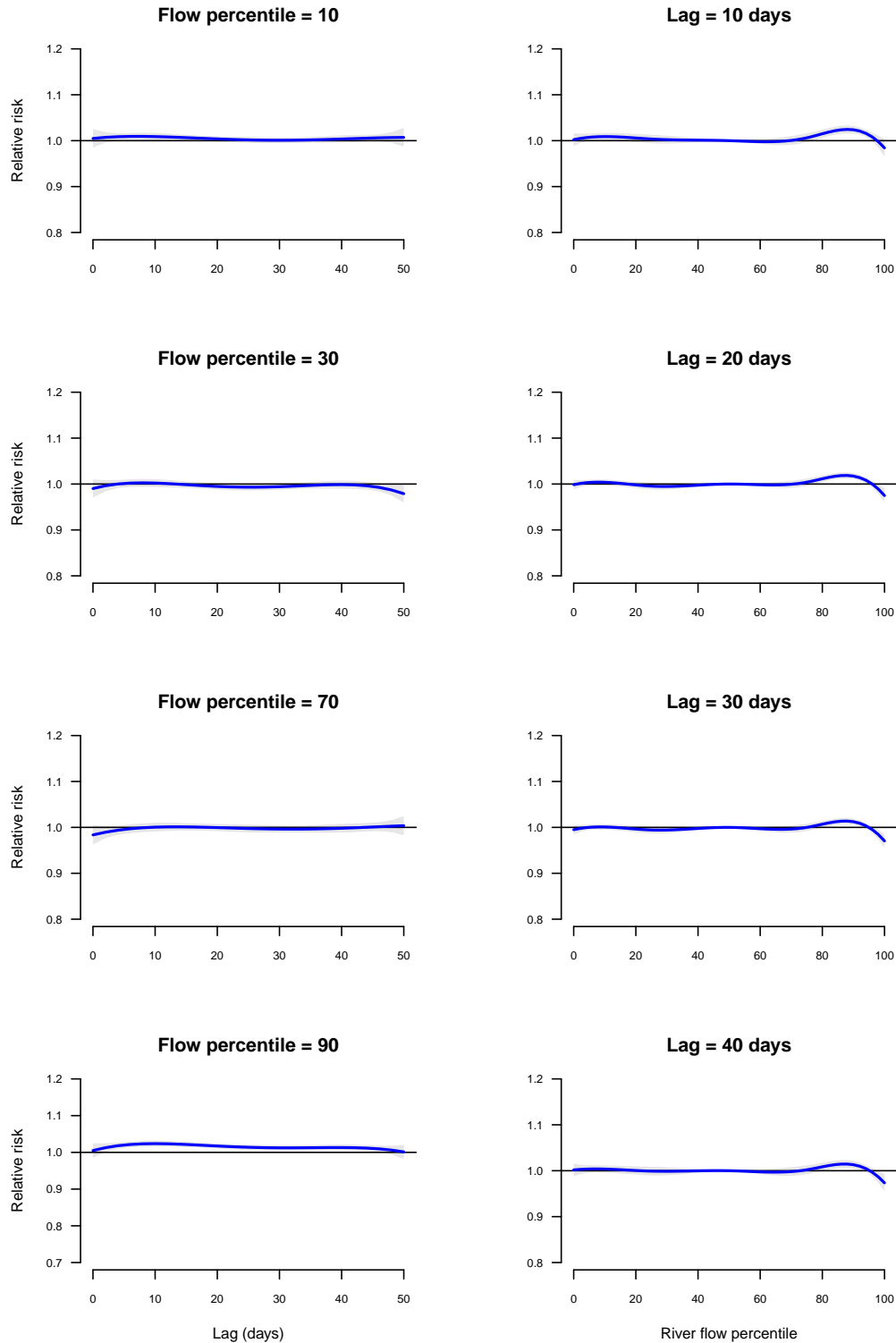
(b) Cryptosporidiosis

(c) Giardiasis

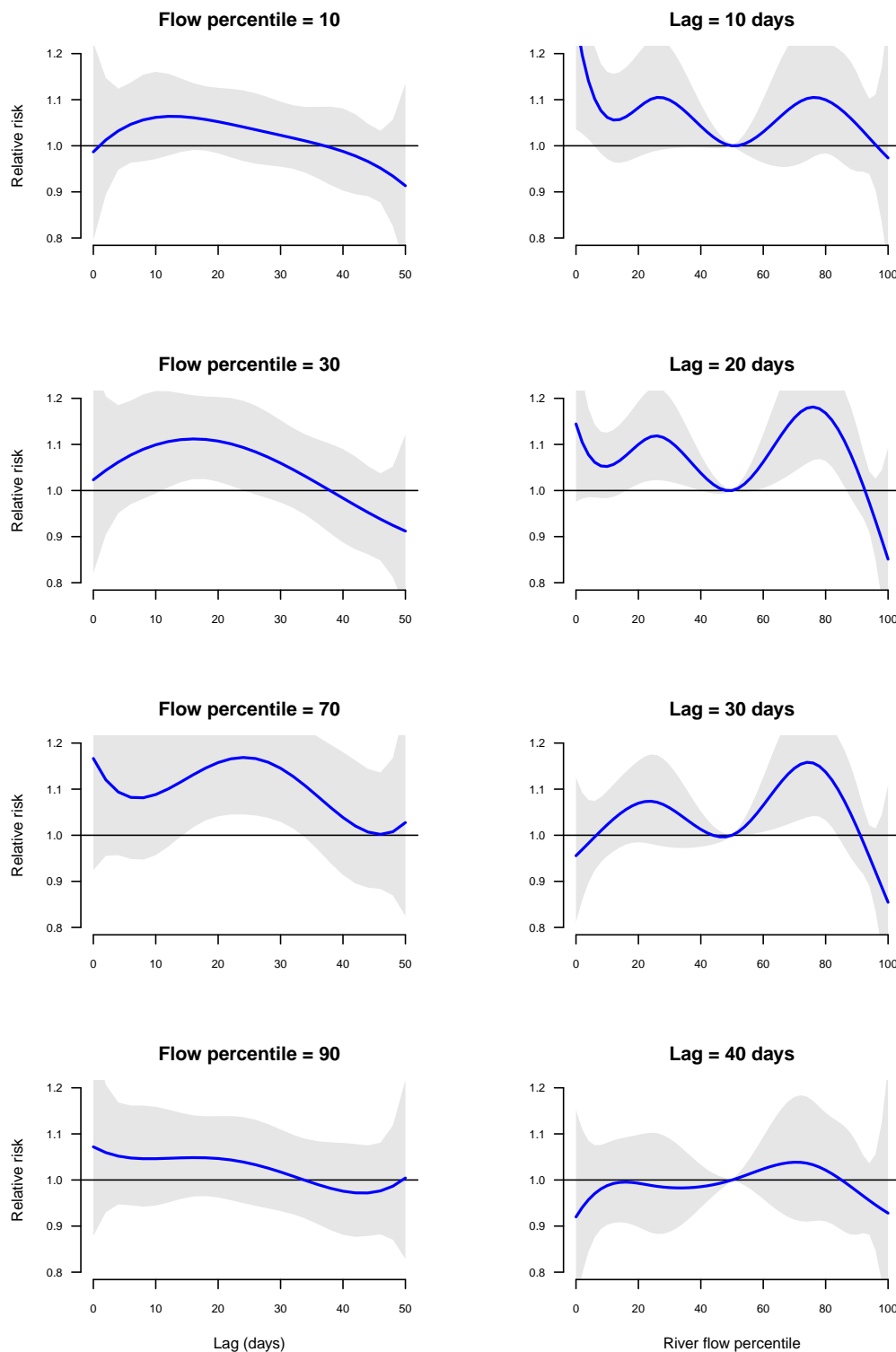
(d) Salmonellosis

**Figure 4.7:** Kriged median annual gastrointestinal illness case incidence rates for the four study diseases during the ten-year period 1997–2006, New Zealand. Dark colours represent high rates and light colours represent low rates. The electronic version of this document also shows the kriged rates for each year.

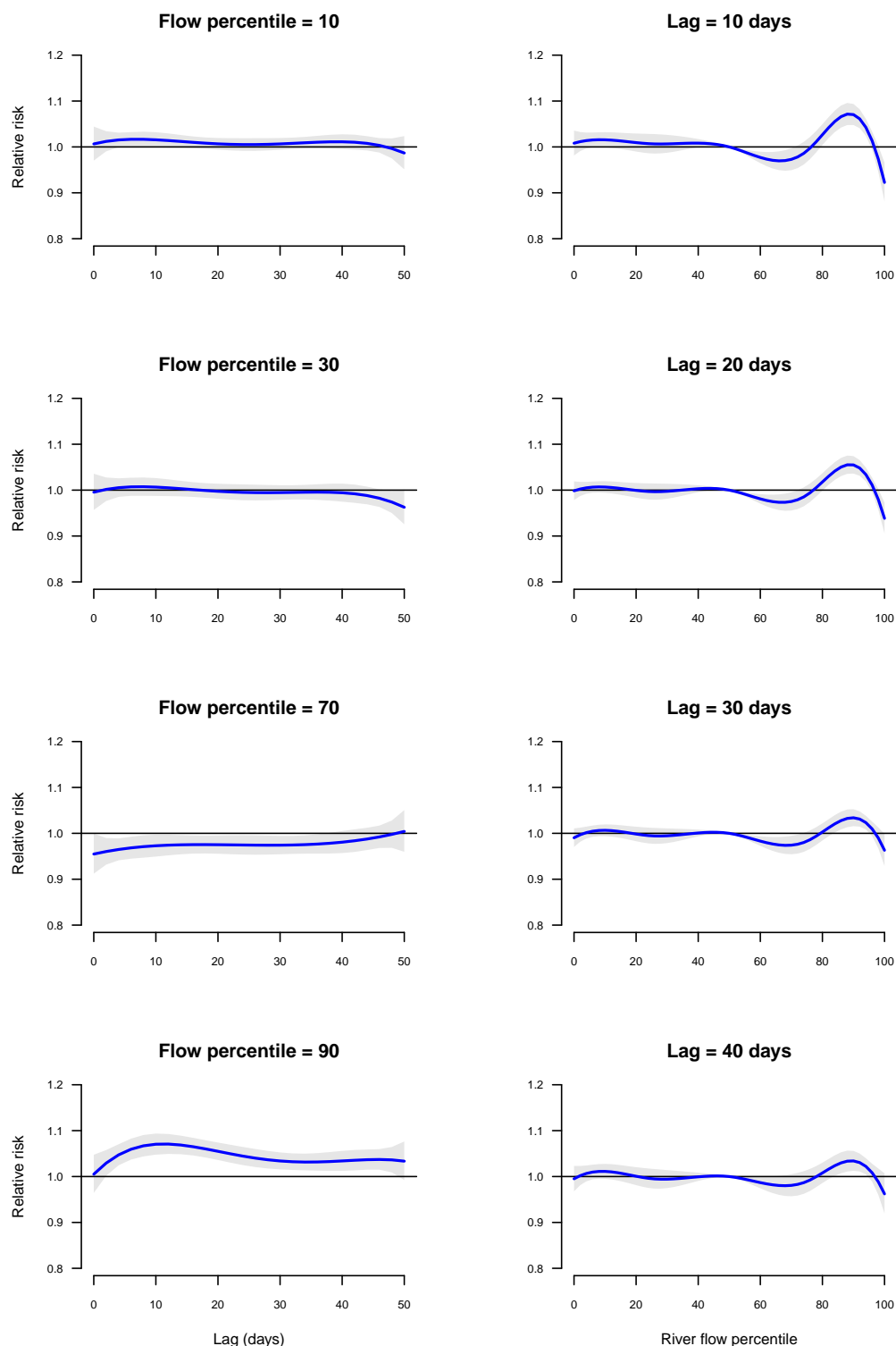




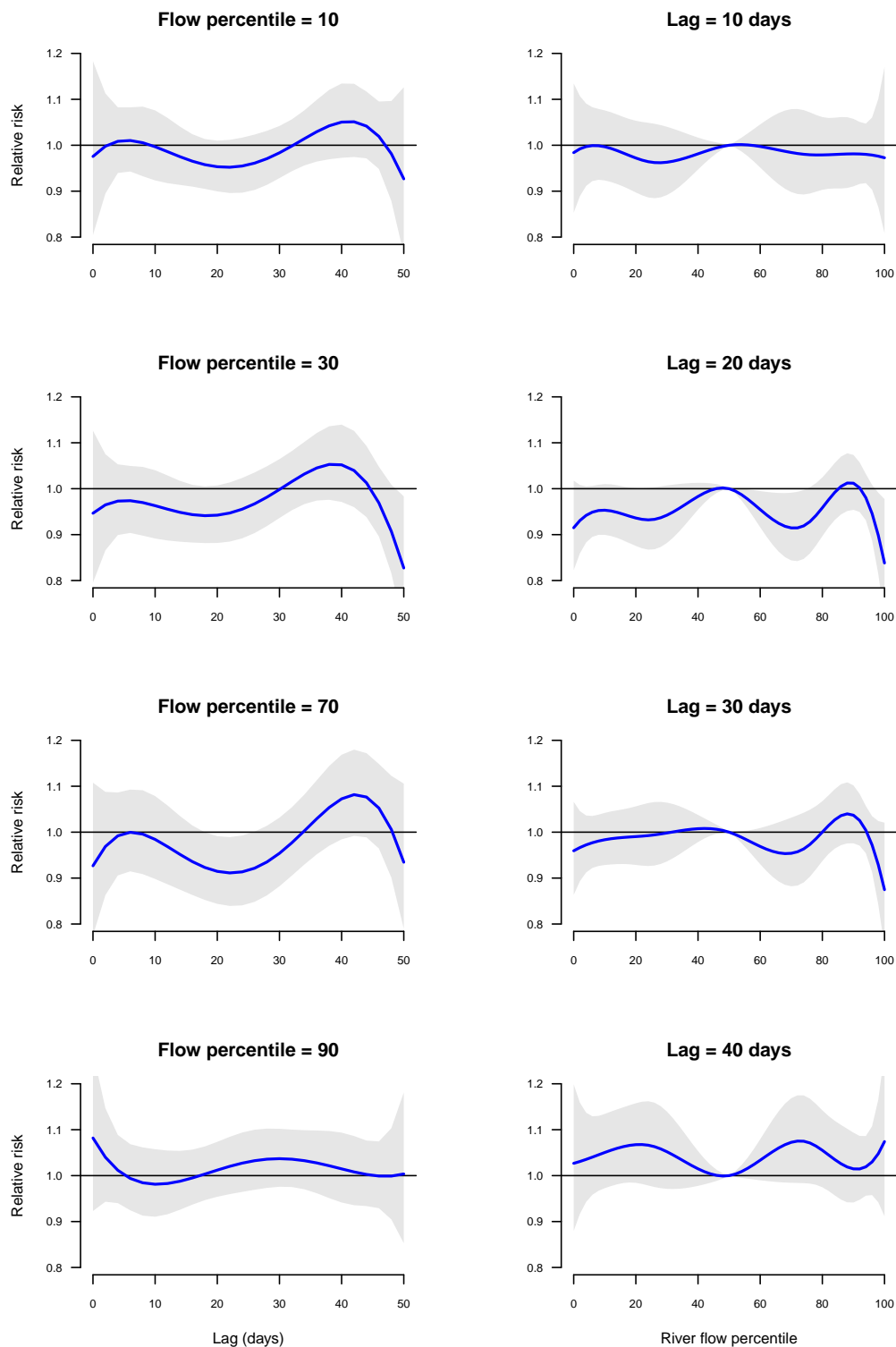
**Figure 4.8: Multiple sites** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month and site, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence intervals. The DLNM was fitted to data from eleven different drinking water abstraction sites within New Zealand, 1997–2007.



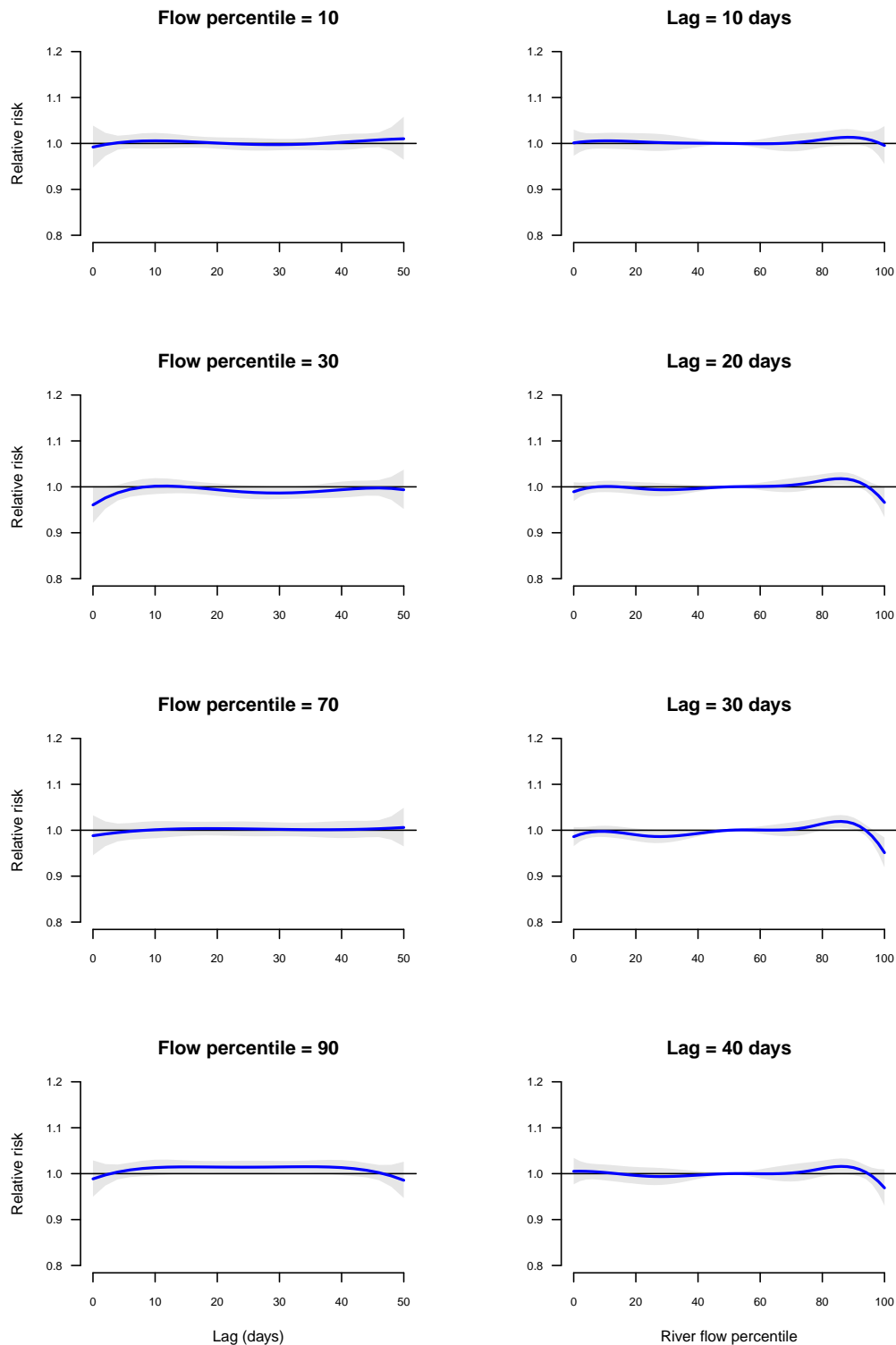
**Figure 4.9: Ohau River (S00079)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Ohau River, for the period 1997–2007, New Zealand.



**Figure 4.10: Hutt River (S00118)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Hutt River, for the period 1997–2007, New Zealand.



**Figure 4.11: Whakatane River (S00217)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95% confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Whakatane River, for the period 1997–2007, New Zealand.



**Figure 4.12: Mangatangi River (S00735)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95% confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Mangatangi River, for the period 1997–2007, New Zealand.

# Five

## The culture-based microbiology of drinking water on campgrounds in New Zealand

### 5.1 Background

Outdoor and nature-based activities are at the centre of New Zealand culture and lifestyle. In addition, New Zealand is a destination for tourists from many countries worldwide who often participate in the outdoor lifestyle. Among the most popular nature-based activities with tourists are visiting beaches, lakes, geothermal attractions/hot pools, glaciers, glow worm caves and national parks. Other activities include sighting wildlife, fishing, scenic boat cruises, scenic drives, trekking/bush walks and camping. In 2008, 2 million tourists took part in nature-based activities, that produced 11.2 million nature-based trips, of which 1.6 million involved international tourists and 9.6 million involved domestic tourists. The most popular period for tourist visits was between December and March, with nearly 44 % of all international visitors arriving during these months (New Zealand Ministry of Business, Innovation and Employment, [2009](#)).

The tourism industry makes a significant contribution to the New Zealand economy. It is estimated that in 2013 the tourism industry generated NZ\$24 billion; NZ\$14 billion of which was attributed to domestic visitor expenditure and NZ\$10 billion to international visitor expenditure. This amount of revenue translated into 3.7 % of gross domestic product (GDP). The industry provided employment equivalent to 110 800 full-time jobs or 5.7 % of the New Zealand workforce (New Zealand Ministry of Business, Innovation and Employment, [2013](#)).

The Department of Conservation (DOC) plays a key role in the promotion of nature-based tourism as it is the central government agency responsible for the conservation of natural and historic heritage in New Zealand. Among the key functions of DOC is to promote the use of natural and historic resources not only for recreation but also for tourism. The legislative mandate for DOC to conserve natural and historic heritage is provided through the [Conservation Act 1987](#), the [National Parks Act 1980](#) and the [Reserves Act 1977](#). Among the activities that DOC undertakes in order to fulfill its mandate are construction and management of outdoor recreation facilities as well as collaboration with private businesses operating on public conservation land. Another key role for the Department is to promote science and research in conservation. It is estimated that DOC protects flora and fauna in 30 % ( $8.4 \times 10^6$  ha) of New Zealand.

Since most of the outdoor activities, particularly camping, take place during summer months when water consumption is high, there is the potential for waterborne gastroenteritis outbreaks. A previous report linked an outbreak of campylobacteriosis at a school camp to drinking water (Bohmer, 1997) while another reported an outbreak of *Norovirus* gastroenteritis due to an unusual water supply contamination at a ski resort (Hewitt et al., 2007). However, no outbreaks of gastroenteritis related to drinking water on DOC-operated campgrounds have been reported in New Zealand.

Elsewhere, outbreaks of waterborne gastroenteritis associated with outdoor activities or recreation facilities have been reported (Arvelo et al., 2012; Boccia et al., 2002; Boulware, 2004). Arvelo et al. (2012) reported that 77% of the 119 persons that participated in a first-grade school excursion developed acute gastroenteritis during the three-day event outside of Guatemala City, Guatemala. The cause of the outbreak was attributed to tap water which had a most probable number (MPN) of 146 and 3 of coliform and *E. coli* cells per 100 mL, respectively. Water was pumped from a well on the premises and treated using an ozone purification system. Laboratory tests confirmed *Norovirus* genogroups I and II as the cause of this particular outbreak. The outbreak reported by Boccia et al. (2002) involved 344 cases at a tourist resort in the Gulf of Taranto, Italy. Although the resort was supplied with water from the main public water network, a few days prior to the outbreak there was a break in the supply system and the resort was connected to an unused irrigation system. A Norwalk-like virus (*Norovirus*) was found to be the cause of illness. A prospective surveillance study, conducted by Boulware (2004), involving 228 Appalachian Trail backpackers who hiked for at least 7 days found that 56% of the participants had experienced gastrointestinal illness. Consumption of untreated water was implicated as the major contributing factor to experiencing gastrointestinal illness. Other risk factors were those related to unhygienic practices e.g. not washing hands after defecating and not routinely cleaning cooking utensils.

Other examples of outdoor-related gastroenteritis outbreak investigations include those conducted by (Morens et al., 1979; Nygård et al., 2004; Waarbeek et al., 2010). In the investigation conducted by Morens et al. (1979), 55% of the 760 persons who had been to a resort camp in Colorado, USA, developed gastrointestinal illness within a week of leaving the camp. Consumption of water or ice-containing beverages was the common factor among the cases. Virus infection was the most probable cause of illness. Drinking water was sourced from a nearby spring in a meadow. An investigation by Nygård et al. (2004) into an outbreak of acute gastroenteritis at a camp in western Norway found that water consumption and taking showers were correlated with the illness. Of the 205 that answered the investigator's questionnaire 134 (65%) had gastroenteritis. Drinking water for the camp was abstracted from a local well and was not treated. *Norovirus* was found in 8 of 10 (80%) faecal samples examined; other pathogens found in faecal samples were

*Campylobacter*, spp. (2/11: 18%), *Rotavirus* (2/11: 18%) and *Adenovirus* (1/11: 9%). An epidemiological investigation conducted by Waarbeek et al. (2010) involved scouts who attended a camp in Belgium. In total, 106 scouts attended camp and 84 returned the questionnaire from the investigators. Of the eighty-four, 85 % had gastrointestinal illness and drinking water was found to be the strongest risk factor. Other factors associated with the illness were latrine use and female gender. *Norovirus* genogroups I and II were detected in 75 % of the cases. Water for the camp was sourced from a nearby farmer's well in canisters and had coliforms, *E. coli* and *Enterococcus* spp. at concentrations > 100, > 70 and 20 per 100 mL, respectively.

These investigations show the potential risk associated with water consumption at facilities used for outdoor recreation. Naturally, circumstances leading to the outbreaks varied from facility to facility but generally involved consumption of inadequately treated water. Direct abstraction of water from contaminated sources and unusual events such as supply disruptions or greater than normal loads of pathogens appear to be common underlying factors in waterborne illness outbreaks at outdoor recreation facilities. In this regard, it is important to develop measures that prevent water source contamination. In order to develop effective methods for prevention of microbial contamination it is important to have an in-depth understanding of the aquatic microbial profile and local factors that influence its variation. The current study employed conventional cultured-based microbiology and statistical techniques to gain an in-depth understanding of the drinking water quality at campgrounds. The study had three objectives: the first objective was to investigate the microbial quality of both source (raw) and tap water intended for human consumption at campgrounds operated by DOC in New Zealand, through the use of conventional microbial contamination indicators. The second objective was to determine the likely sources of faecal contamination of campground drinking water using conventional microbial indicators. The third objective was to recommend possible measures for managing microbial safety of drinking water at DOC-operated campgrounds.

## 5.2 Materials and methods

### 5.2.1 Study campground selection

This was a serial cross-sectional study conducted on DOC-operated campgrounds during the 2011/2012 and 2012/2013 summer months (December-February). A list of all DOC-operated campgrounds was obtained from the Department and used as a sampling frame. The sampling frame was accompanied with descriptive information that included campground class (backcountry, basic, standard or serviced), point-location global positioning system (GPS) coordinates, number of camping sites per campground, type of water supply, rubbish (collected, recycled or not collected) and whether dogs were permitted or not. Individual campground managers were contacted by email to obtain information related



to the type of drinking water treatment employed by each campground. Campgrounds classified as basic had very limited facilities that included toilets and water supplied from a roof or surface source (e.g. stream or lake). Backcountry campgrounds were those that had toilets and water supply (possibly directly from a stream). Picnic tables, cooking shelters or fireplaces were also available at some of the backcountry campgrounds. Standard campgrounds had facilities found at backcountry and basic campgrounds but in addition had showers and rubbish bins. Serviced campgrounds had more facilities and services than all the other types of campgrounds. The extra facilities and services included flush toilets, hot showers and rubbish collection. Laundry facilities and cookers were available at some serviced campgrounds.

Campgrounds that were eligible for selection into the current study were those classified as standard or serviced and had a surface or roof water source. Campgrounds that had a surface water source through a town or city water reticulation system were not eligible. A stratified random procedure was applied to select 15 study campgrounds from among those that were eligible. The number of study campgrounds was limited to fewer than 20 as this was considered to be a practical number of campgrounds for the principal investigator to drive to for sample collection. Some campgrounds were in remote locations with long distances between campgrounds and also long distances between campgrounds and the laboratory at Massey University (Figure 5.1).

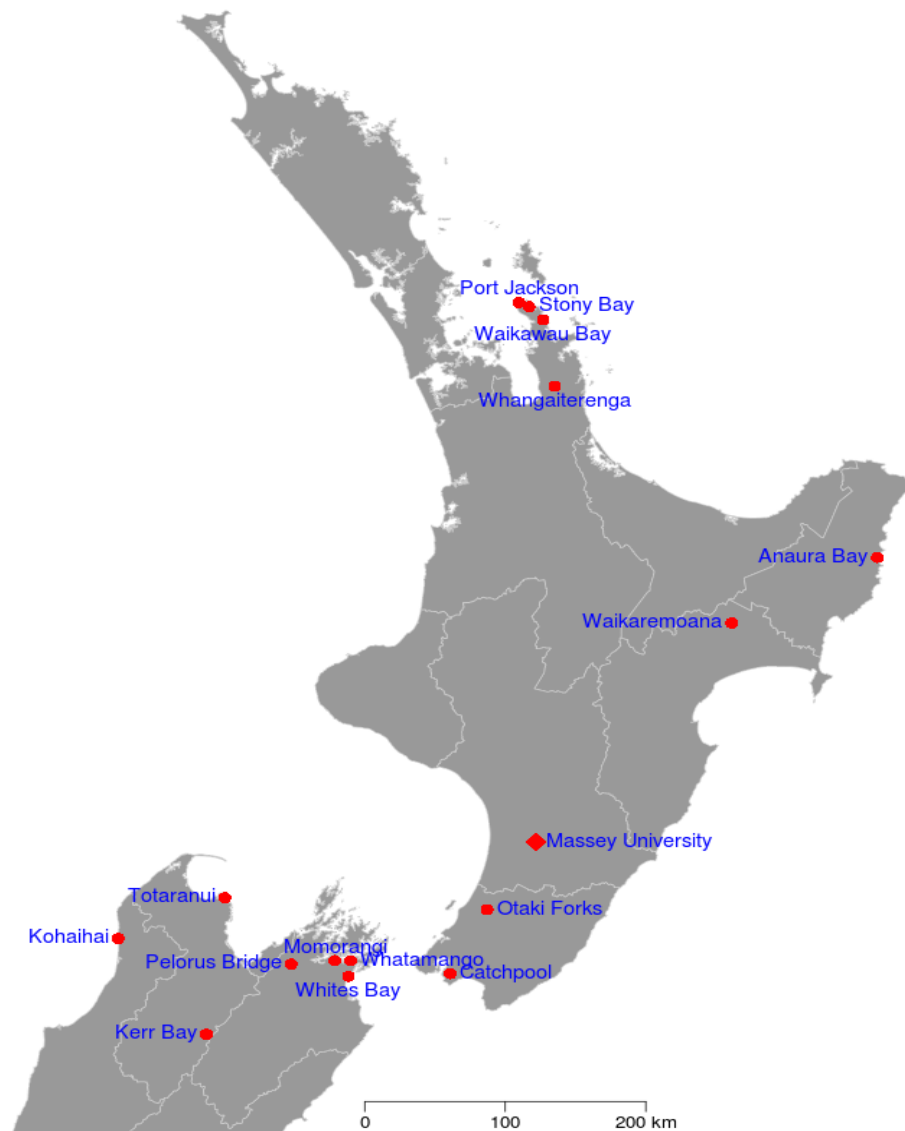
### 5.2.2 Campground water catchment geospatial characteristics

Geospatial characterisation of campground water catchments was performed using GPS coordinates for water abstraction (intake) points and attribute tables from various digital map files (Environmental Systems Research Institute (ESRI) shapefiles). Digital maps for campground water catchments were created in a similar way as described for public drinking water catchments in Section 3.2.4 (page 52), with water abstraction points as the initial reference points. The stream from which Whangaiterenga campground abstracted its drinking water was a very small creek covered by a heavy tree canopy and was missing in the rivers and streams shapefile. For this campground, a 500 m-buffer zone around the abstraction point was used as its catchment. Kohaihai and Whatamango Bay campgrounds were supplied by roof and spring water, respectively, hence did not have a water catchment. The Whatamango Bay campground was erroneously labelled as having a roof water source when recruited into the study. However, by the time the error was noticed samples had already been collected and processed from the campground, it was thus retained in the study. The Totaranui campground had three separate abstraction points, two were used regularly while the third one was used as a backup. Catchment spatial characterisation was only performed for the two regularly used sources at Totaranui, from which samples were collected.

### 5.2.3 Sample collection

A risk-based sampling strategy was adopted, i.e. sample collection coincided with the summer season, the peak camping season in New Zealand. Collection of both water and faecal samples from the study campgrounds was carried out in four rounds over two summer seasons. The first two sampling rounds were conducted during the 2011/2012 summer season, one at the beginning of the season and another at the end. Similarly, the last two rounds took place during the 2012/2013 summer season.

Collaboration with the Department of Ecology at Massey University was established for the purpose of determining an appropriate sampling protocol for wildlife faeces and identification of their animal source. Faecal samples, of any origin, were collected from campground



**Figure 5.1:** Map of the North Island and northern parts of the South Island of New Zealand showing the locations of the study campgrounds and Massey University.

water catchments, for instance, possum faeces shown in Figure 5.2a collected at Otaki Forks campground. Two sets of faecal samples were collected from each campground; swab samples for *Campylobacter* isolation and scoop samples for metagenomic analysis as well as *Cryptosporidium* and *Giardia* screening (Figure 5.2b). Copan Transystem<sup>®</sup> swabs with transport medium containing Amies<sup>®</sup> agar gel with charcoal (Copan Italia Spa; Brescia, Italy) were used for *Campylobacter* sample collection. All samples were geo-referenced using a hand-held GPS receiver, GPSmap 62 (Garmin Limited; Kansas, USA). Samples were stored and transported on ice to molecular epidemiology and public health laboratory (<sup>m</sup>EpiLab) at Hopkirk Research Institute, Massey University, within 36 hours of collection. Samples that were not processed immediately were stored at 4 °C in the laboratory.

Water samples were aseptically collected from a tap within the camping area and at the abstraction point at each campground. Only tap samples were collected at Kohaihai and Whatamango because they had roof and spring sources, respectively. Each tap was disinfected with 70 % ethanol before samples were collected as shown in Figure 5.2c. Water samples were stored and transported to the laboratory on ice in the same way as faecal samples. Three grab and one filter samples were collected from the tap and from the abstraction point. The grab samples (Figure 5.2d) included 100 mL for *Campylobacter* isolation, 100 mL for *E. coli* enumeration and for metagenomics<sup>1</sup> (2 L in the first two rounds and 10 L in the last two rounds). One hundred litres were filtered using a Filtamax<sup>®</sup> filter (IDEXX Laboratories Inc.; Maine, USA) for protozoal screening. Water reticulation pressure was used to force tap water through the filter (Figure 5.2e) while at the abstraction point a portable 12 V battery-driven pump was used (Figure 5.2f). A flow meter attached to the filtration unit measured the volume of water filtered. After filtration, the filter module was removed from the housing, together with residual fluids, and placed in a sealable plastic transportation bag which was then stored and transported to <sup>m</sup>EpiLab laboratory on ice.

#### 5.2.4 Laboratory techniques

Laboratory techniques used for characterising bacterial isolates in the current study included polymerase chain reaction (PCR) and multilocus sequence typing (MLST). PCR has been described in Section 2.4.3 on page 25. MLST is a technique that involves sequence typing of multiple loci in order to identify strains of bacterial species (Dingle et al., 2001). For each bacterial species a set of house-keeping genes, e.g. seven for *Campylobacter*, is selected for the typing scheme and a locus is sequenced on each of the genes. A given sequence at a locus is assigned an allele number. Thus, loci sequences are given different allele numbers if they differ by one or more nucleotides. A strain is identified by a unique combination of the seven allele numbers known as allelic profile or sequence type (ST). Similarity of a core allelic profile is used to categorise STs into clonal complexes (Maiden et al.,

---

<sup>1</sup>Metagenomic sample analyses are discussed in the next Chapter



(a) Possum faeces



(b) Sample bottle & swab



(c) Tap disinfection



(d) Water samples



(e) Tap water filtration



(f) Stream water filtration

**Figure 5.2:** Types of samples collected from campgrounds operated by the New Zealand Department of Conservation during the 2011/12 and 2012/13 summer seasons.

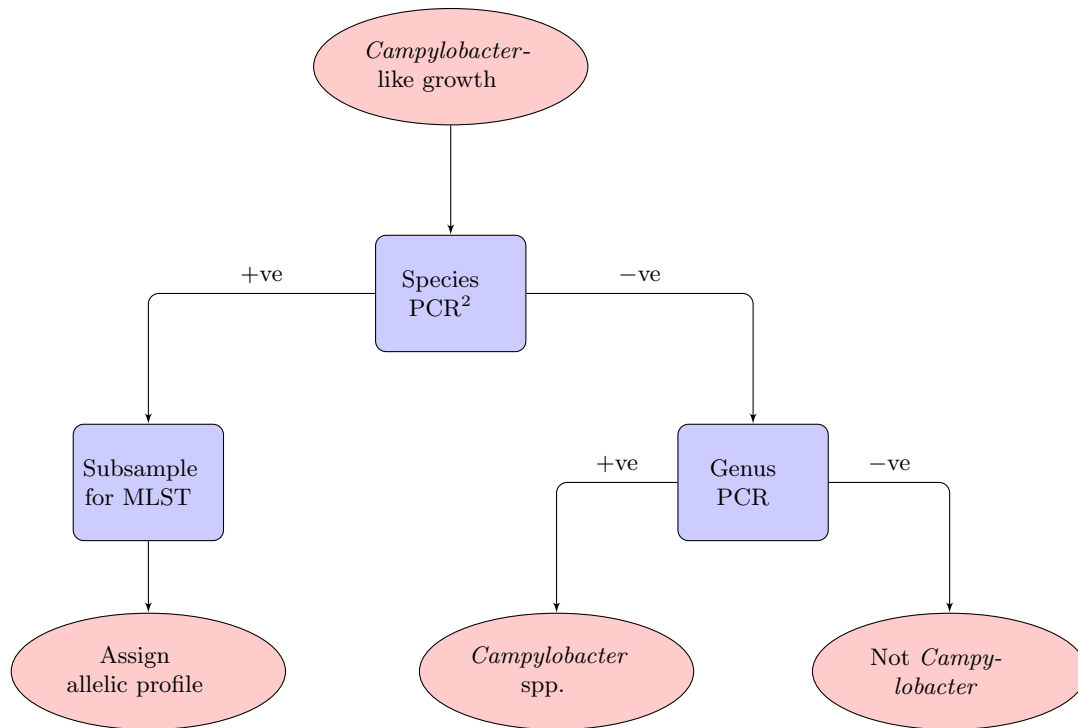


1998; Urwin and Maiden, 2003). In the current study, isolates that had one or more loci not sequenced were classified as *incomplete typing*. Such isolates were assigned an allelic profile if the core loci were sequenced and provided enough information for a particular ST to be assigned. Allelic profiles that could not be matched to existing ones in the PubMLST database were flagged as *new* or referred to as *unknown*.

### 5.2.5 Laboratory processing: Faecal samples

***Campylobacter*:** On arrival at <sup>m</sup>EpiLab laboratory *Campylobacter* enrichment was carried out by incubating faecal swabs in Bolton’s broth under microaerophilic conditions (85 % N<sub>2</sub>, 5 % O<sub>2</sub>, 0 % H<sub>2</sub> and 10 % CO<sub>2</sub>) at 42 °C for 48 h using a Macs-VA500 microaerophilic workstation (Don Whitley Scientific Limited; Yorkshire, UK). After 48 h of enrichment, the broth was cultured onto selective medium, modified charcoal cefoperazone deoxycholate agar (mCCDA), for a further 48 h. From each mCCDA plate with *Campylobacter*-like growth two colonies were subcultured onto separate horse blood agar plates and incubated in the microaerophilic workstation for another 48 h. Deoxyribonucleic acid (DNA) was extracted from a loopful of colonies on each blood agar plate using the protocol outlined in List 1 (page 194) and stored at −20 °C while the rest of the colonies were preserved in glycerol using the protocol outlined in List 2 (page 194) and stored at −80 °C.

PCRs specific to *Campylobacter nova* I (French et al., 2014), *C. coli* and *C. jejuni* were applied to DNA from the isolates. A sample was declared PCR-positive if at least one of the two isolates returned a positive result. Samples that were negative for all three species were subjected to *Campylobacter* genus PCR. Figure 5.3 shows how samples were processed with regards to taxonomic designation. The PCR protocol involved constituting a master mix to which DNA was added to make up to 20 µL (Table A.4 on page 202). A Labcycler<sup>®</sup> machine (SensoQuest GmbH; Göttingen, Germany) was used to perform the reactions on the mixture. The *C. coli* and *C. jejuni* PCR primers targeted the *ceuE* and *mapA* genes, respectively. The *C. sp. nova* I PCR primers targeted a short section of a C4-dicarboxylate trans-membrane transport gene believed to be found only in *C. sp. nova* I. The *Campylobacter* genus PCR targeted the 16S ribosomal ribonucleic acid (rRNA) gene (Linton et al., 1996). The sequences for the pairs of PCR primers used in the current study are shown in Box 3.



**Figure 5.3:** Flow diagram showing the *Campylobacter* taxonomic designation process for samples collected from campgrounds operated by the Department of Conservation surveyed during the 2011/2012 and 2012/2013 summer months, New Zealand.

**Box 3:** *Campylobacter* primer sequences

*Campylobacter* genus ([ibid.](#))

Forward (C412F): 5' GGATGACACTTTTCGGAGC 3'  
 Reverse (C1288R): 5' CATTGTAGCACGTGTGTC 3'

*C. coli* (Denis et al., [2001](#))

Forward (*ceuE*): 5' AATTGAAAATTGCTCCAACATG 3'  
 Reverse (*ceuE*): 5' TGATTTTATTATTGTAGCAGCG 3'

*C. sp. nova* I (mEpiLab; not published)

Forward (Aot10724): 5' GGTGTGTTTGCTGGTCTTGTTGTC 3'  
 Reverse (Aot10724): 5' AAATCCACTCCCCGTTTGC GA 3'

*C. jejuni* (Stucki et al., [1995](#))

Forward (*MapA*): 5' CTTGGCTTGAAATTTGCTTG 3'  
 Reverse (*MapA*): 5' GCTTGGTGCGGATTGTAAA 3'

<sup>2</sup>Including *C. sp. nova* I, *C. coli* & *C. jejuni* PCR

### ***Cryptosporidium* and *Giardia***

Electrofluorescence microscope screening of faecal samples was conducted within 72 h of arrival at <sup>m</sup>EpiLab, following a modified United States Environmental Protection Agency (USEPA) method 1623 (United States Environmental Protection Agency, 2012) protocol outlined in List 5 (page 198). Briefly, a pea-size faecal sample was thoroughly mixed with 700 µL phosphate buffered saline (PBS) in a micro-centrifuge and 50 µL of the supernatant was placed on a microscope slide and incubated at 37 °C for 30–40 min. The slide was fixed with 20 µL methanol and re-incubated at 37 °C for 10 min. The slide was then placed in a humidity chamber and further processed as described in Section 3.2.3 on page 50 but without enumeration.

Faecal samples that returned a positive *Giardia* result (no sample returned a positive *Cryptosporidium* result) on microscopy were confirmed by PCR and assemblages were determined by sequencing. DNA for PCR was extracted using the Nucleospin<sup>®</sup> Soil kit (Macherey-Nagel GmbH & Co. KG; Düren, Germany) (List 4 on page 196). The constituents of the PCR master mix are listed in Table A.4 (page 202) and the PCR conditions are given in Table A.5 (page 203). PCR amplification was carried out using primers targeted at the glutamate dehydrogenase (GDH) locus (Read et al., 2004). The PCR amplicons were sent for sequencing to the Massey Genome Service.

## **5.2.6 Laboratory processing: Water samples**

### ***Campylobacter* and *E. coli***

Samples for *Campylobacter* isolation were filtered through 0.45 µm-pore, 47mm-diameter disks (Millipore Corporation; Massachusetts, USA) soon after arrival at <sup>m</sup>EpiLab. The filter disks were immediately incubated in Bolton’s broth under microaerophilic conditions as described for faecal samples. After broth incubation, processing of samples processed in the same way as described for the faecal samples. The samples for *E. coli* enumeration were submitted to the Central Environmental Laboratories, accredited regional laboratories for water quality testing located in Palmerston North. The Central Environmental Laboratories used a modified Colilert<sup>®</sup> (IDEXX Laboratories Inc.; Maine, USA) enzyme substrate method for enumerating *E. coli*. In summary, contents of one Colilert<sup>®</sup> pack were thoroughly mixed with 100 mL water sample in a sterile bottle. The sample-reagent mixture was poured into a Quanti-tray and sealed using an IDEXX Quanti-tray sealer. The sealed trays were incubated at 35 °C for 24 h. The number of positive wells within the trays were recorded on a worksheet and the IDEXX MPN computer program was used to calculate the results in MPN per 100 mL.

### ***Cryptosporidium* and *Giardia***

Within 72 h of arrival at <sup>m</sup>EpiLab, the Filtia-Max<sup>®</sup> filters were screened for *Cryptosporidium* and *Giardia* following a modified USEPA method 1623 (United States Environmental

Protection Agency, 2012) (List 6 on page 199). A summary of the method is provided in Section 3.2.3 on page 50 although in the current study enumeration was not conducted.

### 5.2.7 *Campylobacter* MLST

Fifty faecal and water *Campylobacter* isolates were subjected to MLST in order to determine their genotype. The isolates included all *C. coli* (15 faecal and one water), all six water *C. jejuni* and 29 randomly selected faecal *C. jejuni* isolates. The master mix and volume of DNA used in the amplification of the seven house-keeping genes are shown in Table A.4 (page 202) while the PCR conditions are outlined in Table A.5 (page 203). The sequences of the primer sets used for the amplification of the house-keeping genes are provided in Box 4. In summary, the MLST protocol involved adding 18.0  $\mu\text{L}$  of master mix containing each of the seven primers into separate wells of a 96-well plate within a DNA-free room. The 96-well plate was then transferred into a PCR room where 2.0  $\mu\text{L}$  DNA was added to each well. The plate was covered with a plastic sealer and spun briefly ( $\sim 500$  rpm) then placed in a Labcycler<sup>®</sup> machine for reaction using the conditions outlined in Table A.5 (page 203).

### 5.2.8 Public health risk assessment

In order to perform the public health risk assessment, the maximum acceptable value (MAV)s for indicator organisms stipulated under the drinking water standards for New Zealand (DWSNZ) 2008 were used to determine the microbial quality of tap water. The standards specify that no *E. coli* should be detectable (i.e.  $< 1$  MPN/100 mL) in treated (tap) water (New Zealand Ministry of Health, 2008). The MAVs defined for recreational water as specified under the microbiological water quality guidelines for marine and fresh-water recreational areas (New Zealand Ministry for the Environment, 2003) were used to determine the microbial quality for source water because no MAVs for source water are defined under DWSNZ.



**Box 4:** MLST primer sequences

Aspartase (*aspA*)

Forward (NZaspF): 5' GARAGAAAAGCWSAWGAATTTAAAGAT 3'  
Reverse (NZaspR): 5' TTTYTTTCATTWGCSTRATRCCATC 3'

Glutamine synthetase (*glnA*)

Forward (NZglnF): 5' TGATAGGMACTTGGCAYCATATBAC 3'  
Reverse (NZglnR): 5' ARRCTCATATHMACATGCATDCCR 3'

Citrate synthase (*gltA*)

Forward (NZgltF): 5' GARTGGCTTGCHGAAAAYAARCTTT 3'  
Reverse (NZgltR): 5' TATAAACCCCTATGYCCAAARCCCAT 3'

Serine hydroxy methyl transferase (*glyA*)

Forward (NZglyF): 5' ATTCWGGTTCTCAAGCWAATCAAGG 3'  
Reverse (NZglyR): 5' GCYAAATCHGCATCTTTKCCRCTAAA 3'

Phospho glucomutase (*pgm*)

Forward (NZpgmF): 5' CWTTRCGYGTDTGTTTTAGATGTVGC 3'  
Reverse (NZpgmR): 5' AATTTTCHGTBCCWGAATAGCGRAA 3'

Transketolase (*tkt*)

Forward (NZtktF): 5' GCWAAAYTCRGGHCAYCCDGGTGC 3'  
Reverse (NZtktR): 5' TTTTAAYVAVHTCTTCRCCCAAAGGT 3'

Adenosine triphosphate synthase alpha subunit (*uncA*)

Forward (NZuncF): 5' GHCAAGGDGTTTRYTGATHTATGTWGC 3'  
Reverse (NZuncR): 5' TTTAADAVYTCWACCATTCTTTGHCC 3'

**NB:** Nucleotide ambiguity codes were used in the primer sequences, see Table A.6 on page 204 for definition of the codes. These primers were developed at <sup>m</sup>EpiLab by Dr P. Biggs based on the New Zealand *Campylobacter* genomes (personal communication).

## 5.3 Data analysis

The types of data available for analysis in the current study included:

1. Campground water catchment attributes (spatial and non-spatial).
2. The concentration of *E. coli* in water (MPN per 100 mL).
3. Presence/absence of *Campylobacter* in 100 mL of water.
4. Presence/absence of *Campylobacter* in faecal samples.
5. Presence/absence of *Giardia* cysts in 100 L of water.
6. Presence/absence of *Giardia* cysts in faecal samples.
7. Allelic profiles for 50 *C. coli* and *C. jejuni* isolates.
8. *Giardia* assemblages.

The data were summarised using standard descriptive data analysis procedures and analysed using regression methods. The allelic profiles, excluding those without an assigned ST, were used to construct a minimum spanning tree using **BioNumerics 7.1**. This was done in order to investigate clustering among the isolates.

### 5.3.1 Regression analysis

Regression analysis is a statistical technique for estimating the strength of relationships among variables. Often the relationship is described using an equation (model) linking the outcome (dependent or response) variable to one or more explanatory (independent or predictor) variables. Regression provides an indication of how the typical (mean) value of an outcome variable changes when any one of the explanatory variables is varied, while other explanatory variable(s) are held constant. It is also used for describing the data structure, predicting future observations or making inferences from a sample to a population from which the sample was drawn (Berk, 2003). Another perspective of regression is that the goal is to find a set of explanatory variables with the largest magnitude of explanatory ability as measured through goodness of fit<sup>3</sup>, i.e. a set of the explanatory variables that can explain most of the variation in the outcome variable. If the regression coefficients are large then as the values of the explanatory variable(s) change from observation to observation, the expected value of the outcome variable varies greatly.

Various types of regression are available and linear regression is among the most commonly used type. Linear regression utilises *linear predictor* functions to model the data and unknown model parameters which are estimated from the data. The linear predictor is the quantity which incorporates the information about the explanatory variables into the model. If one explanatory variable is used then it is called simple linear regression and multiple linear regression if more than one explanatory variables are used. In standard linear regression, the outcome variable is numerical (Gelman and Hill, 2007).

Supposing we have a data array with an outcome variable  $y$  (e.g. body weight) and two explanatory variables  $x_1$  (e.g. body height) and  $x_2$  (e.g. age) for  $n$  individuals:

$$\begin{array}{ccc} y_1 & x_{11} & x_{12} \\ y_2 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} \end{array}$$

a linear regression describing their relationship would be represented by Equation 5.1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{5.1}$$

---

<sup>3</sup>Goodness of fit is how well the model fits a set of observations

where  $\beta_0$  is the intercept,  $\beta_1$  and  $\beta_2$  are unknown parameters (coefficients) and  $\epsilon$  is the error term. The latter accounts for the model's inability to fit the data exactly i.e. what is left over after the conditional mean of  $y$  is subtracted from the observed  $y$ .

Using the data array shown on the previous page, the  $y_i$ 's (for  $i = 1, \dots, n$ ) are called the observations (of variable  $y$ ) while  $x_{i1}$  and  $x_{i2}$  constitute the *design point* corresponding to  $y_i$ . A case or a data point is constituted by  $y_i$ ,  $x_{i1}$  and  $x_{i2}$  together (Sen and Srivastava, 1990). Using the vector-matrix notations, Equation 5.1 can also be written as follows:

$$y = X_i\beta + \epsilon \quad (5.2)$$

where  $X$  is the design matrix and  $\beta$  is the matrix for the coefficients.

The major assumptions in standard linear regression modelling include:

- **Linearity.** This assumes that the outcome-explanatory variable relationship is linear in the regression parameters. The explanatory variables do not necessarily have to be linear, for example they could be transformed using a quadratic function (*ibid.*).
- **Weak exogeneity.** The assumption is that the explanatory variables are composed of fixed values rather than random ones. This implies, for instance, that measurements of the explanatory variables contain no errors or are selected in advance (Chatterjee and Simonoff, 2013).
- **Constant variance (homoscedasticity).** The error term  $\epsilon$  is assumed to have a constant but unknown variance  $\sigma^2$  (Montgomery et al., 2012). This is equivalent to saying that the variance in the outcome variable errors is the same regardless of the values of the explanatory variables.
- **Independence** of errors. This assumes that the errors in the error term are not correlated. Alternatively, the value of one error does not depend on the value of any other error.
- **Normality** of errors. This assumes that the errors have a normal distribution.
- The errors are assumed to have a mean zero.
- **Lack of multicollinearity** in the predictors. This can be triggered by having two or more perfectly correlated predictor variables (e.g. if the same predictor variable is mistakenly given twice, either without transforming one of the copies or by transforming one of the copies linearly). It can also happen if there is too little data available compared to the number of parameters to be estimated (e.g. fewer data points than regression coefficients). This is also referred to as the 'small  $n$  large  $p$ ' problem.

In practice, it is difficult not to violate some or all of the assumptions underlying the basic linear regression model. Thus, numerous extensions of linear regression that allow for

relaxation of some or all of the assumptions have been developed. Among such extensions are generalised linear model (GLM)s which enable modelling of a outcome variable that is bounded or discrete. This type of regression analysis was used to analyse the data in the current study, hence it is discussed below.

### Generalised linear modelling concept

GLM is a framework for statistical analysis, for both linear and non-linear regression modelling, that allows for outcome variables having error distributions other than a normal distribution to be modelled. The framework includes linear and logistic regression as special cases. In addition to variables (outcome and explanatory) and coefficients described on page 117, a generalised linear model also involves (Gelman and Hill, 2007):

1. A *link function*  $g$ , yielding a vector of transformed data  $\hat{y} = g^{-1}(X\beta)$  that are used to model the data.
2. A data distribution,  $p(y|\hat{y})$ .
3. Other parameters, such as variances, overdispersions, and cut-points, that are involved in the predictors, link function, and data distribution are sometimes incorporated.

In a GLM, each outcome of the dependent variables,  $y$ , must be generated from a particular distribution in the exponential family (Montgomery et al., 2012). The most common distributions from this family include the normal, Poisson, binomial, exponential and gamma distributions. The options in a GLM are the transformation  $g$  and the data distribution  $p$ . For example, in linear regression the identity transformation ( $g(u) \equiv u$ ) is used and the data distribution is normal, with standard deviation  $\sigma$  estimated from the data. Logistic regression utilises the inverse-logit ( $g^{-1}(u) = \text{logit}^{-1}(u)$ ) for transformation and the data distribution is defined by the probability for binary data:  $\Pr(y = 1) = \hat{y}$ .

A GLM can be extended to a generalised linear mixed model (GLMM) (also known as hierarchical linear model, nested model, mixed model, multilevel model or random-effects model) by including a random effect in addition to the usual fixed effects. These kinds of models are suitable for analysing data that are structured in groups, such as students within schools and schools in cities. Mixed modelling can be viewed as a generalisation of linear regression in which intercepts (and sometimes the coefficients) are allowed to change by group. Another way of viewing mixed modelling is that it is regression that incorporates a categorical variable depicting group membership. In a regression modelling sense, mixed models may be thought of as techniques for compromising between *complete pooling* (exclusion of the grouping variable) from the model and *not pooling* (estimating separate models within each group). Both these extremes have problems: the effect of no pooling tends to be overestimation of the variation among groups and make the individual groups appear more different than they really are (Gelman and Hill, 2007).

**Generalised linear modelling implementation**

**Faecal model:** The **R** package `lme4` (Bates et al., 2014) was used to implement GLMs in the current study. A GLM was fitted to the *Campylobacter* faecal data in order to estimate the odds of obtaining a PCR-positive result from a non-passerine faecal sample compared to that of a passerine faecal sample while accounting for the island in which samples were collected. In this model, the outcome variable was a binary PCR result indicating whether or not a sample was positive for one or more of the four *Campylobacter* PCRs: *C. sp. nova* I, *C. coli*, *C. jejuni* or *Campylobacter* spp.. The explanatory variables were the animal source of the faeces and Island (North or South) in which the sample originated. The types of animals included were passerine (including sparrow and starling), duck (including swan), rail (pukeko and weka), seagull and domestic ruminant (cattle and sheep). Deer, dog, possum, rabbit and rodent were excluded from the analysis because none of their faecal samples were PCR-positive for *Campylobacter*.

**Water models:** The odds ratio of a water sample exceeding the *E. coli* MAV in tap water, i.e. containing at least one *E. coli*, to that of not exceeding were estimated using a GLMM belonging to the binomial family. Explicitly, in the tap water GLMM the outcome variable was a binary result indicating whether or not *E. coli* was detected in a given sample. The fixed effects terms were a categorical variable for the type of water treatment available at a campground and a binary variable indicating whether domestic ruminants (cattle and sheep) were present or absent in the campground water catchment. Water treatment was categorised into three groups: no treatment, filter-only and other types of treatment (combination of filter and ultra violet (UV), UV-only and combination of filter and chemical) (Table 5.1). Campground was introduced in the GLMM as a random effects term to account for the fact that samples collected from the same campground were more likely to be similar than samples from different campgrounds. The threshold for the outcome variable, i.e. *E. coli* concentration, in the GLMM model was set to reflect the compliance threshold as determined by the DWSNZ 2008, which require that no *E. coli* should be detected in treated water (tap water). Water treatment and presence/absence of ruminants in the campground catchment were study factors (factors of interest) hence were introduced in the model regardless of their statistical significance in the variable selection process. Another GLMM with campground as a random effects term was applied to the intake water *E. coli* concentration data. However, in this GLMM the estimated random effects were all zero, therefore, a GLM was used. In this GLM dichotomised *E. coli* concentrations ( $< 200$  MPN/100mL versus  $\geq 200$  MPN/100mL) formed the outcome variable while ruminants (present/absent) and Island (North/South) were the explanatory variables. The 200 MPN/100 mL threshold for *E. coli* concentrations reflected the MAV set for recreational water use by the New Zealand Ministry for the Environment (2003). The recreational water MAV was used because no water quality threshold is set for raw drinking water in DWSNZ 2008.

## 5.4 Results

### 5.4.1 Campground descriptive statistics

There were 220 campgrounds operated by the New Zealand Department of Conservation (DOC) during the study period. Of these, 44.5 % (98) were located in the North Island and 55.5 % (122) in the South Island. The campgrounds were categorised into five classes: backcountry, basic, great walk, serviced and standard. Fifteen campgrounds were recruited into the study (Table 5.1), three of which were classified as serviced and 12 as standard; three were located in forest parks, four in national parks, six in recreation reserves and one in a scenic reserve; nine did not have water treatment facilities, two used filters only, two had UV light treatment, one combined filtration with UV light irradiation while a combination of filter and chemical treatment was used at another campground. The number of camping sites available to campers at each campground ranged from 20 to 850.

**Table 5.1:** Description of study campgrounds operated by the Department of Conservation surveyed for microbial drinking water quality in the 2011/2012 and 2012/2013 summer months, New Zealand.

Campground	Type	Class	Water treatment	Sites	Water supply
<b>North Island</b>					
Anaura Bay	Recreation Reserve	Standard	None	75	Stream
Catchpool	Forest Park	Standard	None	150	Stream
Otaki Forks	Forest Park	Standard	Filter	150	Stream
Port Jackson	Recreation Reserve	Standard	None	130	Stream
Stony Bay	Recreation Reserve	Standard	None	75	Stream
Waikaremoana Motorcamp	National Park	Serviced	Filter/Chemical	59	Stream
Waikawau Bay	Recreation Reserve	Standard	Filter	350	Stream
Whangaiterenga	Forest Park	Standard	None	90	Stream
<b>South Island</b>					
Kerr Bay	National Park	Serviced	UV	20	Lake
Kohaihai	National Park	Standard	None	50	Roof
Momorangi Bay	Recreation Reserve	Serviced	None	131	Stream
Pelorus Bridge	Scenic Reserve	Serviced	Filter/UV	46	Stream
Totaranui	National Park	Standard	UV	850	Stream
Whatamango Bay	Recreation Reserve	Standard	None	50	Spring
Whites Bay	Recreation Reserve	Standard	None	20	Stream

### 5.4.2 Geospatial descriptives

The campground water sources ranged from a very small streams, e.g. at Anaura Bay, Catchpool and Whangaiterenga (Figure A.18 on page 192), to a complex network of streams draining into a lake at Kerr Bay (Figure A.19a on page 193). Catchment sizes ranged from 0.1 km<sup>2</sup> at Totaranui Unnamed to 186.9 km<sup>2</sup> at Kerr Bay. Eleven (out of 14) surface water catchments had 85 % or more area covered with indigenous forest (Figures A.18 and A.19).

### 5.4.3 Water samples

In total, 104 water samples were collected (Table 5.2), 51 were from abstraction points (intakes) while 53 were from taps. Only tap samples were collected from Kohaihai and Whatamango Bay campgrounds because they had roof and spring water sources, respectively. Source water was not collected at Pelorus Bridge in the last two rounds because of restricted access to the water abstraction site due to logging activities in the area.

#### *E. coli*

Enumeration of *E. coli* was conducted in 102 (49 intake and 53 tap) water samples (Table 5.2) — enumeration was not performed in two samples, one from Pelorus Bridge intake and another from the Totaranui Unnamed intake. Intake samples had a higher median MPN of *E. coli* organisms/100mL, 38 (range: 0; 2800), than tap samples, 1 (range: 0; 1900). The recreational water maximum acceptable value (MAV) of 200 MPN/100mL of *E. coli* organisms was exceeded in 14 (29%) of the 49 intake water samples. Intake samples from three campgrounds, Waikaremoana Motorcamp, Kerr Bay and Pelorus Bridge, did not have detectable concentrations of *E. coli*. Among the campgrounds with detectable concentrations of *E. coli* in intake water samples, Port Jackson had the highest median MPN followed by Anaura Bay, Stony Bay, Waikawau Bay and Catchpool (Figure 5.4). Overall, 30/53 (57%) of the tap samples exceeded the DWSNZ regulatory MAV of no detectable *E. coli* in tap water. However, no *E. coli* was detected in tap samples from three campgrounds: Waikaremoana Motorcamp, Kerr Bay and Totaranui. In contrast, tap water samples from five other campgrounds (Catchpool, Port Jackson, Stony Bay, Waikawau Bay and Whites Bay) had one or more *E. coli* organisms per 100 mL of tap water on all sampling occasions.

#### *Campylobacter*

Overall, 9.6% (10/104) of the water samples tested *Campylobacter*-positive using PCR. By sample source, 14% (7/51) intake and 6% (3/53) tap samples were *Campylobacter*-positive. Of the 104 samples, two (1.9%) were positive for *C. sp. nova* I, one (1%) *C. coli*, six (5.8%) *C. jejuni* and four (3.8%) *Campylobacter* spp.. One sample tested positive for both *C. sp. nova* I and *C. jejuni*, i.e. one of the two sample isolates tested positive for *C. sp. nova* I and the other *C. jejuni*; two samples tested positive for *C. jejuni* and *Campylobacter* spp.. The seven positive intake samples were collected from: Catchpool (one *C. jejuni*), Otaki Forks (one *C. jejuni*), Pelorus Bridge (one *C. jejuni*/*Campylobacter* spp.), Port Jackson (one *Campylobacter* spp.) and three from Stony Bay (one *C. sp. nova* I/*C. jejuni* and two *C. jejuni*). The three positive tap samples were collected from Whatamango Bay (one *C. jejuni*/*Campylobacter* spp.) and Stony Bay (one *C. sp. nova* I and one *C. jejuni*).

#### *Cryptosporidium* and *Giardia*

Flourescence microscopy for *Cryptosporidium* and *Giardia* was conducted on 44 (24 intake and 20 tap) water samples. None of these samples were positive for *Cryptosporidium*. All



**Table 5.2:** Number of water samples collected from campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand. Also shown are the number of *E. coli*-positive samples collected from the tap and intake.

Campground	Tap ( $\geq 1$ MPN/100mL)		Intake ( $\geq 200$ MPN/100mL)	
	+ve	Total	+ve	Total
<b>North Island</b>				
Anaura Bay	3	4	3	4
Catchpool	4	4	1	4
Otaki Forks	2	4	1	4
Port Jackson	3	3	2	3
Stony Bay	3	3	2	3
Waikaremoana Motorcamp	0	4	0	4
Waikawau Bay	4	4	1	4
Whangaiterenga	3	4	1	4
<b>South Island</b>				
Kerr Bay	0	2	0	2
Kohaihai	1	2		
Momorangi Bay	2	4	1	4
Pelorus Bridge	1	4	0	2
Totaranui	0	4	1	8
Whatamango Bay	1	4		
Whites Bay	3	3	1	3
Total	30	53	14	49

tap samples were negative for *Giardia* while two of the 24 (8%) intake samples were positive for *Giardia*. One of the *Giardia*-positive samples was collected from Waikawau Bay while the other was collected from Whangaiterenga.

#### 5.4.4 Faecal samples

A total of 668 faecal samples were collected, 76.0% (508/668) were from wild birds (duck, passerine, pukeko, seagull, shag, sparrow, starling, swan, weka), 15.7% (105/668) from wild mammals (deer, possum, rabbit, rodent) and 7.9% (53/668) from domestic ruminants (cattle, sheep) (Table 5.3).

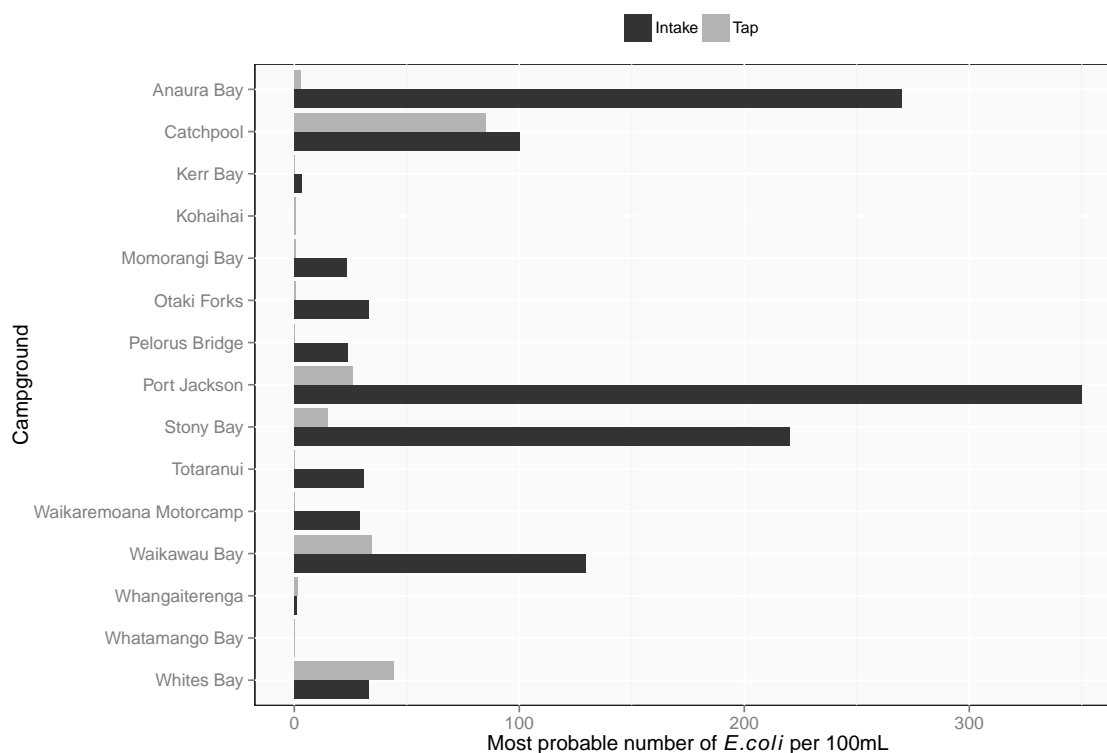
#### *Campylobacter*

A total of 206 faecal samples had *Campylobacter*-like growth on blood agar plates. Of these, 29.6% (61) were PCR-positive for *C. sp. nova* I, 7.3% (15) were positive for *C. coli* and 29.1% (60) were positive for *C. jejuni*. *C. sp. nova* I/*C. jejuni* mixed PCR results were obtained in six samples.

#### *Cryptosporidium* and *Giardia*

Flourescence microscopic examination for *Cryptosporidium* and *Giardia* was conducted on 612 faecal samples. No sample was positive for *Cryptosporidium* while 2.5% (15) samples were positive for *Giardia*. Of the fifteen *Giardia*-positive samples, twelve (all from ducks





**Figure 5.4:** Median most probable number of *E. coli* in intake and tap water samples collected from campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand.

from Waikaremoana Motorcamp) were subjected to a PCR confirmatory test, eleven of which returned positive test results. The PCR-positive samples were sequenced to determine their assemblage. Three were assemblage AII, one assemblage BII and seven assemblage BIV.

#### 5.4.5 Multilocus sequence typing analysis

A subset of 50 (43 faecal and 7 water) *Campylobacter*-positive samples were strain-typed using the MLST scheme. The strains are presented in Table 5.4 and are stratified by sample source.

#### Faecal isolate sequence types

Six (12%) of the 50 isolates on which MLST was performed were from cattle (domestic ruminants) and 74% (37) were from wild birds (passerines, pukeko and seagulls). All the isolates from cattle, passerines and seagulls had sequence type (ST) already known (existing in the PubMLST database) while about half of the pukeko isolates had STs that have not been reported before.

**Table 5.3:** Number of faecal samples, stratified by animal source, collected from campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand.

Campground	Cattle	Deer	Dog	Duck	Passerine	Possum	Pukeko	Rabbit	Rodent	Seagull	Sheep	Weka	Total
<b>North Island</b>													
Anaura Bay	0	0	0	0	52	2	0	0	0	0	0	0	54
Catchpool	2	0	2	0	11	20	0	2	2	0	0	0	39
Otaki Forks	0	0	0	0	35	22	0	0	0	0	0	0	57
Port Jackson	41	0	0	1	7	0	0	0	0	0	0	0	49
Stony Bay	2	0	0	8	9	0	26	3	0	0	8	0	56
Waikaremoana Motorcamp	0	5	0	40	11	3	0	3	0	0	0	0	62
Waikawau Bay	0	0	0	3	22	9	11	1	0	0	0	0	46
Whangaiterenga	0	0	0	0	20	1	0	6	0	0	0	0	27
<b>South Island</b>													
Kerr Bay	0	0	0	10	2	0	0	2	0	15	0	0	29
Kohaihai	0	0	0	0	6	0	0	0	0	0	0	5	11
Momorangi Bay	0	0	0	12	35	3	0	6	0	0	0	0	56
Pelorus Bridge	0	0	0	0	39	3	0	0	0	0	0	0	42
Totaranui	0	1	0	11	16	10	25	1	0	0	0	0	64
Whatamango Bay	0	0	0	7	15	0	32	0	0	0	0	0	54
Whites Bay	0	0	0	0	22	0	0	0	0	0	0	0	22
Total	45	6	2	92	302	73	94	24	2	15	8	5	668

### Water isolate sequence types

All the isolates from water (intake and tap) samples had STs that were already known. However, two stream water isolates had new, previously unreported allele sequences and allelic profiles that were similar to ST-2381 and ST-3672 allelic profiles. Table 5.5 shows the strains isolated from water and indicates the campground from which the strain originated, previous reported hosts and whether or not the strain had previously been associated with illness in humans. The strain (ST-538) isolated from the Catchpool water source has previously been reported to cause human infection while the remaining strains are associated with wild birds and are potential human pathogens.

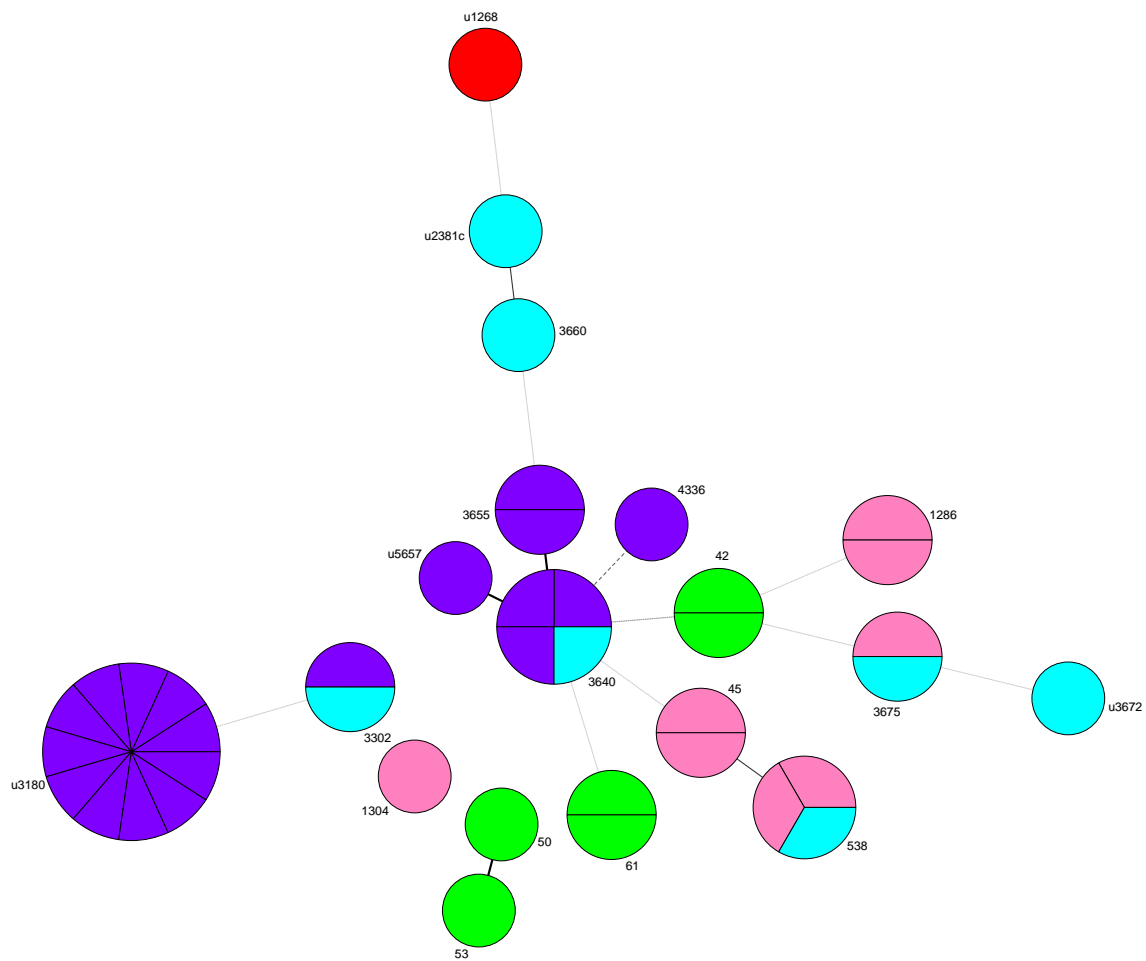
### Minimum spanning tree

Figure 5.5 shows clustering of 41 isolates that had adequate ST information for the analysis to be performed. The number of sectors in each pie (node) correspond to the number of isolates while the thickness of the lines connecting the nodes corresponds to the similarities between the STs.

### 5.4.6 Regression analysis

#### Faecal *Campylobacter*

The faecal *Campylobacter* GLM results are displayed in Table 5.6. The intercept indicates that the odds of a faecal sample from a passerine located in the South Island returning



**Figure 5.5:** Minimum spanning tree of *Campylobacter jejuni* and *C. coli* isolated from campgrounds operated by the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand. Colour codes: blue = water, green = cattle, pink = passerine, purple = pukeko, red = seagull. The node size is proportional to the number of isolates while the thickness of the connecting lines is proportional to the similarities between the sequence types.

a *Campylobacter*-positive test were 0.19. Faecal samples from New Zealand rails (pukeko and weka)<sup>4</sup> were 35 times more likely to return a *Campylobacter*-positive test compared to faecal samples from passerines, regardless of the Island from which they were collected. In contrast, faecal samples from ducks were nine times less likely to be *Campylobacter*-positive compared to those from passerine, regardless of the island in which they were collected. The odds of a faecal sample returning a *Campylobacter*-positive result seemed to increase by 15 % if the faecal samples was collected from the South Island compared to the odds of a sample collected from the North Island after controlling for sample source, however, this result was statistically non-significant at  $P$  value=0.05.

### Water *E. coli*

The tap water *E. coli* GLMM results are displayed in Table 5.7. The intercept indicates that the odds of a water sample from a tap with no water treatment at a campground without ruminants in the water catchment having an *E. coli* concentration  $\geq 1$  MPN/100 mL were 3.61. A sample from a tap whose water was treated using UV or a combination of filtration and UV or chemicals was 100 (i.e.  $\frac{1}{0.01}$ ) times less likely to have detectable concentrations of *E. coli* compared to a sample from a tap without any treatment, regardless of whether or not domestic ruminants were present in the campground water catchment. In contrast, there was a statistically non-significant (at  $P$  value=0.05) 2 % increase in the odds of obtaining a water sample with detectable concentrations of *E. coli* if a tap had a filter-only treatment system installed compared to a sample from a tap without any treatment, regardless whether or not domestic ruminants were present in the campground water catchment. The intake water *E. coli* GLM results are displayed in Table 5.8. Both ruminants (present/absent) and Island (North/South) were statistically non-significant at  $P$  value=0.05, however, there appeared to be an increase in the odds of obtaining a water sample with *E. coli* concentrations  $\geq 200$  MPN/100 mL from the intake if ruminants were present in the catchment compared to when ruminants were absent regardless of the island in which the campground was located.

## 5.5 Discussion

The current study quantified the public health risk associated with drinking water supplied at fifteen campgrounds managed by the Department of Conservation (DOC) in New Zealand. By coinciding sample collection with the peak camping season, the public health risk was estimated during a period when the highest number of members of the public engaged in outdoor activities were most likely to be exposed to drinking water-related risk at the campgrounds. The public health risk assessment was achieved by examining the occurrence of four organisms associated with water quality (*Campylobacter*, *E. coli*, *Cryptosporidium*

<sup>4</sup>Also known as purple swamphen (*Porphyrio porphyrio*) and Maori hen or woodhen (*Gallirallus australis*), respectively.

and *Giardia*) in campground drinking water. Detection and analysis of these organisms at the point of water abstraction offered insight in the level of pollution in the source water and gave an indication of possible polluting animals in the catchment. While the microbial content of tap water provided a framework for assessing the direct public health risk. Tap water microbial analysis also provided an indication of how effective the available water treatment methods were in making the water microbiologically safe for human consumption.

The four microbes used for public health risk assessment in the current study are recommended for determining water quality in New Zealand (New Zealand Ministry for the Environment, 2003; New Zealand Ministry of Health, 2008). Elsewhere the four organisms have been used in previous research to perform risk assessment related to drinking water (Dechesne and Soyeux, 2007; Ferguson et al., 2007). Conventional cultured-dependent laboratory techniques were employed in the detection and identification of the microbes under investigation. The laboratory techniques used included phenotypic identification, microscopy, polymerase chain reaction (PCR) and multilocus sequence typing (MLST).

Analysis of data was conducted using a generalised linear modelling approach, an established statistical analysis framework. Specifically, a generalised linear mixed model (GLMM) in which campground was the random effects term was used to identify animals likely to contaminate DOC-operated campground drinking water catchments. A similar GLMM was used to estimate the potential public health risk posed by tap water provided at the campgrounds. This was achieved by designing the modelling process so that there was indication whether or not the regulatory drinking water quality threshold was exceeded or not in tap water. Modelling campground as a random effect meant that the results of the model could be extrapolated to other campground operated by DOC. In addition, MLST data analysis provided information on the pathogenicity and potential pathogenicity of the *Campylobacter* strains isolated from drinking water at the campgrounds.

Summarised *E. coli* data showed that on more than half of the sampling occasions the drinking water provided by the campgrounds was not compliant with drinking water standards for New Zealand (DWSNZ) regulations. However, tap water from three campgrounds (Waikaremoana Motorcamp, Kerr Bay and Totaranui) complied with the regulations on all sampling occasions. In contrast, five campgrounds (Catchpool, Port Jackson, Stony Bay, Waikawau Bay and Whites Bay) had tap water that did not comply with the regulations on all sampling occasions. All the campgrounds that were not compliant with DWSNZ regulations on all occasions did not have any water treatment facilities installed, except for Waikawau Bay which had a filtration-only system. Evidence from the regression analysis revealed that water treatment was necessary for ensuring that tap water was compliant with the regulations. However, treatment that involved filtration only was ineffective in eliminating microbial contamination. This means that in order for campgrounds to supply

drinking water that is microbiologically safe and compliant with DWSNZ regulations, water treatment systems should have a combination of filtration and other water treatment methods such as ultra violet (UV) or chemical treatment.

The isolation of *C. jejuni* and *C. coli* in tap water samples not only provided evidence of faecal contamination but also exposed a potentially serious public health risk. The public health significance of the presence of *C. sp. nova* I in drinking water is not known. Both Whatamango Bay and Stony Bay campgrounds, where *C. jejuni* and *C. coli* were isolated from tap water samples, did not have any water treatment installed. These findings suggest that remedial measures at the two campgrounds are urgently required. The remedial measures could include, for instance, installation of filtration and UV treatment facilities. Further, a spring supplied water to the Whatamango Bay campground and detection of *Campylobacter* in tap water implies that the spring was unsecured or there was infiltration into the system, possibly due to ageing pipes. Hence remedial measures could include securing the spring and/or replacement of the pipe infrastructure. Ample examples are available of drinking water-related gastrointestinal illness outbreaks at outdoor recreational facilities providing untreated (Boccia et al., 2002; Nygård et al., 2004) or inadequately treated (Arvelo et al., 2012; Bohmer, 1997; Hewitt et al., 2007; Waarbeek et al., 2010) potable water. Therefore, precautionary measures should be seriously considered.

Regression analysis revealed that rails (pukeko and weka) were the most likely sources of *Campylobacter* organisms within and around the campground water catchments. Further, evidence from the MLST data show that pukeko and ducks were most likely to pollute water sources at DOC-operated campgrounds with *Campylobacter*. Although *Campylocater* strains isolated from birds were not known to be zoonotic, in that they had not previously been reported as causing human illness, their close relationship with known pathogens presents a potential public health risk. Another potential public health risk from wild birds exposed in the present study is that of *Giardia* assemblages A and B isolated from ducks. This study is the first to report these assemblages in wild birds in New Zealand although the assemblages are known to be human pathogens (Learmonth et al., 2003; Winkworth, 2010; Winkworth et al., 2008). Elsewhere, studies have also reported that *Giardia* assemblages A and B are zoonotic and are predominantly found in humans (Amar et al., 2002; Feng and Xiao, 2011; Lalle et al., 2005).

Limitations in the current study include the fact that a small number of observations were available for both intake (49 observations) and tap (53 observations) water regression analysis. At campground level, there were only 2–4 samples from either the intake or tap. The small number of observations may have adversely affected the models' ability, for example, to effectively estimate the relationship between the microbial quality of water and the presence of ruminants in the campground water catchment. Another limitation is that fresh

samples were required for the isolation of microbes especially *Campylobacter*. On rainy days it was, however, difficult to distinguish between fresh faeces and old ones that were moistened by rain. This could have lead to under-detection of microbes in faecal samples. The strengths of the study include the fact that the direct public health risk associated at the study campgrounds was estimated. The study also identified measures that can be used to effectively mitigate the risk.

In summary, the current study established the public health risk associated with drinking water on DOC-operated campgrounds. Campgrounds that posed the highest public health risk were identified. Remedial measures proposed for these campgrounds include protection of drinking water sources, where possible, and installation of water treatment facilities such as a combination of filtration and UV or chemical treatment. With ample examples of gastrointestinal illness outbreaks at outdoor recreational facilities elsewhere, the need for providing microbiologically safe drinking water at campgrounds at all times cannot be overemphasized.

**Table 5.4:** Multilocus sequence types for a subset of *Campylobacter* isolates ( $n = 50$ ), stratified by sample source, for faecal and water samples collected from the Department of Conservation-operated campgrounds, 2011/2012 and 2012/2013 summer seasons, New Zealand.

ST	Cattle	Passerine	Pukeko	Seagull	Stream	Tap	Total
<b>Known sequence types</b>							
ST-42	2						2
ST-45		2					2
ST-50	1						1
ST-53	1						1
ST-61	2						2
ST-538		2			1		3
ST-1286		2					2
ST-1304		1					1
ST-3302			1			1	2
ST-3640			3		1		4
ST-3655			2				2
ST-3660						1	1
ST-3675		1			1		2
ST-4336			1				1
Total	6	8	7	0	3	2	26
<b>New allelic profiles<sup>§</sup></b>							
ST-1275				1			1
ST-2381					1		1
ST-3640			1				1
ST-3672					1		1
ST-5609		1					1
Total	0	1	1	1	2	0	5
<b>Incomplete typing<sup>†</sup></b>							
ST-1275				3			3
ST-1324		1					1
ST-3655			2				2
Total	0	1	2	3	0	0	6
<b>Unknown sequence types</b>							
NEW			13				13
Total	0	0	13	0	0	0	13
Grand Total	6	10	23	4	5	2	50

<sup>§</sup>New allelic profiles, but most similar to the five listed STs

<sup>†</sup>STs with missing alleles, but most similar to the three listed STs



**Table 5.5:** Multilocus sequence types for *Campylobacter* isolated from water samples collected from the Department of Conservation-operated campgrounds, 2011/2012 and 2012/2013 summer seasons, New Zealand. Also shown are the hosts from which the strains have previously been isolated and whether the strain has previously caused human infection (Yes) or not (Possible).

Source	Campsite	Species	Strain	Known host	Human pathogen
Stream	Catchpool	<i>C. jejuni</i>	ST-538	Human/chicken/cattle	Yes
Stream	Otaki Forks	<i>C. jejuni</i>	ST-3675	Dotterel/duck/passerine	Possible
Stream	Pelorus Bridge	<i>C. jejuni</i>	ST-3672 New	Dotterel/duck	Possible
Stream	Stony Bay	<i>C. jejuni</i>	ST-2381 New	Pukeko	Possible
Stream	Stony Bay	<i>C. jejuni</i>	ST-3640	Pukeko	Possible
Tap	Stony Bay	<i>C. coli</i>	ST-3302	Pukeko	Possible
Tap	Whاتمango Bay	<i>C. jejuni</i>	ST-3660	Pukeko	Possible

**Table 5.6:** Generalised linear mixed model estimating the presence of *Campylobacter* in faecal samples collected from the Department of Conservation-operated campgrounds during 2011/2012 and 2012/2013 summer seasons, New Zealand.

Variable	Odds ratio	Lower 95%CI*	Upper 95%CI*	P value	Observations
Intercept	0.19	0.13	0.28	0.00	
<b>Faecal source</b>					
Passerine	1.00				302
Duck	0.11	0.02	0.36	0.00	92
Rails	34.81	18.36	72.24	0.00	99
Ruminants	1.22	0.53	2.61	0.62	53
Seagull	1.67	0.44	5.26	0.41	15
<b>Island</b>					
North	1.00				309
South	1.15	0.67	1.95	0.61	252

\*Confidence interval

**Table 5.7:** Generalised linear mixed model estimating the concentrations of *E. coli* above or below 1 MPN per 100 mL in tap water samples collected from the Department of Conservation-operated campgrounds, 2011/2012 and 2012/2013 summer seasons, New Zealand.

Variable	Odds ratio	Lower 95%CI*	Upper 95%CI*	P value	Observations
Intercept	3.61	0.73	17.72	0.11	
<b>Water treatment</b>					
None	1.00				31
Filter only	1.02	0.07	13.94	0.99	8
Other types	0.01	0.00	0.32	0.01	14
<b>Ruminants in catchment</b>					
Absent	1.00				37
Present	1.28	0.15	11.03	0.82	16

\*Confidence interval

**Table 5.8:** Generalised linear model estimating the concentration of *E. coli* above or below 200 MPN per 100 mL in intake water samples collected from the Department of Conservation-operated campgrounds, 2011/2012 and 2012/2013 summer seasons, New Zealand.

Variable	Odds ratio	Lower 95%CI*	Upper 95%CI*	<i>P</i> value	Observations
Intercept	0.48	0.19	1.19	0.11	
<b>Ruminants in catchment</b>					
Absent	1.00				37
Present	1.74	0.42	7.23	0.45	12
<b>Island</b>					
North	1.00				30
South	0.37	0.08	1.61	0.18	19

\*Confidence interval



# Six

## The metagenome of drinking water on campgrounds in New Zealand

### 6.1 Background

Microorganisms, including bacteria and archaea, are ubiquitous and play an important role in the ecosystem of living organisms, for they are not only primary sources of nutrients but also primary converters of dead matter into a form that can be readily utilised by higher forms of life. Human life has a profound relationship with microorganisms that includes both positive and negative aspects. Examples of positive aspects of the human-microorganism relationship include the fact that in the human gut microbes are able to digest substances that are not digested by human enzymes (Guarner and Malagelada, 2003; Stevens and Hume, 1998), and that the growth of harmful bacteria is often suppressed by non-harmful microbes (Guarner and Malagelada, 2003). Further, Neish (2009) reported that microbes play an important role in the modulation of the human immune system. Another example of the positive aspect of the relationship is that for many generations microbes have been used in the production of medicine for human use (Garrod, 1960). One of the major negative effects of microorganisms on humans is causing illness. Diseases in which microbes are implicated range from infectious diseases like campylobacteriosis (Bohmer, 1997; Muellner et al., 2013; Nelson, 2010; Sheppard et al., 2009) to non-communicable conditions like cancers (Blaser, 2008; Karnes and Usatine, 2014; Morales-Sánchez and Fuentes-Panana, 2014).

It is generally accepted that to better understand the role of microbes in the complex human-microorganism relationship, it is necessary to undertake a holistic approach such as a genomic study, as illustrated by the human microbiome project (Huttenhower et al., 2012; Qin et al., 2010; Turnbaugh et al., 2007). This approach has been used to study other ecosystems beyond the human body e.g. in drinking water (Bai et al., 2013; Gomez-Alvarez et al., 2012), faeces (Hand et al., 2013; Oikonomou et al., 2013), river water (Ghai et al., 2011), soil (Daniel, 2005; Sangwan et al., 2012; Zablocki et al., 2014), permafrost (MacKellprang et al., 2011) and sea (Kerkhof and Goodman, 2009; Venter et al., 2004). The genomic holistic approach has been employed to answer the *who?*, *what?* and *how?* questions. Answering the *who?* question leads to the identification of organisms present in an ecosystem

being studied while the *what?* question deals with what the organisms do. How the organisms interact with each other and with the environment is dealt with by the *how?* question.

Fundamental to genomic studies is the partial or complete sequencing of individual organism genomes. The sequencing of bacteriophages MS2 (Fiers et al., 1976) and  $\Phi$ X174 (Langeveld et al., 1978) marked the commencement of the study of microbial genomes (Wooley et al., 2010) and since then full genomes of many organisms have been sequenced (Baltrus et al., 2009; Biggs et al., 2011; Blattner et al., 1997; Cole et al., 1998; Qin et al., 2012; Stinear et al., 2008; Tomb et al., 1997) including that of humans (Lander et al., 2001; Venter, 2001). Whole organism genomics provides opportunities for the study of the evolution of not only single genes, but also whole transcriptional units, chromosomes, and cellular networks. However, the single organism genome sequencing approach has limitations, among which is the reliance on cloning cultured organisms. Since only about 1 % of the microbes in nature have been successfully cultured, it implies that full genomic data are highly biased and not suitable for studying microbial communities (Pace, 1997; Rappé and Giovannoni, 2003), which are habited by numerous microbial species. This limitation can be overcome by employing methods that do not require prior cultivation of microbes in order to extract genetic material. In this way, unbiased samples of microbial community members can be studied and inferences made about a given microbial community or population. Genetic material extracted through such unbiased methods is known as a metagenome and the study thereof metagenomics.

Although metagenomic sequencing overcomes some of the limitations encountered by single organism genome sequencing, it has its own challenges. Among the challenges is the fact that a metagenome usually has fragmented sequences or incomplete individual genomes. While in a single organism (or clone) genome study it is certain that all extracted deoxyribonucleic acid (DNA) fragments belong to the same genome, it is not the case with metagenomic studies as environmental samples are composed of many microbial species and it is not possible to determine the true origin of each fragment. Because of a lack of complete species information it is difficult to map individual metagenomic sequences (also known as reads) to their species of origin, therefore, assembly of a full individual genome is highly complex. This means that genomic analysis of metagenomic data is generally limited to a small range of genomic aspects such as single nucleotide polymorphism (SNP)s, short functional signatures and single domain genes (Wooley et al., 2010).

Two main sequencing approaches are employed in metagenomic studies. The first approach is metabarcoding which targets a specific segment of a genome or specific genes such as the 16S ribosomal ribonucleic acid (rRNA) or 18S rRNA genes. The second approach is where an entire metagenome is sequenced and an attempt is made to reconstruct individual organism genomes composing the metagenome. The latter approach is known as whole genome

shotgun (WGS) sequencing or random shotgun sequencing. In shotgun sequencing, DNA is broken up randomly into numerous small segments, which are sequenced to obtain reads. Multiple overlapping reads for the DNA are obtained by performing several rounds of fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequences (contigs). Metabarcoding is useful for microbial community profiling and helps answer the *who?* question. In comparison, the shotgun approach not only provides microbial community profiles but also microbial community functional information which helps answer the *what?* and *how?* questions. Overall, metagenomics provides the ability for one to study the relationship between microbes and their communities on the one hand and habitats in which they live on the other.

The current status of the application of metagenomic techniques in the drinking water industry was assessed by conducting a systematic search for peer-reviewed research literature using the search term: ((metagenom\* OR metabiom\*) AND ('drinking water' OR drinking-water OR freshwater OR groundwater)). This search term resulted in retrieval of 166 articles through the Scopus search engine and 222 articles through the Web of Knowledge search engine. The two sets of retrieved articles were combined and duplicates were removed. Of these articles, eleven were considered to be relevant to the topic under review. Relevant articles were those that conducted their investigations along the drinking water supply chain (from the water source to the tap) and applied metagenomic techniques for sample processing and analysis. Of the relevant articles, five used 454 sequencing technology (Delafont et al., 2013; Gomez-Alvarez et al., 2012; Kwon et al., 2011; Oh et al., 2011; Pinto et al., 2012), another five Illumina (Bai et al., 2013; Chao et al., 2013; Oh et al., 2011; Shi et al., 2013; Wang et al., 2012b) and one ABI (Schmeisser et al., 2003). Common to all the studies in the eleven articles was the use of the 16S rRNA gene to describe microbial community profiles at various stages of the drinking water supply system. Additionally, five studies (Bai et al., 2013; Chao et al., 2013; Gomez-Alvarez et al., 2012; Oh et al., 2011; Shi et al., 2013) analysed functional genes using WGS sequences.

The present study used metabarcoding and WGS to test the null hypothesis which stated that microbial communities do not vary with varying habitats. The alternative hypothesis stated that microbial communities vary with varying habitats. The alternative hypothesis implied that microbial communities could serve as signatures for their own habitats. The specific objectives of the study were, firstly, to investigate aquatic microbial community metagenomes associated with drinking water at campgrounds operated by Department of Conservation (DOC) in New Zealand. Secondly, to perform source tracking of microbial metagenomes found in campground drinking water. Thirdly, to recommend appropriate schemes for managing microbial contamination of campground water systems. In order to achieve these objectives, metagenomic (metabarcoding and random shotgun) sequencing and analytical techniques were employed.

## 6.2 Materials and methods

### 6.2.1 Study sites and sample collection

This was a serial cross-sectional study conducted on DOC-operated campgrounds during the 2011/2012 and 2012/2013 summer months (December-February). The study sites and sample collection procedure are described in Section 5.2.1 (page 107). In summary, 15 DOC-operated campgrounds with surface or roof water supply were included in the study. Sample collection was conducted in four rounds over the two summer seasons. Water and faecal samples were collected from each campground. Water samples were collected from both taps and abstraction points while scoop faecal samples, of any origin, were collected from the within and around the campground water catchment. Water samples for metagenomic analyses were aseptically collected; 2L in the first two rounds and 10L in the last two rounds. All samples were stored and transported on ice to molecular epidemiology and public health laboratory (*m*EpiLab) at the Hopkirk Research Institute, Massey University.

### 6.2.2 Laboratory processing

#### Faecal samples

Five scoop faecal samples were randomly selected from each campground for metagenomic analyses in the last two rounds of sampling. DNA extraction was performed using the NucleoSpin<sup>®</sup> Soil kit according to the manufacturers' instructions (List 4 on page 196). Briefly, a pea-size portion of the sample was placed in a NucleoSpin<sup>®</sup> tube containing ceramic beads. The cells in the sample were then lysed by vortexing and contaminants precipitated before the lysate was filtered out. The DNA in the lysate was bound to a column and washed. After drying the column, the DNA was eluted and quantified using a Nanodrop<sup>®</sup> 1000 spectrophotometer (Thermo Fisher Scientific Inc.; Massachusetts, USA) and a Qubit<sup>®</sup> 1.0 fluorometer (Invitrogen Corporation; California, USA) then stored at  $-20^{\circ}\text{C}$ . DNA selected for multiplexed 16S rRNA gene and random shotgun sequencing on the Illumina MiSeq system at New Zealand Genomics Limited (NZGL) had an optical density (OD) of 1.8–2.0 (Nanodrop<sup>®</sup>) and contamination of less than 10 % of both ribonucleic acid (RNA) and protein (Qubit<sup>®</sup>). Validation of the size and concentration of the eluted DNA was performed by comparing to Fosmid Control DNA (40kb; 100 ng/ $\mu\text{L}$ ) via electrophoresis on a 2 % agarose gel.

#### Water samples

Within 48 h of arrival at *m*EpiLab each sample was split into 500 mL subsamples (or 2 L in rounds 3 and 4) and filtered separately. Filtration was done using 0.2  $\mu\text{m}$ -pore, 47 mm-diameter Supor<sup>®</sup>-200 membrane disks (Pall Corporation; New York, USA). After filtration the disks were stored in 50 mL centrifuge tubes at  $-80^{\circ}\text{C}$  or in 2 mL microcentrifuge tubes containing 1 mL RNALater<sup>®</sup> (Ambion, Inc.; Texas, USA) (last two sampling rounds). DNA was extracted from the filter disks using the Metagenomic DNA Isolation Kit for

Water (Epicentre<sup>®</sup>; Wisconsin, USA) or the NucleoSpin<sup>®</sup> Soil kit. The protocol for the Epicentre<sup>®</sup> kit is outlined in List 3 on page 195 while that for the NucleoSpin<sup>®</sup> Soil kit in List 4 on page 196. The resultant DNA was quantified, stored and sent for sequencing as described for faecal samples above.

### 6.2.3 Metagenomic DNA sequencing

#### 16S rRNA gene sequence library preparation

The V4 hypervariable region of the 16S rRNA gene was targeted for sequencing by the Illumina MiSeq system. Sequence libraries were prepared by polymerase chain reaction (PCR) amplification using custom made 16S rRNA gene PCR primers targeted at the regions flanking the V4 hypervariable region. The components of both the forward and reverse PCR primers as well as the PCR conditions used during library preparation are outlined below:

Components of the forward (515F) PCR primer:

1. 5' Illumina adapter sequence.
2. Forward primer pad.
3. Forward primer linker.
4. Forward primer sequence which bound to the 515F region flanking the V4 hyper variable region.

Components of the indexed reverse (806R) PCR primers:

1. Reverse complement of the 3' Illumina adapter sequence.
2. Index sequence.
3. Reverse primer pad.
4. Reverse primer linker.
5. Reverse primer sequence which bound to the 806R region flanking the V4 hyper variable region.

The PCR conditions used were as follows:

1. Holding: one cycle at 95 °C for 300 s.
2. Denaturing: 15 cycles at 95 °C for 15 s.
3. Annealing: 15 cycles at 64 °C for 15 s.
4. Extension: 15 cycles at 72 °C for 60 s.
5. Holding: one cycle at 72 °C for 420 s.

#### Whole genome shotgun library preparation

The WGS sequencing libraries were prepared using the TruSeq method on the MiSeq platform. The method involved 8 main steps (Illumina Inc., 2011):

1. **Input DNA fragmentation:** The genomic DNA was mechanically sheared into fragments ranging from 200–400 base pairs. These fragments had 3' to 5' overhangs.



2. **Fragment end repair:** The overhangs generated during fragmentation were converted into blunt ends. The 3' overhang was removed by using exonuclease while polymerase was used to remove the 5' overhang.
3. **3' end adenylation:** This step was aimed at reducing the rate of chimera formation. A single **A** nucleotide was added to the 3' ends of the blunt fragments and a corresponding single **T** nucleotide on the 3' end of the adapter provided a complementary overhang for ligating the adapter to the fragment.
4. **Indexed paired-end ligation:** During this step multiple indexing adapters were ligated to the ends of the DNA fragments.
5. **Ligation product purification:** Unligated adapters as well as adapters that were ligated to each other were removed during this process. An appropriate size range of the sequence library for clustering was selected.
6. **Product amplification:** This process used PCR to selectively enrich the DNA fragments that had adapter molecules on both ends and to amplify the amount of DNA in the library.
7. **Library validation:** Quality control analysis of the sample library was performed during this process. Also, DNA library templates were quantified in order to create appropriate cluster densities across every lane of every flow cell.
8. **Library pooling:** Multiplexed DNA libraries were normalized to 10 nM and then pooled in equal volumes.

#### 6.2.4 Sequence Data

The output from the Illumina MiSeq system were 16S rRNA gene (hereafter referred to simply as 16S) and WGS reads (sequences), organised in fastq files. In every fastq file, each nucleotide was accompanied by a corresponding quality score. A single fastq file contained tens of thousands to millions of unpaired 16S or WGS reads from a single sample. Since the reads were obtained using a pair of primers (forward and reverse), two fastq files per sample were produced, one for each primer. These unpaired or unaligned raw reads were 150 bases long for 16S and 250 bases long for WGS. Once paired, the 16S reads were expected to have a length of 250 base pairs and 400–500 base pairs long for WGS.

#### Sequence classification

Sequences from next-generation sequencing (NGS) platforms, such as MiSeq, are classified using numerical taxonomical methods before downstream analyses can be performed. For reasons of clarity some terms used in metagenomics are specified here based on the definitions provided by Sneath and Sokal (Sneath and Sokal, 1973; Sokal and Sneath, 1963). *Numerical taxonomy* is the grouping of taxonomic units by numerical methods into taxa

on the basis of their character states. *Taxon* (plural *taxa*) is the taxonomic group of any nature or rank. *Operational taxonomic unit (OTU)s* are the units of study, which could be individual organisms, taxonomic groups such as species, genus and so on. A collection of OTUs make up a taxon. In the current study, taxonomic ranks used were (from highest to lowest): kingdom, phylum, class, order, family, genus, and species. In 16S sequence studies OTUs are commonly composed by clustering reads that are  $\geq 97\%$  similar, although this threshold can be adjusted by the user. The potential consequences of this are threefold: firstly, different species that are  $\geq 97\%$  similar on the sequenced gene(s) are merged resulting in OTUs with multiple species. Secondly, species that have paralogs with  $< 97\%$  similarity are split across multiple OTUs. Thirdly, artifacts including read errors and chimaeras may result in spurious clusters. After clustering, the OTUs are matched to a database in order to be assigned to a species. OTUs that are not matched to a species are flagged as novel or unknown.

### 6.2.5 Data analysis

#### Sequence statistics and quality

The quality of the input sequences was ascertained using standard quality analyses that included the analysis of base call accuracy, base content, sequence quality and sequence lengths. The Q-score (Illumina Inc., 2014), also known as Phred quality score, was the main tool used to assess base call accuracy and sequence quality. The Q-score is the probability that a given base was called incorrectly by the sequencer. It is logarithmically related to the base calling error probability and is defined by Equation 6.1.

$$Q = -10 \log_{10} P \quad (6.1)$$

where  $P$  is the estimated probability that a given base call is incorrect. A higher Q-score indicates a smaller probability that a base was incorrectly called. For example a Q-score of 20 represents a probability of 1 in 100 of an incorrect base call or 99.0% accuracy in the base call. Similarly, a Q-score of 30 represents a probability of 1 in 1000 of an incorrect base call or 99.9% accuracy in the base call. The Q-score was used to determine the quality of each position of any given sequence.

#### 16S sequence processing and analysis

The 16S raw sequences in each sample were first paired (aligned or combined) to form 253 base pair overlapping sequences using fast length adjustment of short reads (FLASH) 1.2.6 (Magoč and Salzberg, 2011) and then they were quality trimmed using SolexaQA 2.2 (Cox et al., 2010). From each sample a maximum of 300 000 aligned sequences were randomly selected as input into the quantitative insights into microbial ecology (QIIME) process. The output of this process included an OTU table, phylogenetic tree, representative sequences,

taxa summary charts and alpha rarefaction curves. The OTU table, phylogenetic tree and the set of representative sequences were the main input files for the **R** package **Phyloseq** (McMurdie and Holmes, 2013) which was used to perform various analyses. The OTU table, formatted as a biological observation matrix (BIOM) file (McDonald et al., 2012), contained the 16S OTUs with their corresponding abundance scores (counts of taxa) on a per-sample basis.

Species or taxa richness of the samples was measured in order to investigate how the number of species (taxa) varied across the sample sources. Three diversity indices were used for this purpose: the Chao1 (Chao, 1984), Shannon (Molles, 2013; Tuomisto, 2010) and Inverse Simpson indices which is derived from the Simpson index (Simpson, 1949; Southwood and Henderson, 2009). The Chao1 index calculates the estimated true species diversity of a sample. The Shannon index quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset. The Inverse Simpson index indicates the effective number of species that is obtained when the weighted arithmetic mean is used to quantify average proportional abundance of species in the dataset.

### Public health hazard assessment using 16S sequences

Initially, a multivariate analysis using canonical correspondence analysis (CCA) was applied to all the 16S metagenomes. This was done in order to investigate whether the 16S sequence abundance scores could be used to determine similarities and/or differences among metagenomes of different origins. Thereafter, the public health hazard associated with drinking water supplied at the campgrounds was assessed using 16S taxa in two ways. Figure 6.1 is a schematic representation of the procedure used to assess the public health significance of the 16S metagenomes found at the campgrounds. The first approach was based on taxa belonging to the Family *Campylobacteraceae*. This bacterial Family was chosen because it includes *Campylobacter* species which are the leading causes of gastrointestinal illness in New Zealand (Environmental Science and Research, 2014). The *Campylobacteraceae* phylogenetic tree was extracted and overlaid with taxa abundance scores according to sample sources. This allowed for the visualisation of the phylogenetic relatedness of the *Campylobacteraceae* taxa from different sources. Then the *Campylobacteraceae* taxa abundance scores, per sample source, were used to calculate proportional similarity index (PSI) and construct a tree using the neighbor-net algorithm (Bryant and Moulton, 2004). The second approach was based on taxa related to drinking water-associated bacterial pathogens recognised by World Health Organization (WHO) (Table 2.1 on page 14). The 16S OTUs (both faecal and water) were queried for eight bacterial genera (*Burkholderia*, *Campylobacter*, *Escherichia*, *Francisella*, *Legionella*, *Leptospira*, *Mycobacterium*, *Salmonella*) in order to retrieve the related taxa. However, no taxa under the *Francisella* and *Salmonella* genera queries were retrieved. This could be that members of the *Francisella* and *Salmonella* genera were misclassified into other genera (refer to Section 6.2.4 for possible explanation).

The sequences corresponding to the identified taxa were retrieved from the representative set and matched against the National Center for Biotechnology Information (NCBI)-nr (Wheeler et al., 2000) database in an automated process using basic local alignment search tool (BLAST) via the internet. This allowed for identification of taxa that have previously been reported elsewhere. Further, the retrieved taxa with the accompanying abundance scores were analysed using PSI. Inclusion of spurious sequences into the PSI analysis was minimised by setting the minimum number of sequences per taxa to twenty.

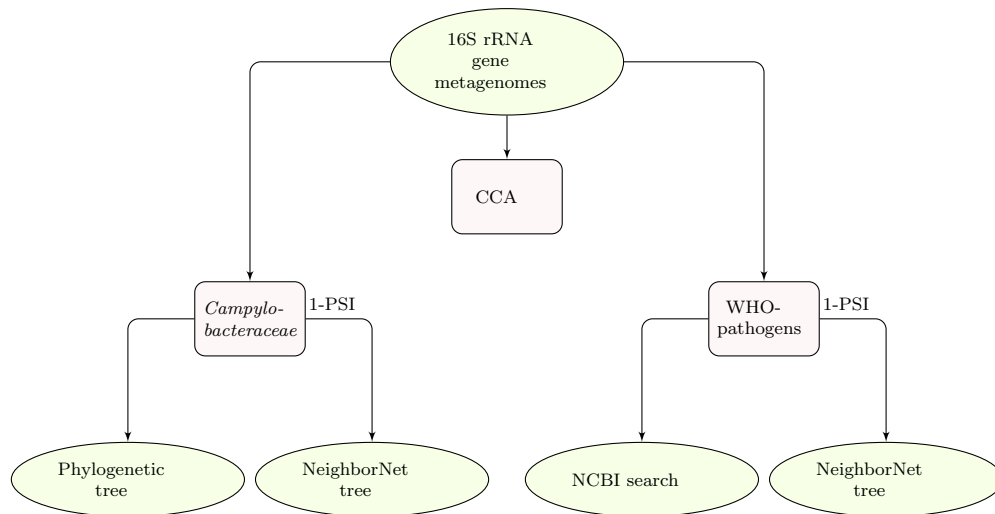
PSI is a measure of similarity that estimates the area of congruence between two frequency distributions (Feinsinger et al., 1981). PSI values range from zero to one, with zero indicating distributions with no common elements and one indicating distributions containing the same elements. The percentile method described by Efron and Tibshirani (1986) was employed to calculate the bootstrapped 95 % confidence intervals for PSI values using 2000 iterations. To demonstrate taxa dissimilarity (divergence) among metagenomes of different origins, values of 1-PSI were used to construct a NeighborNet tree in *Splitstree* 4.13.1.

### WGS sequence processing and analysis

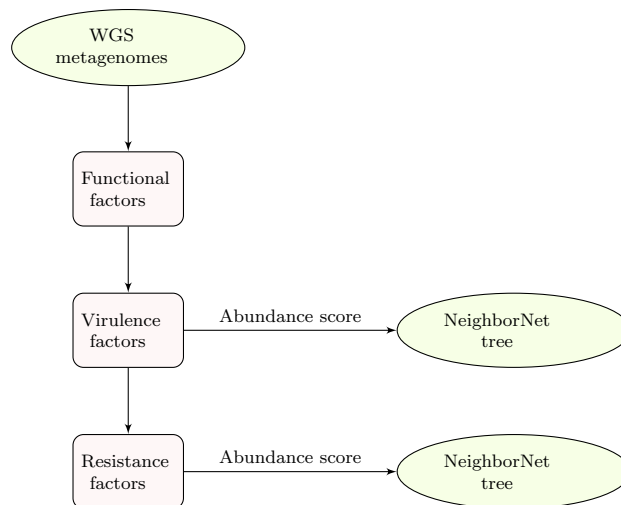
Due to computer resource constraints, only a subset of 20 out of 69 WGS metagenomes were processed using a large computer server available at Massey University. Processing all the 69 WGS metagenomes would have resulted in both the storage capacity and computational capability of a desktop computer to analyse the data being exceeded. However, the WGS sequences in each of the 20 samples were paired using *FLASH* to form 400–500 base pair overlapping sequences which were quality trimmed using *SolexaQA*. The paired sequences were then matched against the NCBI-nr database using protein alignment using a DNA aligner (*PAUDA*) 1.0.1 (Huson and Xie, 2013) in order to assign taxonomic ranks. The functional content of the resultant metagenome files was analysed using the *SEED* classification system (Overbeek et al., 2005) within metagenome analyzer (*MEGAN*) 5.7.0 (Huson et al., 2007). This process resulted in production of abundance scores for each functional factor on a per-sample basis.

### Public health hazard assessment using WGS sequences

Since microbial community profiling of metagenomes was performed using 16S sequences, the WGS analysis was focused on the functional content. Figure 6.2 is a schematic representation of the procedure used to assess the public health significance of WGS metagenomes found at the campgrounds. Among the functional factors identified in the metagenomes, virulence factors were isolated and a tree constructed based on their abundance scores using the neighbor-net algorithm within *MEGAN*. This was done in order to identify similarities/differences among metagenomes of different origins. The same process was repeated for antimicrobial/toxic compound resistance factors (a subset of virulence factors).



**Figure 6.1:** Flow diagram showing how 16S rRNA gene metagenomes were analysed.



**Figure 6.2:** Flow diagram showing how whole genome shotgun metagenomes were analysed.

## 6.3 Results

### 6.3.1 Descriptive statistics

A description of the study campgrounds is provided in Section 5.4.1, therefore, here the summary statistics are limited to those related to metagenomic sequences. A total of 117 (42 faecal and 75 water) samples were successfully sequenced for the 16S rRNA gene (referred to simply as 16S) and 69 (27 faecal and 42 water) samples were also sequenced using the WGS method (Table 6.1). In general, water samples were more evenly sequenced across the study campgrounds than faecal samples.

#### 16S sequences

The median number of raw sequence pairs per sample was  $6.4 \times 10^5$  (95 %CI:  $3.8 \times 10^5$ ;  $1.2 \times 10^6$ ) and these had 94.6 % nucleotides with a Q-score of 30 or more (Figure A.20a). Figure A.20b shows the average Q-score at each of the 150 sequence positions in the raw sequences. These Q-scores indicate that the 16S sequences were of high quality. The median percentage of read pairs combined per sample was 94.2 (95 %CI: 88.6, 97.9) giving a total of  $7.6 \times 10^7$  combined sequences that yielded  $2.4 \times 10^{10}$  nucleotides.

#### WGS sequences

The median number of raw sequence pairs per sample was  $1.5 \times 10^6$  (95 %CI:  $7.5 \times 10^5$ ;  $3.3 \times 10^6$ ) and these had 86.5 % nucleotides with a Q-score of 30 or more (Figure A.21a). Figure A.21b shows the average Q-score at each of the 250 sequence positions in the unpaired reads. A total of  $2.3 \times 10^6$  sequences were obtained from twenty samples whose raw sequences were combined.

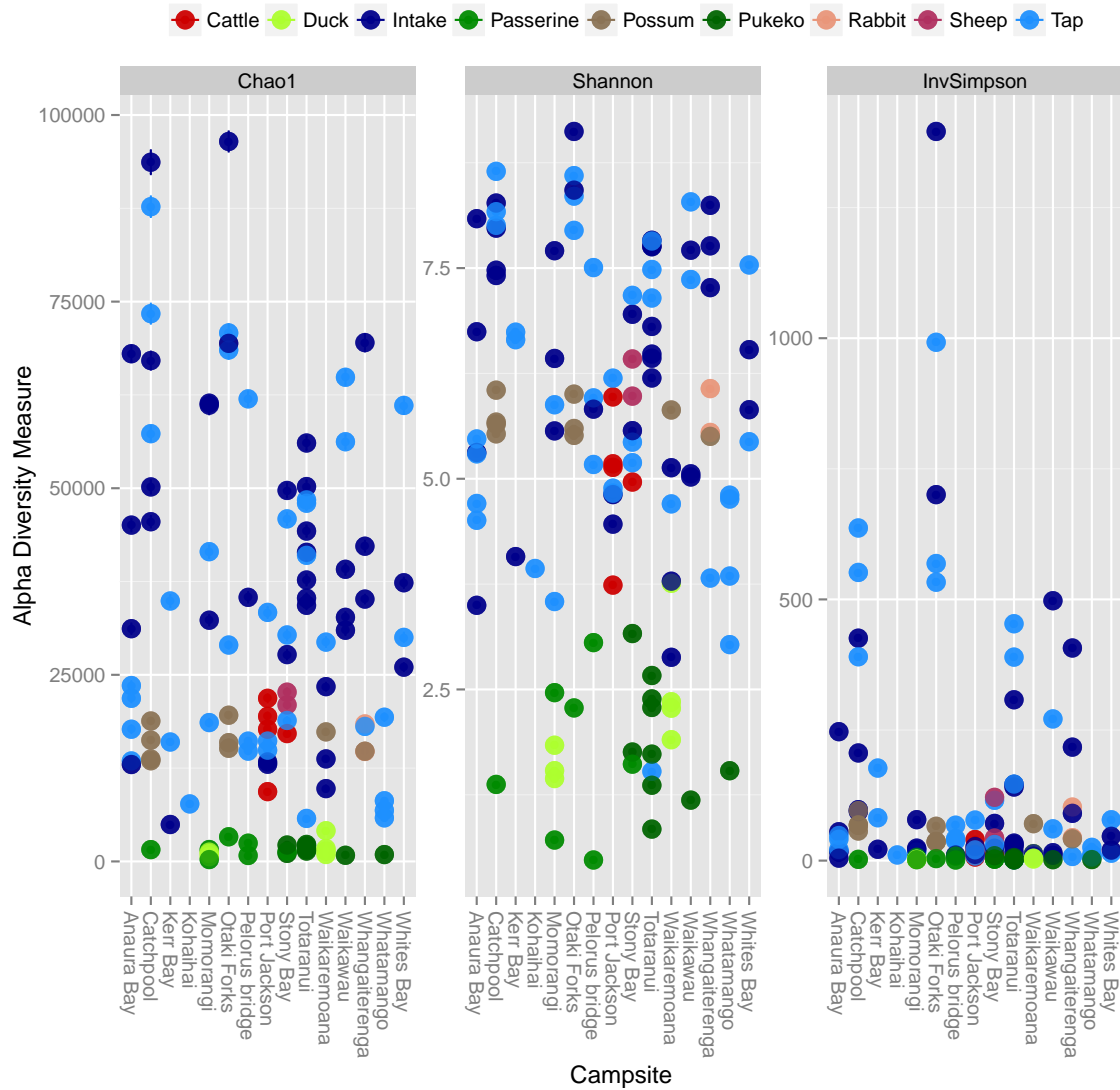
#### Species richness

The species richness indices based on 117 16S metagenomes are summarised in Figure 6.3 and all the indices show that on average water samples had higher richness index scores compared to faecal samples.

### 6.3.2 Public health hazard assessment

#### 16S sequence multivariate analysis

The canonical correspondence analysis (CCA) involving all 117 16S metagenomes revealed clustering based on sample origin (Figure 6.4). Figure 6.4a shows the wild bird (duck, passerine and pukeko) cluster at the top with the domestic ruminant (cattle and sheep) cluster in close proximity below. The cluster on the left is composed of both intake (stream and lake) and tap water metagenomes while wild mammals (possum and rabbit) metagenomes are clustered in the right bottom corner. Figure 6.4b is an enlargement of the water cluster while Figure 6.4c is an enlargement of the wild bird cluster. Subclustering is observable



**Figure 6.3:** Taxa richness indices, stratified by sample source and campground, for 16S metagenomes extracted from samples collected from campgrounds operated by the Department of Conservation surveyed during the 2011,2012 and 2012/2013 summer seasons, New Zealand.

**Table 6.1:** Number of samples sequenced for 16S rRNA gene and whole genome shotgun in a survey conducted on campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand.

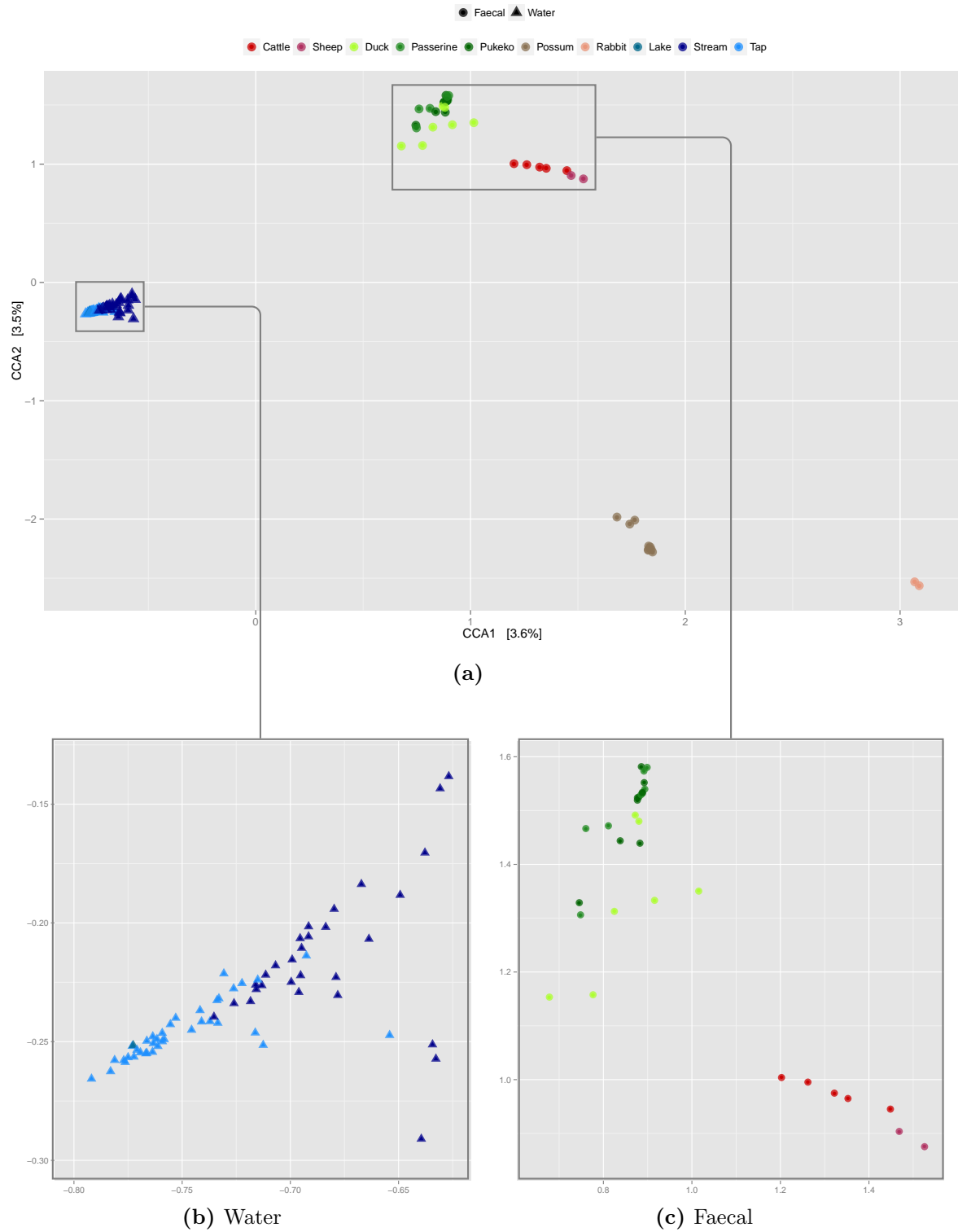
Campground	Duck	Passerine	Pukeko	Possum	Rabbit	Ruminants	Intake	Tap	Total
<b>16S sequencing</b>									
Anaura Bay	0	0	0	0	0	0	4	4	8
Catchpool	0	1	0	4	0	0	4	3	12
Kerr Bay	0	0	0	0	0	0	1	2	3
Kohaihai	0	0	0	0	0	0	0	1	1
Momorangi Bay	3	2	0	0	0	0	3	2	10
Otaki Forks	0	1	0	3	0	0	2	3	9
Pelorus Bridge	0	2	0	0	0	0	1	3	6
Port Jackson	0	0	0	0	0	4	2	3	9
Stony Bay	0	1	2	0	0	3	2	3	11
Totaranui	0	0	6	0	0	0	7	4	17
Waikaremoana	4	0	0	1	0	0	3	1	9
Waikawau Bay	0	0	1	0	0	0	3	2	6
Whangaiterenga	0	0	0	1	2	0	3	1	7
Whathamango Bay	0	0	1	0	0	0	0	4	5
Whites Bay	0	0	0	0	0	0	2	2	4
Total	7	7	10	9	2	7	37	38	117
<b>WGS sequencing</b>									
Anaura Bay	0	0	0	0	0	0	2	3	5
Catchpool	0	0	0	4	0	0	1	1	6
Kerr Bay	0	0	0	0	0	0	1	1	2
Kohaihai	0	0	0	0	0	0	0	1	1
Momorangi Bay	1	2	0	0	0	0	1	0	4
Otaki Forks	0	0	0	3	0	0	1	2	6
Pelorus Bridge	0	1	0	0	0	0	0	1	2
Port Jackson	0	0	0	0	0	3	2	2	7
Stony Bay	0	1	1	0	0	3	1	2	8
Totaranui	0	0	3	0	0	0	3	1	7
Waikaremoana	1	0	0	1	0	0	3	0	5
Waikawau Bay	0	0	0	0	0	0	2	1	3
Whangaiterenga	0	0	0	1	2	0	2	2	7
Whathamango Bay	0	0	0	0	0	0	0	3	3
Whites Bay	0	0	0	0	0	0	1	2	3
Total	2	4	4	9	2	6	20	22	69

within these clusters e.g. the tap water metagenomes are located more towards the lower left corner with intake water metagenomes spreading towards the upper right corner. This is an indication that microbial community profiles can be used to identify the origin of the microbes in a given environment based on their taxa abundance scores.

### 16S *Campylobacteraceae* phylogeny

The *Campylobacteraceae* phylogenetic tree, overlaid with abundance scores, provides more evidence of taxa clustering by metagenome origin (Figure 6.5). Taxa from water metagenomes are predominant in the top and bottom branches of the tree while the middle branches are occupied predominantly by taxa from faecal metagenomes. Further, the top branches are composed of the genus *Arcobacter* and the middle branches are composed of the genus *Campylobacter*. The genus *Sulfurospirillum* is predominant in the bottom branches. The five most abundant *Arcobacter* taxa, starting with the most abundant, included denovo422499, denovo258834, denovo390851, denovo288261 and denovo14187 while denovo401932,





**Figure 6.4:** Canonical correspondence plot (a) showing clustering of 117 metagenomes based on 16S rRNA gene sequences extracted from faecal and water samples collected from campgrounds operated by the Department of conservation surveyed during the 2011/2012 and 2012/2013 summer seasons, New Zealand. (b) and (c) are enlargements of the water (intake and tap) and faecal (wild bird and ruminant) clusters, respectively.

denovo141548, denovo179758, denovo89995 and denovo279614 were the most abundant *Campylobacter* taxa. The abundance scores for the five most abundant *Arcobacter* taxa ranged from 24 to 9646 while the abundance score range for the five most abundant *Campylobacter* taxa was 108–15 727. The *Sulfurospirillum* taxa that had abundance score greater than ten were denovo414613 (590) and denovo151428 (46).

### 16S NCBI database matches

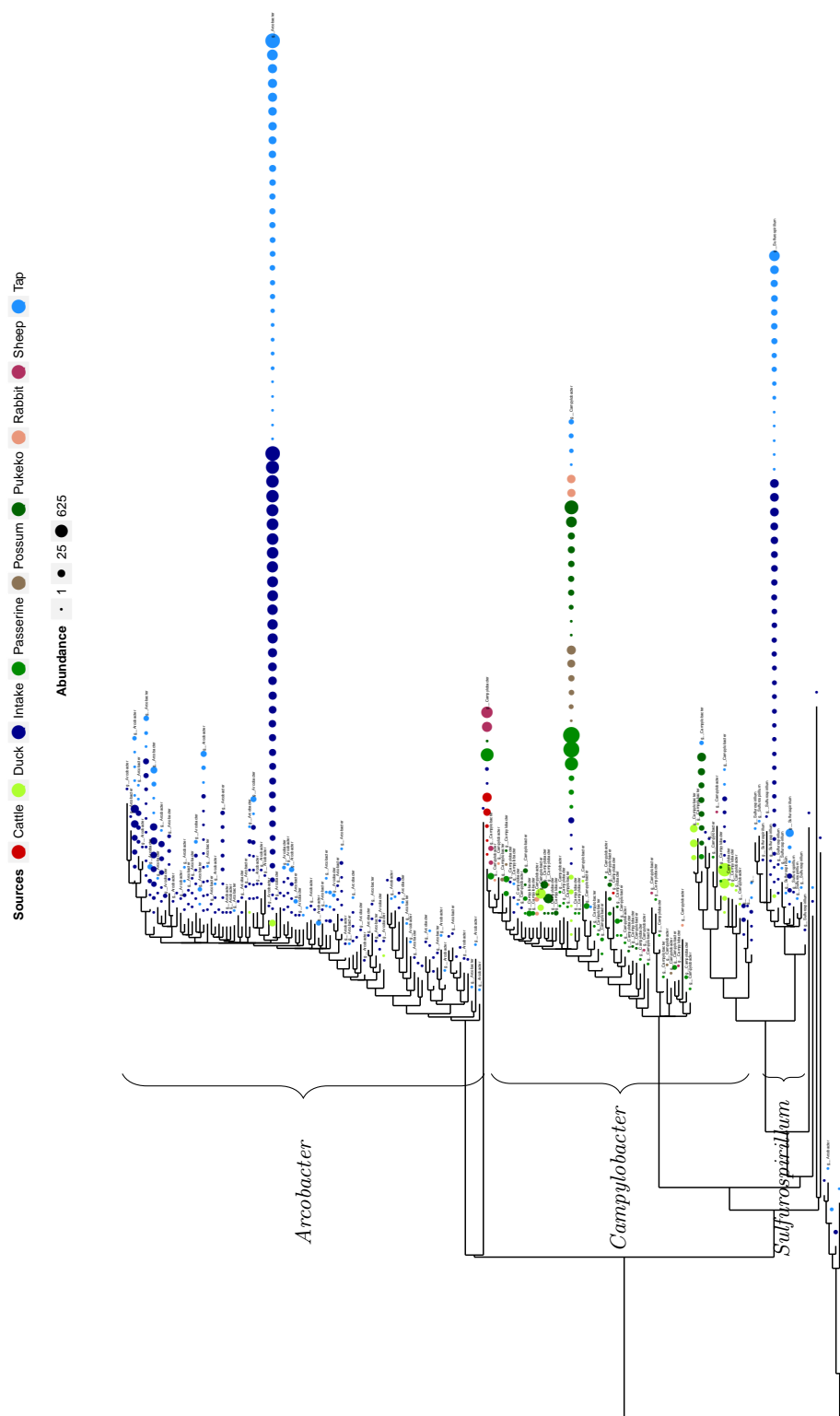
Taxa related to WHO-recognised pathogens associated with drinking water that matched bacterial species found in the NCBI-nr database are shown in Table 6.2. All the matches (hits) had 100 % similarity except for the *Legionella* sp. D2863-denovo297986 match which had 99 % similarity. Some taxa sequences matched multiple species in the database e.g. denovo256673 matched *Burkholderia diffusa*, *B. gladioli*, *B. sp.* BDU8 and *B. sp.* I12B-02616. Among the matched bacterial species were those that had been isolated in human clinical patients while others were novel species. Examples of bacteria previously isolated from human patients include *B. gladioli*, which is generally considered to be a plant pathogen. However, Segonds et al. (2009) reported that this organism was isolated from eighteen French patients, some of whom had cystic fibrosis. *Mycobacterium austroafricanum* is a non-tuberculosis species that has been reported to cause opportunistic infection in humans and has previously been isolated from water in South Africa (Croce et al., 2014; Tsukamura et al., 1983). Another non-tuberculosis mycobacterium that has been isolated from both human patients and water is *M. lentiflavum* (Marshall et al., 2011). Marshall and co-workers reported that *M. lentiflavum* was isolated from 36 patients, among whom were four that exhibited clinical illness and that *M. lentiflavum* was cultured in 13 of 206 samples from drinking water sites in Brisbane, Australia. In a study by Van Ingen et al. (2008), *M. simiae* was isolated from 28 Dutch patients although no clinical illness was observed. Examples of novel species are those with a name ending in capitalised letters and digits such as *Burkholderia* sp. BDU8, *Campylobacter* sp. BV-R1, *Legionella* sp. D2863 and *Mycobacterium* sp. AFPC-000167.

### 16S proportional similarity index analysis

The divergence among metagenomes of different origins based on the proportional similarity indices of taxa related to both the *Campylobacteraceae* Family and WHO-recognised pathogens associated with drinking water is illustrated by the NeighborNet trees in Figure 6.6. The water metagenomes were dissimilar to those of faecal origin. The PSI value and their 95 % confidence intervals are provided in Tables A.7 and A.8 on page 206.

### Metagenome functional analysis

The five most abundant functional factors identified in the WGS metagenomes included those related to carbohydrate metabolism (16.5 %), amino acids and derivatives (11.6 %), protein metabolism (9.9 %), DNA metabolism (8.6 %) as well as cell wall and capsule

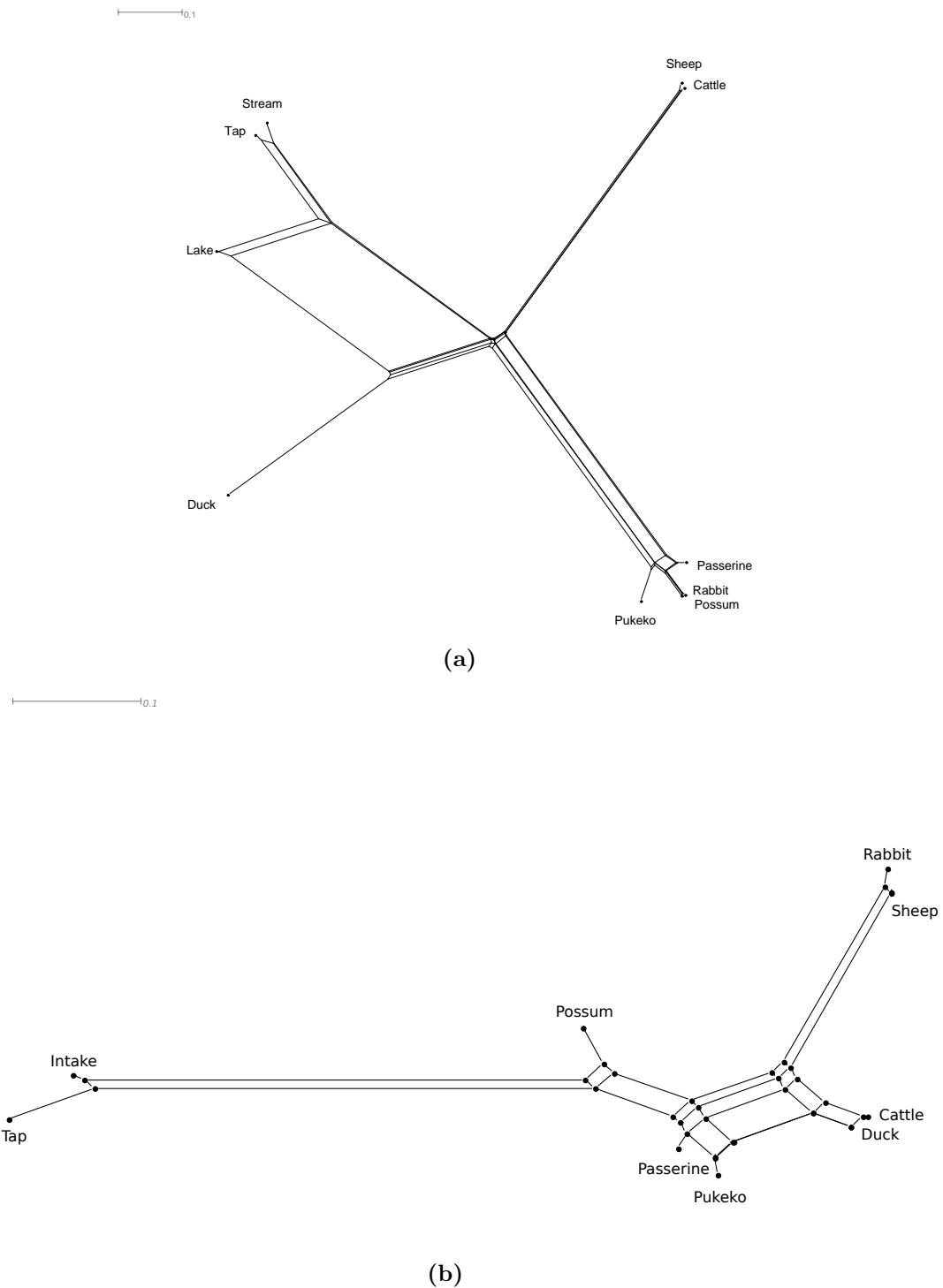


**Figure 6.5:** Phylogenetic tree for the Family *Campylobacteraceae* constructed using 16S metagenomes collected from campground operated the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand.

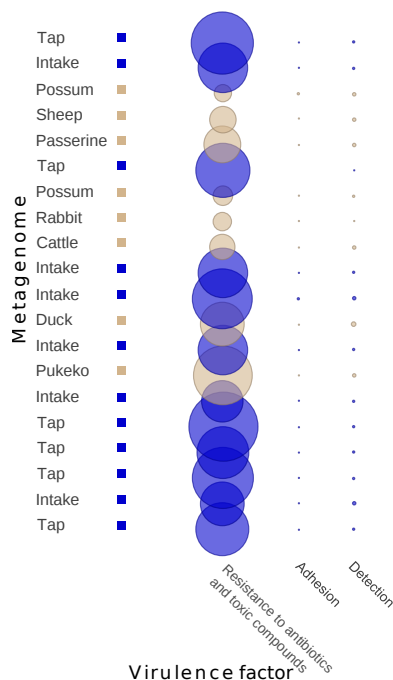
**Table 6.2:** Bacterial species deposited in the NCBI database that matched with taxa found in 16S metagenomes collected from campgrounds operated by the Department of Conservation surveyed during the 2011/2012 and 2012/2013 summer seasons, New Zealand. Also shown is the NCBI accession number, taxa number, OTU sequences, number of taxa in the OTU table and published research articles related to the organism. All hits had 100 % similarity except for denovo297986 which had 99 % similarity with *Legionella* sp. D2863.

Accession no.	Species	Taxa*	Taxa counts	Reference
KF475808.1	<i>Burkholderia diffusa</i>	denovo256673	71	Vanlaera et al., 2008
KF527218.1	<i>Burkholderia gladioli</i>	denovo256673	71	Segonds et al., 2009
KC820505.1	<i>Burkholderia</i> sp. BDU8	denovo256673	71	
KC589249.1	<i>Burkholderia</i> sp. I12B-02616	denovo256673	71	
JF958162.1	<i>Burkholderia</i> sp. JSC-R3-522-9	denovo45043	284	
KF305636.1	<i>Burkholderia</i> sp. KN2-3	denovo88144	401	
HF674704.1	<i>Burkholderia</i> sp. Kb12	denovo88144	401	
KF551161.1	<i>Burkholderia</i> sp. PTGT-5	denovo88144	401	
HF674683.1	<i>Burkholderia</i> sp. RAU2l	denovo88144	401	
HQ628642.1	<i>Campylobacter lanienae</i>	denovo141548	1235	Logan et al., 2000
KF192319.1	<i>Campylobacter lanienae</i>	denovo141548	1235	Logan et al., 2000
KF192320.1	<i>Campylobacter lanienae</i>	denovo141548	1235	Logan et al., 2000
KF192321.1	<i>Campylobacter lanienae</i>	denovo141548	1235	Logan et al., 2000
JQ863067.1	<i>Campylobacter</i> sp. BV-R1	denovo401932	15727	
JQ863069.1	<i>Campylobacter</i> sp. BV-R2	denovo401932	15727	
JQ863072.1	<i>Campylobacter</i> sp. BV-R3	denovo401932	15727	
JQ863073.1	<i>Campylobacter</i> sp. BV-R4	denovo401932	15727	
JN380984.1	<i>Legionella</i> sp. D2863	denovo297986	2319	
JN380995.1	<i>Legionella</i> sp. Edu-2	denovo196715	1265	
KF019697.1	<i>Mycobacterium austroafricanum</i>	denovo208543	2335	Croce et al., 2014
KF019694.1	<i>Mycobacterium lentiflavum</i>	denovo58140	315	Marshall et al., 2011
KF028776.1	<i>Mycobacterium simiae</i>	denovo58140	315	Van Ingen et al., 2008
KC113104.1	<i>Mycobacterium</i> sp. AFPC-000167	denovo58140	315	
KC113105.1	<i>Mycobacterium</i> sp. AFPC-000172	denovo58140	315	
JX566888.1	<i>Mycobacterium</i> sp. AW6	denovo240695	215	
KF019695.1	<i>Mycobacterium</i> sp. AW7-2	denovo240695	215	
JX469387.1	<i>Mycobacterium</i> sp. BJC15-C31	denovo38454	361	
JX469393.1	<i>Mycobacterium</i> sp. BJC15-C37	denovo38454	361	
FN386730.1	<i>Mycobacterium</i> sp. Sco-B08	denovo38454	361	

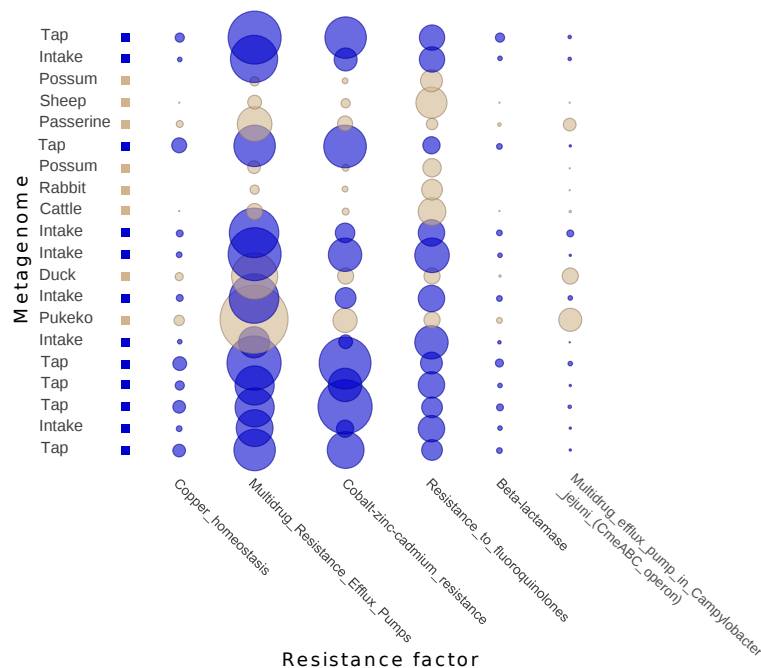
\*Replicate taxa indicate sequences matching multiple species in the NCBI database



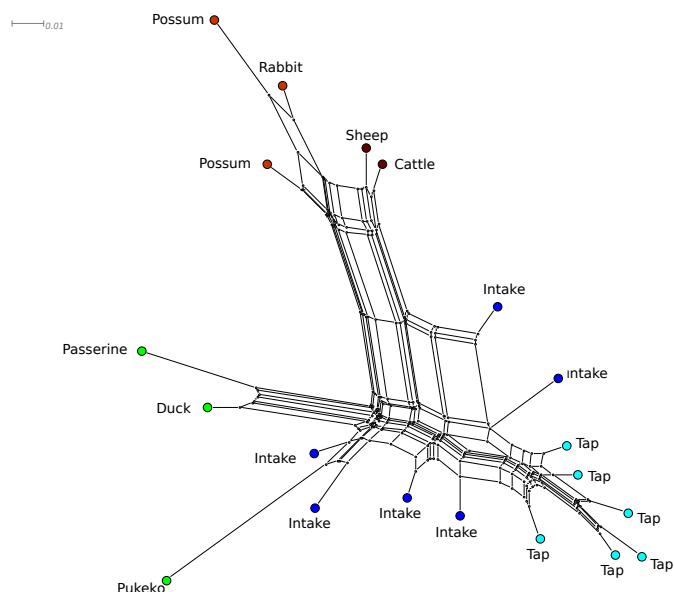
**Figure 6.6:** NeighborNet trees illustrating divergence of metagenome sources based on taxa related to **a** *Campylobacteraceae* Family and **b** WHO-recognised pathogens associated with drinking water.



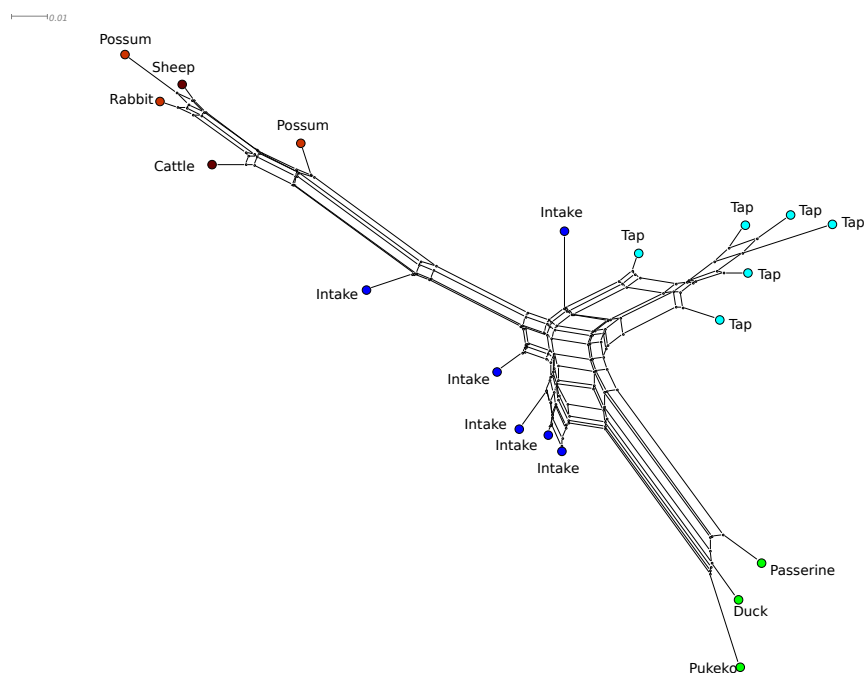
**Figure 6.7:** Bubble plot showing the abundance of virulence factors found in each of the twenty WGS metagenomes collected from campground operated by the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand. The colours brown and blue indicate faecal and water samples, respectively. Bubble size is proportional to the abundance of virulence factor in each sample.



**Figure 6.8:** Bubble plot showing the abundance of resistance factors found in each of the twenty WGS metagenomes collected from campground operated by the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand. The colours brown and blue indicate faecal and water samples, respectively. Bubble size is proportional to the abundance of resistance factor in each sample.



**Figure 6.9:** NeighborNet tree illustrating divergence of the virulence factors found in WGS metagenomes collected from campgrounds operated by the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand.



**Figure 6.10:** NeighborNet tree illustrating divergence of the resistance factors found in WGS metagenomes collected from campgrounds operated by the Department of Conservation surveyed in the summer months of 2011/2012 and 2012/2013, New Zealand.

(5.5 %). Among the virulence factors the three most abundant included those related to antibiotic and toxic compound resistance (89.9 %), detection (6.1 %) and adhesion (3.3 %) (Figure 6.7). Multidrug resistance efflux pumps (30.6 %), cobalt-zinc-cadmium resistance (19.8 %), resistance to fluoroquinolones (19.8 %), beta-lactamase (9.9 %), copper homeostasis (5.9 %) and multidrug efflux pump in *Campylobacter jejuni* (4.3 %) were among the most common antimicrobial/toxic compound resistance factors (Figure 6.8).

The relatedness of WGS metagenomes based on virulence factors is illustrated by the NeighborNet tree presented in Figure 6.9. Broadly, three clusters can be recognisable: the domestic ruminant cluster at the top, the water cluster in the bottom right corner and the wild bird cluster on the left. A similar pattern is evident in the NeighborNet tree constructed using antimicrobial and toxic compound factors (Figure 6.10).

## 6.4 Discussion

The present study investigated the microbial content and quality of the drinking water provided at campgrounds operated by the Department of Conservation (DOC) in New Zealand. Microbial source tracking was performed and identified animals likely to contaminate drinking water sources at the campgrounds. For this purpose, twenty campgrounds were recruited and surveyed over two summer seasons (December-February), i.e. periods when the highest public health risk related to drinking water at the campgrounds was expected. The microbial content and quality of the drinking water was examined using metagenomics techniques. This approach was adopted for a variety of reasons. The first reason was that metagenomics provides a much more unbiased perspective of the microbial profile of a given habitat compared to conventional cultured-based methods currently used in water quality testing. The second reason was that metagenomics techniques have the ability to directly detect multiple pathogens in a single test as opposed to the use of indicator organisms as is the case with current water quality testing methods. In this way the direct public health hazard associated with drinking water was revealed. The third reason was that metagenomics can be used for microbial source tracking e.g. through comparing microbial community profiles and/or functional factors from different habitats.

The tools used in the present study included high quality 16S ribosomal ribonucleic acid (rRNA) gene (16S for short) and whole genome shotgun (WGS) sequences as illustrated by the Q-scores. The sequences were analysed using three main techniques in order to highlight similarities/differences among metagenomes of different origins. The first technique was a multivariate approach through the application of canonical correspondence analysis (CCA). This technique is commonly used in metagenomic and ecology studies (Carpi et al., 2011; Gianoulis et al., 2009; Liang et al., 2011; Rakocinski et al., 1996; Ter Braak and Verdonschot, 1995). The second technique involved the estimation of proportional



similarity index (PSI) and construction of trees based on the estimated indices. Previous studies have used various types of indices to discriminate between metagenomes of different origins (Mitra et al., 2010; Nalbantoglu et al., 2011; Somboonna et al., 2012). Functional factor abundance scores were also used to construct trees. The public health hazards were highlighted through the direct detection and identification of known pathogenic organisms and also through the identification of virulence factors of the microbial communities.

Water samples were found to have higher taxa (species) richness index scores than faecal samples. Similar findings have previously been reported regarding viral metagenomes (Allen et al., 2013). Further, CCA showed that abundance scores can be used to discriminate metagenomes of different origins. These findings suggest that certain microbes favour particular environments. Therefore, identifying signature markers of such microbes could greatly enhance microbial source tracking. Previous studies have identified such markers, for example, Gomi et al. (2014) reported identifying host-specific markers in the *E. coli* genome specific to chicken, cow, human and pig hosts. A combination of multiplex polymerase chain reaction (PCR) and dual index sequencing were used to successfully identify these markers in multiple isolates. In another study, Kildare et al. (2007) developed an assay for quantitative detection of faecal contamination from cow, dog and human hosts based on the *Bacteroidales* 16S markers.

The *Campylobacteraceae* phylogenetic tree provided evidence that water is dominated by members of the genus *Arcobacter*. Most species under this genus are regarded as non-pathogenic and are of low public health significance. However, this confirms that the majority of the microbes living in water do not cause disease. The National Center for Biotechnology Information (NCBI) hits provided direct evidence of the microbial public health hazard attributed to drinking water provided at the campgrounds. The single-test multiple pathogen detection ability of metagenomics, illustrated by the NCBI hits, is an important feature that could be exploited for developing enhanced water quality testing assays. Currently, the use of indicator organisms, is common practice in the drinking water industry, however, the practice has over the years been shown to lack specificity and sensitivity in pathogen detection (Bonadonna et al., 2002; Harwood et al., 2005, 2014; Lund, 1996), hence metagenomics offers a credible alternative.

The metagenome functional analysis further revealed that virulence factors were among the most common factors in the metagenomes found at the campgrounds. This analysis revealed that the capability to cause disease or resist antimicrobial action was widespread among the members of the microbial communities. It is, however, not possible to establish through metagenomics whether these capabilities were active or latent. The possibility of transfer of virulence factors among members of the microbial communities implies that microbes that have previously been known to be non-pathogenic could become pathogenic, these are

sometimes referred to as emerging pathogens. While sharing of antimicrobial resistance factors among microbes implies that microbes that were previously sensitive to drugs could become resistant leading to untreatable diseases. Previous studies have provided evidence of antimicrobial resistance transfer among members of a given microbial community (Babic et al., 2006; Martínez, 2008; Nikaido, 2009; Shi et al., 2013). The metagenomic functional analysis findings of the present study demonstrate that metagenomics can provide pertinent information for the understanding of microbial community interactions and the evolution of new pathogens as well as antimicrobial resistance. In a recent report on antimicrobial resistance, the World Health Organization (WHO) highlighted the fact that globally there are high rates of antimicrobial resistance being observed in bacteria causing infections in both the hospital and community setting. The report also revealed that there are major gaps in information regarding the evolution, transmission and impact of antimicrobial resistance (World Health Organization, 2014a). This shows the importance of conducting research, such as the current study, that will help fill the information gap regarding emerging pathogens and antimicrobial resistance. The accumulated evidence could then be used for developing strategies to mitigate both the emerging disease and antimicrobial resistance problems.

The strengths of the current study include the fact that the ability of metagenomics to detect multiple pathogens in a single assay was demonstrated. This could be used to overcome one of the major shortcomings of the water quality testing methods currently in use i.e. the use of indicator organisms. One source of weakness in the current study is that environmental faecal samples were used. It is possible that the samples were contaminated with microbes from the surrounding environment e.g. soil. The possibility was greatest for small-size faeces such as those from birds. Contamination could have resulted in erroneous assignment of metagenomes to sources, e.g. soil metagenomes being assigned to faecal sources. However, for large-size faeces such as those from cattle, effort was made to minimise contamination by collecting samples from uncontaminated portions such as the centre of the faecal pad. Another source of weakness is the small number of samples used in the WGS metagenome functional analysis. This was due to operational limitations that included computer resource constraints. The small dataset may have adversely affected the study power in robustly highlighting similarities/differences among the metagenomes. Further, caution should be exercised when interpreting the public health hazard posed by the organisms identified through metagenomics because environmental deoxyribonucleic acid (DNA) is used as the input material. This means that it is not possible to determine whether the DNA was from live organisms, dead cells or was naked ancient DNA floating in the environment. One consequence of this is that the public health hazard is likely to be overstated. Further, recent research has shown that reagent contamination can influence the metagenomic profile, particularly in samples with low concentrations of DNA (Salter et al., 2014). To assess the level of reagent contamination, negative controls are recommended

during sample processing. In the present study no negative controls were run, however, the Illumina sample quality control requirements ensured that samples with very low concentrations of DNA ( $< 100 \text{ ng}/\mu\text{L}$ ) were excluded from the sequencing process. Further, the distinct clustering of the 16S metagenomes of different origins exhibited in the CCA plot suggest that contamination of common source had minimal effect in the current analysis. In addition, the public health hazard assessment in the current study was conducted using selected bacterial species that are not included on the reagent contaminant list provided by Salter et al. ([2014](#)).

In summary, the metagenome related to drinking water provided at the DOC-operated campgrounds in New Zealand was described, revealing the quality of the water at the molecular level. The microbial community functional analysis revealed the potential public health hazards attributable to aquatic metagenomes associated with drinking water. Virulence factors, particularly antimicrobial resistance factors, were found to be widespread among drinking water metagenomes. However, the present study did not establish whether these functional factors were active or latent. The microbial community functional factor evidence, though, adds to the body of knowledge that could help in the understanding of the evolution and transmission of both emerging pathogens and antimicrobial resistance. In turn the knowledge could be used to develop better strategies for combating emerging diseases and antimicrobial resistance. The present study demonstrates that metagenomics has the potential to dramatically improve water quality testing.

# Seven

## General discussion

### 7.1 Background

This thesis starts with Chapter 1 which provides an introduction to water quality and the sequencing technologies. In Chapter 2 the drinking water delivery system in New Zealand is reviewed. Also reviewed are the common methods used for determining drinking water quality and safety. In addition, the current status regarding the use of metagenomic techniques in the drinking water supply systems is provided.

The first study (Chapter 3) is an investigation of the factors that are associated with the presence/absence of microbes in raw water intended for treatment and public consumption. The investigation revealed that *E. coli*, the water industry standard microbial contamination indicator, was a good predictor for the presence/absence of *Campylobacter* (bacteria) but not for the concentrations of *Cryptosporidium* or *Giardia* (protozoa) in raw drinking water. This implies that *E. coli* is not suitable for use as a universal microbial contamination indicator. Cattle densities were generally found to be associated with raw water microbial contamination, particularly dairy cattle densities. However, the strength of the association was minimal. Season did not appear to be a very important predictor for determining the concentration of the monitored microbes in raw drinking water, meaning that targeting pathogen removal at certain times of the year was of limited value. The relationship between river flow on rivers supplying drinking water and reports of gastrointestinal illness in the community was described using a distributed lag non-linear modelling approach in Chapter 4. Analysis of this relationship revealed that an increase in the number of gastrointestinal illness reports was likely to be observed around ten days after high flow. It was further established that small drinking water distribution networks were likely to experience a higher increase in the number of gastrointestinal illness reports compared to large networks. These findings suggest that upgrading water treatment facilities in small drinking water distribution networks is an important exercise in order that delivery of microbiologically safe drinking water is enhanced.

Chapter 5 investigated the quality of drinking water provided at campgrounds operated by the Department of Conservation (DOC) using culture-based laboratory methods. The investigation established that application of water treatment using methods such as ultra

violet (UV) or chemical treatment was the most effective way of ensuring that tap water was compliant with drinking water regulatory requirements as stipulated in the drinking water standards for New Zealand (DWSNZ) 2008. This implies that strategies such as installing effective drinking water treatment facilities at campgrounds that currently have none, particularly those that are popular with campers, should be instituted in order to further reduce the risk of waterborne disease outbreaks. The metagenomes found at the campgrounds are described in Chapter 6. Evidence in this chapter supports the hypothesis that microbial communities vary from habitat to habitat. This phenomenon was exploited in order to track sources of contamination in drinking water. The potential for the metagenomic techniques to provide enhanced drinking water quality assays of high sensitivity and specificity was also demonstrated.

### 7.1.1 Types of data

This thesis utilised different types of data from various sources and these data can be grouped into seven main categories:

- **Geospatial:** The data in this category were mainly in form of digital maps with accompanying attribute tables. The digital map layers were those related to land cover, lithology, soil temperature and farming activities throughout New Zealand.
- **River flow rates:** Daily average flow rates recorded on various rivers within New Zealand were obtained for a ten-year period, 1997–2006. Flow rates were converted from  $\text{m}^3/\text{s}$  to percentile in order to make them comparable between large and small rivers.
- **Climatic:** These types of data comprised daily average rainfall and ambient temperature for a ten-year period, 1997–2006, recorded at weather stations located closest to the river flow recording sites.
- **Disease cases:** These were human cases of gastroenteritis (campylobacteriosis, cryptosporidiosis, giardiasis and salmonellosis) extracted from the New Zealand national notifiable disease surveillance (EpiSurv) database for a ten-year period, 1997–2006.
- **Drinking water supply:** These data were extracted from the 2001 register of community drinking-water supplies in New Zealand.
- **Laboratory data:** Laboratory results from the various analyses performed on water and faecal samples collected through field studies (presented in Chapters 3, 5 and 6) composed these types of data.
- **DNA sequence-related data:** These data included metagenomic sequence data generated by sequencing deoxyribonucleic acid (DNA) extracted directly (without culturing microbes in the laboratory) from environmental samples and multilocus sequence typing (MLST) data generated through sequencing of seven house-keeping genes of *Campylobacter* cultured isolates.

Utilisation of the data described above varied from study to study. Some of the data were specific to particular studies while others were commonly used across the studies presented in this thesis. Examples of data that were specific to particular studies include river flow and metagenomic sequence data that were used exclusively in the river flow study (Chapter 4) and in the campground study (Chapter 6), respectively. In contrast, geospatial data (digital maps) were used in one form or another in all studies.

A variety of computer software (programs) were used to amalgamate and synthesize the data described above for presentation in this thesis. The software can be categorised into five groups based on the tasks performed by each software (Table 7.1). For the sake of clarity, only the main tasks performed by each software are shown, however, in practice multiple tasks were performed by each piece of software.

**Table 7.1:** Computer software used for data processing, data analysis and thesis compilation.

Software	Purpose	Reference
ArcMap 10.0	Digital map display	(Environmental Systems Resource Institute, 2010)
BioNumerics 7.1	MLST data processing	(Applied Maths NV, 2014)
Basic local alignment search tool (BLAST) 2.2.29	Sequence alignment	(Altschul et al., 1990)
Excel 2010	General data management and storage	(Microsoft, 2012)
Fast length adjustment of short reads (FLASH) 1.2.6	Metagenomic data processing	(Magoč and Salzberg, 2011)
Jabref 2.7	Bibliography management	(JabRef Development Team, 2014)
Knitr 1.7	Report writing	(Xie, 2014)
L <sup>A</sup> T <sub>E</sub> X	Report writing	(Mittelbach et al., 2004)
Metagenome analyzer (MEGAN) 5.7	Metagenomic data processing and analysis	(Huson et al., 2011)
MySQL 5.5	General data management and storage	(DuBois, 2008)
Protein alignment using a DNA aligner (PAUDA) 1.0.1	Metagenomic data processing	(Huson and Xie, 2013)
Perl 5.14.2-21	Pipeline execution	(Christiansen et al., 2012)
Python 2.7.3	Pipeline execution	(Python Software Foundation, 2013)
Quantitative insights into microbial ecology (QIIME) 1.7.0	Sequence data processing	(Caporaso et al., 2010)
R 3.1.0	Statistical analysis	(R Core Team, 2013)
SolexaQA 2.2	Metagenomic data quality trimming	(Cox et al., 2010)
Splitstree 4.13.1	Data visualisation	(Huson and Bryant, 2006)

## 7.2 Challenges and pitfalls

During the course of the research work that culminated into this thesis there were a number of challenges and unforeseen occurrences. The biggest challenges were by far those related to the campground study, however, the other studies also presented challenges, e.g. in data management and analysis. The challenges can be divided into three groups i.e. sample collection, sample processing as well as data management and analysis.

### 7.2.1 Sample collection

The challenges outlined here relate to those encountered during the field work in the campground study (Chapters 5 and 6).

#### **Faecal samples**

During the planning stages of the campground study, the epidemiological aspects of the study design were considered in detail in order to develop a suitable faecal sampling scheme. For example, a spatially random sampling scheme for faecal sampling was envisaged. In the proposed scheme, each campground drinking water catchment was to be divided into 20–30 regular grids and a random point in each grid selected from which to collect a faecal sample. The points were to be located using a hand held global positioning system (GPS) receiver during sample collection. The sampling scheme was designed to ensure that representative samples are collected from all parts of a given catchment. In this way the *true* faecal burden upstream of the abstraction point would be estimated.

However, it was realised that the scheme was impractical as soon as digital map layers of the catchments were obtained and catchment areal sizes were calculated. It was realised that some catchments were too large to be covered on foot. An attempt was then made to reduce the coverage area by limiting the sampling areas to 50 m-buffer zones along the riverbeds. Again, some catchments were too large to be covered on foot. A visit to one of the campgrounds confirmed that such an idealistic scheme could not be implemented. It came to light that even for small catchments the terrain and vegetation cover was often prohibitive for one to walk along the entire riverbed. Also, faeces were not uniformly available throughout the catchment, thus, predefining a location (or a zone) at which to collect samples was not practical as there was no guarantee that faeces would be present at the selected location.

Given these challenges, the faecal sampling scheme was modified in two ways. The first modification was that the spatially random sampling scheme was no longer a requirement, instead, samples were collected wherever they were sighted. The second modification was related to campgrounds at which access to areas upstream of the water abstraction point was not possible due to prohibitive terrain and/or vegetation cover e.g. Otaki Forks, Whites

Bay and Anaura Bay campgrounds. On such campgrounds faecal samples were collected in areas downstream of the water abstraction point. Because of these modifications to the sampling scheme, the spatial spread of locations at which faecal samples were collected in relation to the water abstraction point was not uniform across the campgrounds.

Given that New Zealand wildlife is dominated by birds, many of them with very small droppings, sighting of faeces was quite hard. Further, laboratory processes such as *Campylobacter* culturing as well as *Cryptosporidium* and *Giardia* microscopy required fresh faecal samples. These factors made finding appropriate faecal samples difficult and time-consuming especially on campgrounds with little animal life. Apart from this, on rainy days it was hard to differentiate between fresh faeces and old but rehydrated ones.

### Water sample collection

Prior to the onset of sample collection, all study campground managers were contacted and requested to provide GPS coordinates for the water abstraction points. Once provided with the coordinates it seemed to be a straightforward case of arriving at the campground and locating the abstraction point using the provided coordinates and a hand held GPS receiver. In practice it turned out not to be that straightforward and on a number of campgrounds, e.g. Totaranui and Waikawau Bay, the abstraction points were located in nearby hills, with obscure and treacherous access. The reasons for locating the abstraction points uphill with obscure access were twofold. One reason was the desire to have a gravity-driven water supply system that provided adequate pressure without the use of pumps. This was because these remote campgrounds had limited or no electricity supply. The other reason was aimed at minimising contamination of the water sources, thus, access was deliberately left obscure and barely maintained so that unauthorised persons such as trampers and campers did not have access to the water source. For this reason, the assistance of the campground management was required in locating the water sources. During the first sampling round prior arrangements with campground management had to be made for someone in charge of the water supply system to be available on the day of sampling to give directions to the abstraction site.

The actual collection of water samples at the abstraction sites was challenging at some campgrounds, particularly at Totaranui. The sampling equipment comprised a 12 V battery, water pump, filter cartridge, flow meter and sample bottles, weighing over 15 kg in total. This equipment had to be carried to the abstraction sites, often uphill trips that took up to an hour in some cases. On return, the load was even heavier because of the water samples. Faecal sample collection was carried out on return from source water collection, which involved squatting repeatedly with the heavy load on the back as faecal samples were collected.



### 7.2.2 Sample processing

As with sample collection, the challenges outlined here are those related to the campground study.

#### *Campylobacter* culturing

During the course of the study, it emerged that some *Campylobacter* isolates were slow to grow, particularly those from water samples. Such isolates did not always show obvious growth after the standard 48 h incubation period on modified charcoal cefoperazone deoxycholate agar (mCCDA). Therefore, isolates showing signs of slow or poor growth were incubated for a further 24 h on mCCDA, which sometimes resulted in some good growth. These isolates frequently had only countable colonies on blood agar and attempts to obtain sufficient colonies for DNA extraction and isolate preservation resulted in repeated reculturing. Thus a lot more time was spent per sample than normal. Because sample collection was scheduled weekly, the slow-growing isolates would still be requiring attention by the time the next batch of samples was arriving.

#### Metagenomic DNA extraction

Once the samples had been obtained, metagenomic DNA extraction commenced but did not proceed smoothly, especially in the initial stages. During the first two rounds of sample collection, 0.5 L of water sample was filtered per filter disk and stored at  $-80^{\circ}\text{C}$ . No DNA extraction was performed during the sampling period because of time limitations. Thus, extraction only commenced after the first two rounds had been completed. The Metagenomic DNA Isolation Kit for Water (Epicentre<sup>®</sup>; Wisconsin, USA) was the only kit used for the extractions in the initial stages. Repeated extraction attempts failed to yield DNA in sufficient quantities and of quality required for metagenomic sequencing on the MiSeq system. This is despite the fact that the volume of the input water sample had been increased fivefold to 500 mL per filter from the kit manufacturer's recommendation of 100 mL. This led to speculation that the concentration of organisms in the 500 mL sample was not sufficient to yield the required DNA, thus, it was decided that a larger volume of water sample be used. In order to select an appropriate volume of water to be used, samples from Otaki Forks campground were collected and DNA extractions were performed after filtration of 0.5 L, 2.0 L and 5.0 L water sample per filter. As expected, there was an increase in the quantity of DNA yielded with increasing volume of water sample. In the end 2 L was adopted as the new input sample volume.

The increased input sample volume presented a new challenge in that filtration took a longer period of time, sometimes up to 8 hours or until it was decided no more filtration through a given filter was possible. To alleviate this problem, suction pressure was applied to hasten the filtration process. However, even with suction pressure some samples required as long as 6 hours to filter. There was no clear correlation between water turbidity and filter

clogging. It appeared that very fine particulates caused more clogging than large debris; thus, sometimes clear looking water clogged more easily than turbid water.

### Metagenomic DNA sequencing

After increasing the input sample, quantity-related issues of the extracted DNA were largely resolved but quality-related problems persisted. On a number of occasions New Zealand Genomics Limited (NZGL) reported that whole genome shotgun (WGS) sequencing proceeded well but 16S ribosomal ribonucleic acid (rRNA) gene (16S for short) sequencing failed on the same DNA sample. It was reported that 16S libraries could not be prepared because the polymerase chain reaction (PCR)-amplification process often failed. This led to speculation that the water samples contained PCR inhibitors that were not removed by the Epicentre<sup>®</sup> kit. Several attempts were made to rid the DNA samples of the inhibitors using DNA cleanup kits but without success. It was then decided that a different DNA extraction kit be tried. The NucleoSpin<sup>®</sup> Soil kit was procured and worked well for most of both faecal and water samples, i.e. the number of failed 16S library preparations were drastically reduced.

Quantification of the extracted DNA was initially done using a Nanodrop<sup>®</sup> spectrophotometer, however, repeatability of measurements was very low i.e. there was wide variation in quality readings for the same DNA sample if tested multiple times by the same operator. In addition, there seemed to be a disparity between the DNA quality readings obtained using Nanodrop<sup>®</sup> spectrophotometer and those reported by NZGL, who used a Qubit<sup>®</sup> fluorometer. To mitigate this, a new Qubit<sup>®</sup> fluorometer was procured for molecular epidemiology and public health laboratory (*mEpiLab*).

### 7.2.3 Data management and analysis

The data used in this thesis can be divided into two main categories: ordinary data and metagenomic data. The ordinary data were mainly in form of Excel spreadsheets while metagenomic data were composed of sequences, nucleotide quality scores and summary data. The metagenomic data were often large files e.g. in one sequencing run ten samples yielded 43 GB worth of files. This meant that computer storage had to be managed prudently in order to accommodate both working files and backups. External computer hard drives were procured for storing backup data.

#### Data management

Most of the ordinary data were obtained from multiple sources. For instance, the river flow data were obtained from National Institute of Water and Atmospheric Research (NIWA) and sixteen regional councils within New Zealand. This meant that the data were not always in the same format and care had to be taken to ensure that the data were converted to the same format before amalgamation. Among the common differing data formats were

those related to date and GPS coordinates. Many weeks were spent on data clean up and conversion. Once the data had been cleaned, amalgamation into a relational database was first attempted using Microsoft Access 2010, however, this software was found to be deficient in handling the large volume of data and the intricate table connections. For this reason, a relational database was created using MySQL in which 34 tables were linked using 40 table connections.

### Data analysis

Once again, the major challenges encountered in data analysis were those related to the campground project, particularly metagenomic sequence data analysis. In order to analyse metagenomic data specialised programs for data processing and analysis had to be installed on the computer. However, at the time of commencing this research work metagenomic data analysis programs were in their infancy. As such, installations and operation of the programs were not smooth. Often 2–3 days were required to download and complete the installations. For instance, QIIME (Caporaso et al., 2010), the main software used to process 16S data, required a large number of dependencies for the installation to be completed. Unfortunately, most of the times a few dependencies would not install properly making QIIME unable to work. Resolving such issues was time-consuming and did not always end in success, leading to a situation where more time was spent resolving software installation issues than performing analyses. In addition, WGS data processing required enormous amount of computational power, beyond the capacity of a standard desktop computer. Therefore, WGS data had to be processed using a large central computer server available at Massey University. There were not many software options available for analysing metagenomic data especially WGS data. The few that were available required large amounts of computational power. For example, metaAmos required a minimum of 32 GB random-access memory (RAM). This was at a time when a standard desktop computer operated with 4 GB RAM. For this reason, a desktop computer with 32 GB RAM and 2 TB of storage was purchased. Even with such computer resources analysing WGS metagenomes required many days and the output files were large. For instance, processing twenty WGS metagenomes yielded more than 1 TB worth of files, consuming most of the available computer space. This became a major limiting factor in the number of WGS metagenomes that could be analysed for presentation in this thesis.

In conclusion, coming across these challenges and finding solutions was both mentally exhausting and rewarding. Although planning is of great value before a project is implemented sometimes nothing prepares one for the reality on the ground.

### 7.2.4 Future research work

The studies presented in this thesis have exposed areas that could be explored further in future. This could include investigating the value of using predicted raw water micro-

bial concentrations in enhancing the capabilities of water treatment plants in delivering microbiologically safe drinking water. Another area of future research could include an investigation in how to incorporate river flow rates in the water plant's log credit removal calculation. Incorporating machine learning techniques in such a process means that the system could self-learn from previous river flow values and pathogen levels could be developed. In this way, a self-improving water treatment regime could be initiated.

Metagenomics has been shown to have the capacity to revolutionise the way water quality testing is conducted. However, in its current form metagenomics is far from being an ideal water quality test. An ideal water quality test has properties such as easy to use, cheap, quick to produce results, sensitive and specific. While metagenomics possesses the last two properties, it still requires considerable technical expertise in sample processing, data processing and data analysis. In addition, the current techniques are expensive and require lengthy periods of time from sample collection to production of results. Thus the areas of research would be those aimed at simplifying the metagenomic testing protocol that could produce results within a short period of time with minimum costs. Alternatively, research work could be directed towards using metagenomics as a tool for identifying biomarkers that can be incorporated in tests such as multiplexed PCR that are easier to use and produce results more quickly at a cheaper cost. Recent studies in this area have shown this to be feasible (Gomi et al., 2014; Kildare et al., 2007). These properties could be exploited in order to develop a tool that can rapidly identify faecal or microbial pollution, its source and for assessing the public health importance of the pollution. The metagenome functional analysis presented in this thesis is based on less than a third of the extracted metagenomes. The plan is to acquire more computer resources and perform the analyses with all metagenomes included.



# Appendix

```
library(ggplot2); library(RISmed)

# 1.1 Search and retrieve records for metagenomic-related publications
query.mtg <- paste("(high-throughput OR (next AND generation) OR shotgun)",
"AND sequenc* OR metagenom* OR pyrosequenc*")

esearch.mtg <- EUtilsSummary(query.mtg, type = "esearch", db = "pubmed",
                             mindate = 1950, maxdate = 2013, retmax = 75000)

records.mtg <- EUtilsGet(esearch.mtg, type="efetch", db="pubmed")

# 1.2 Search and retrieve records for 16S-related publications
query.16s <- "(16S AND rrna) OR (16S AND rdna) OR (hypervariable AND region*)"
"

esearch.16s <- EUtilsSummary(query.16s, type = "esearch", db = "pubmed",
                             mindate = 1950, maxdate = 2013, retmax = 75000)

records.16s <- EUtilsGet(esearch.16s, type = "efetch", db = "pubmed")

# 1.3 Combine metagenomics and 16S years into one dataframe
topic.yr <- rbind(data.frame(Years = Year(records.16s), Topic = '16Srrna'),
                  data.frame(Years = Year(records.mtg), Topic = 'Metag'))

# 1.4 Retrive total number of publications for each year
yr.pub.total <- data.frame()
for (i in min(topic.yr$Years):max(topic.yr$Years)){
  peryear <- EUtilsSummary("", type = "esearch", db = "pubmed",
                           mindate = i, maxdate = i)
  yr.pub.total <- rbind(yr.pub.total, data.frame(Years = i,
                                                Total.pubs = QueryCount(peryear)))
}

# 1.5 Compute annual publication counts and proportions for each topic
tmp.yr <- merge(yr.pub.total, dcast(topic.yr, Years ~ Topic, length),
               by="Years")
years.df <- melt(tmp.yr, id = c("Years", "Total.pubs"),
                variable.name = "Topic", value.name = "Topic.pubs")
years.df$Prop.pubs <- years.df$Topic.pubs*100000/years.df$Total.pubs

# 1.6 Plot the results
ggplot(subset(years.df, Years< 2014), aes(Years, Prop.pubs, color = Topic)) +
  geom_line(size = 1.5) +
  theme(legend.position = c(0.10, 0.92))
```

```
# ----- Publications by Journal -----
# 2.1 Search and retrieve 16S and metagenomic-related publication records
query.mtg16s <- paste("(high-throughput OR (next AND generation) OR shotgun)",
                      "AND sequenc* OR metagenom* OR pyrosequenc*",
                      "OR (16S AND rrna) OR (16S AND rdna)",
                      "OR (hypervariable AND region*)")

esearch.mtg16s <- EUtilsSummary(query.mtg16s, type = "esearch",
                                db = "pubmed", mindate = 1950,
                                maxdate = 2013, retmax = 75000)

records.mtg16s <- EUtilsGet(esearch.mtg16s, type="efetch", db="pubmed")

# 2.2 Count the number of publications on the query topic per journal
journal.count <- as.data.frame(table(MedlineTA(records.mtg16s)))
names(journal.count) <- c("Journal", "Jo.total")

# 2.3 Select the top 20 Journals
journal.top20 <- journal.count[rev(order(journal.count$Jo.total)),][1:20,]

# 2.3 Retrive total number of publications for each of the top 20 journal
jo20.total <- data.frame()
for(i in unique(journal.top20$Journal)){
  perjournal <- EUtilsSummary(paste0(i, '[jo]'), type = 'esearch',
                              db = 'pubmed', mindate = 1950, maxdate = 2013)
  jo20.total <- rbind(jo20.total, data.frame(Journal = i,
                                              Total.pubs = QueryCount(perjournal)))
}

# Select top 20 journals and compute percentages
journal.df <- merge(journal.top20, jo20.total, by = "Journal")
journal.df$Pub.percent <- journal.df$Jo.total*100 / journal.df$Total.pubs

# 2.5 Plot the results
ggplot(journal.df, aes(Journal, Jo.total, fill = Journal)) +
  geom_bar(stat = "identity") + coord_flip() +
  theme(legend.position = "none") +
  labs(x = "Journal",
       y = "Raw number of 16S & metagenomic-related publications")

ggplot(journal.df, aes(Journal, Pub.percent, fill = Journal)) +
  geom_bar(stat = "identity") + coord_flip() +
  theme(legend.position = "none") +
  labs(x = "Journal",
       y = "Percentage of 16S & metagenomic-related publications")
```

```

# - - - - - Publications by Country - - - - -
# 3.1 Tidy up country names and count the number of publications per country
## Perl installation is required
country.count <- as.data.frame(table(gsub('Germany.*', 'Germany',
                                           gsub('Russia.*|Ussr ', 'Russia ',
                                           gsub('Korea.*', 'S.Korea ',
                                           gsub('China.*', 'China ',
                                           gsub('England|United Kingdom', 'UK',
                                           gsub('United State.*', 'USA',
                                           gsub("(^|[:space:]))([[:alpha:]])",
                                           "\\1\\U\\2", tolower(Country(records.mtg16s)),
                                           perl=TRUE))))))
names(country.count) <- c('Country', 'Country.total')

# 3.2 Select the top 15 countries
country.top15 <- country.count[rev(order(country.count$Country.total))
,][1:15,]

# 3.3 Plot the results
ggplot(country.top15, aes(reorder(Country, Country.total),
                           Country.total, fill = reorder(Country, Country.total))) +
  geom_bar(stat = "identity") + theme(legend.position = "none") +
  labs(x = "Country", y = "Number of publications") +
  theme(axis.text.x = element_text(angle = 30, vjust = 1))

```





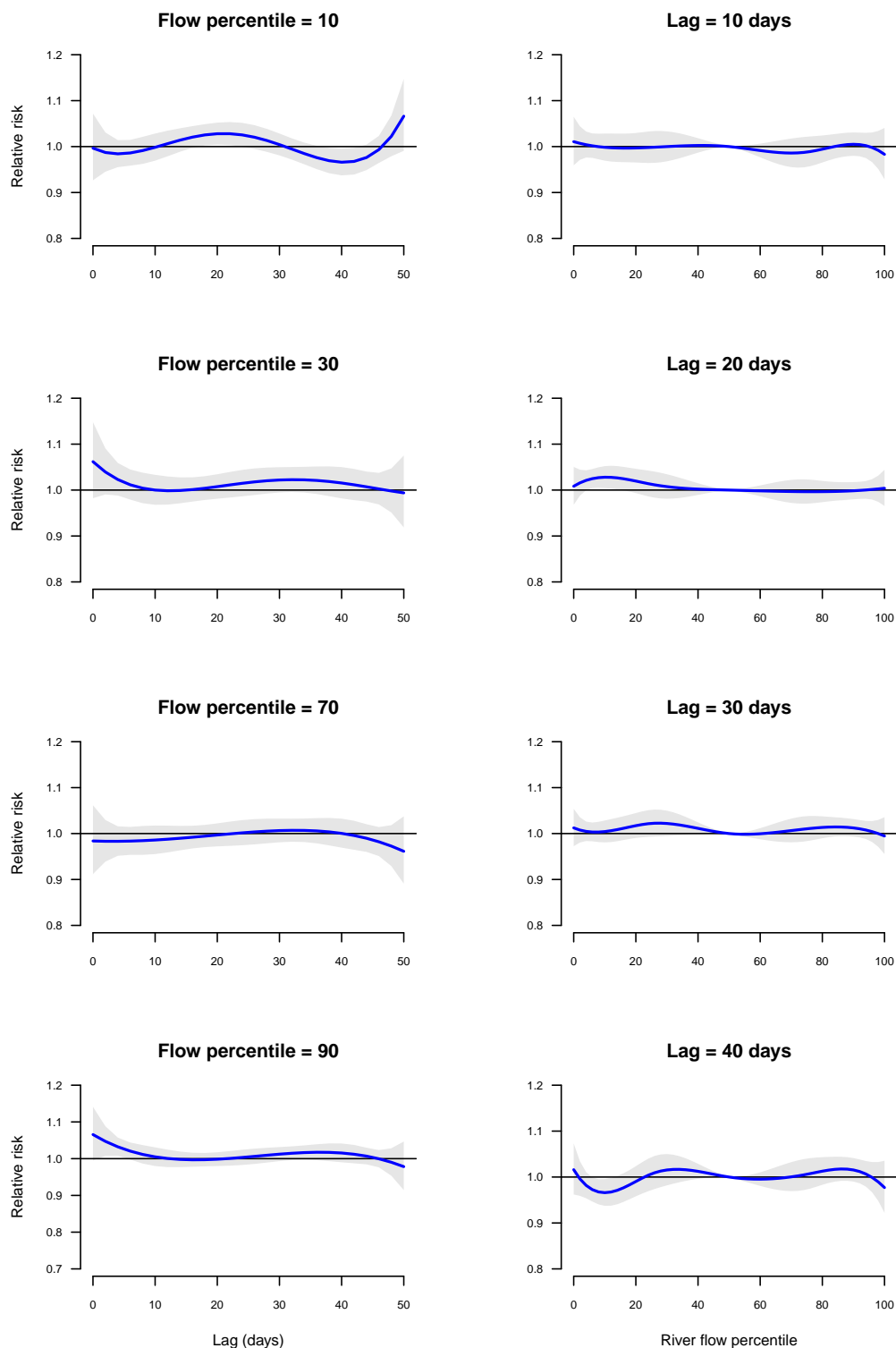
(a) Campylobacteriosis

(b) Cryptosporidiosis

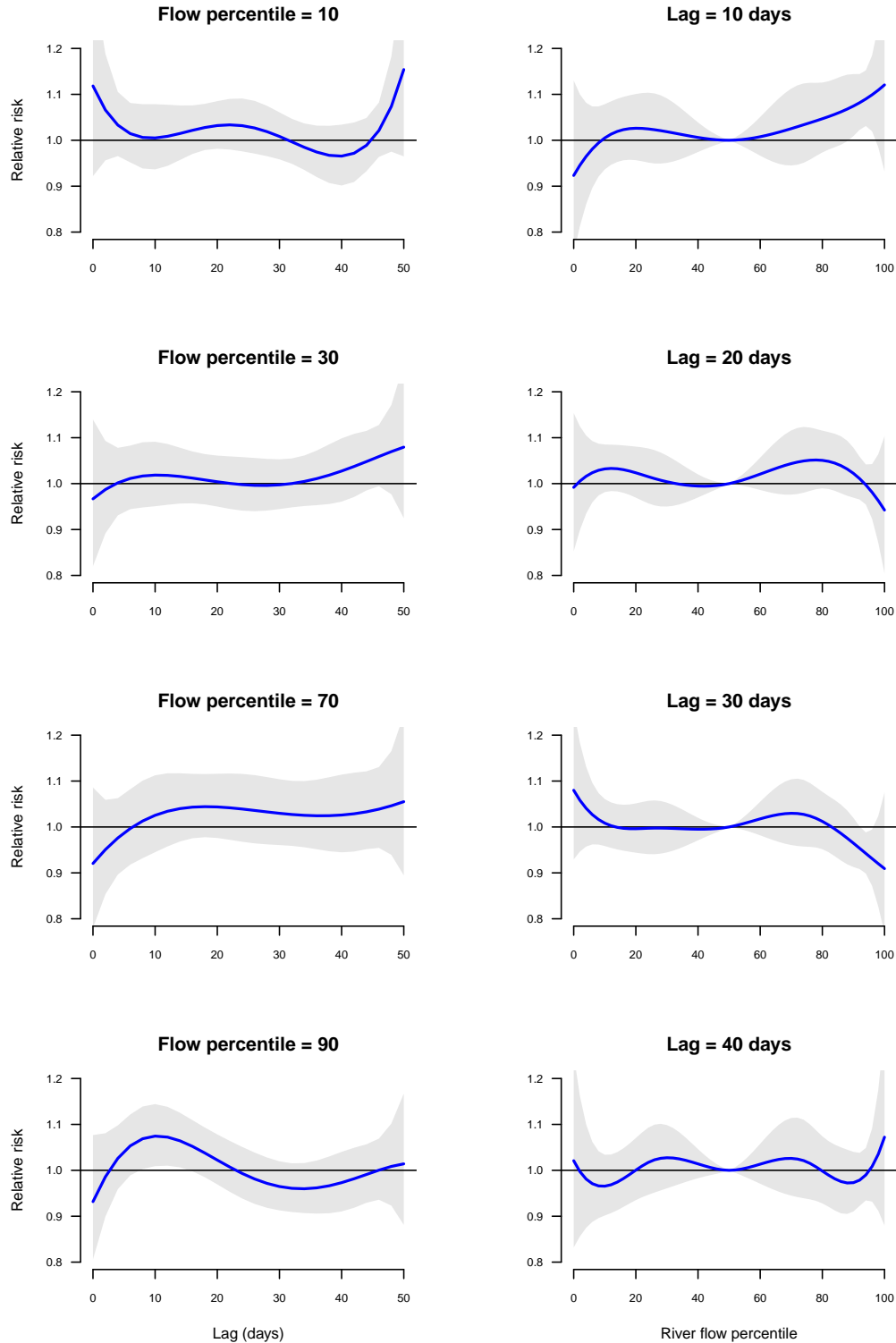
(c) Giardiasis

(d) Salmonellosis

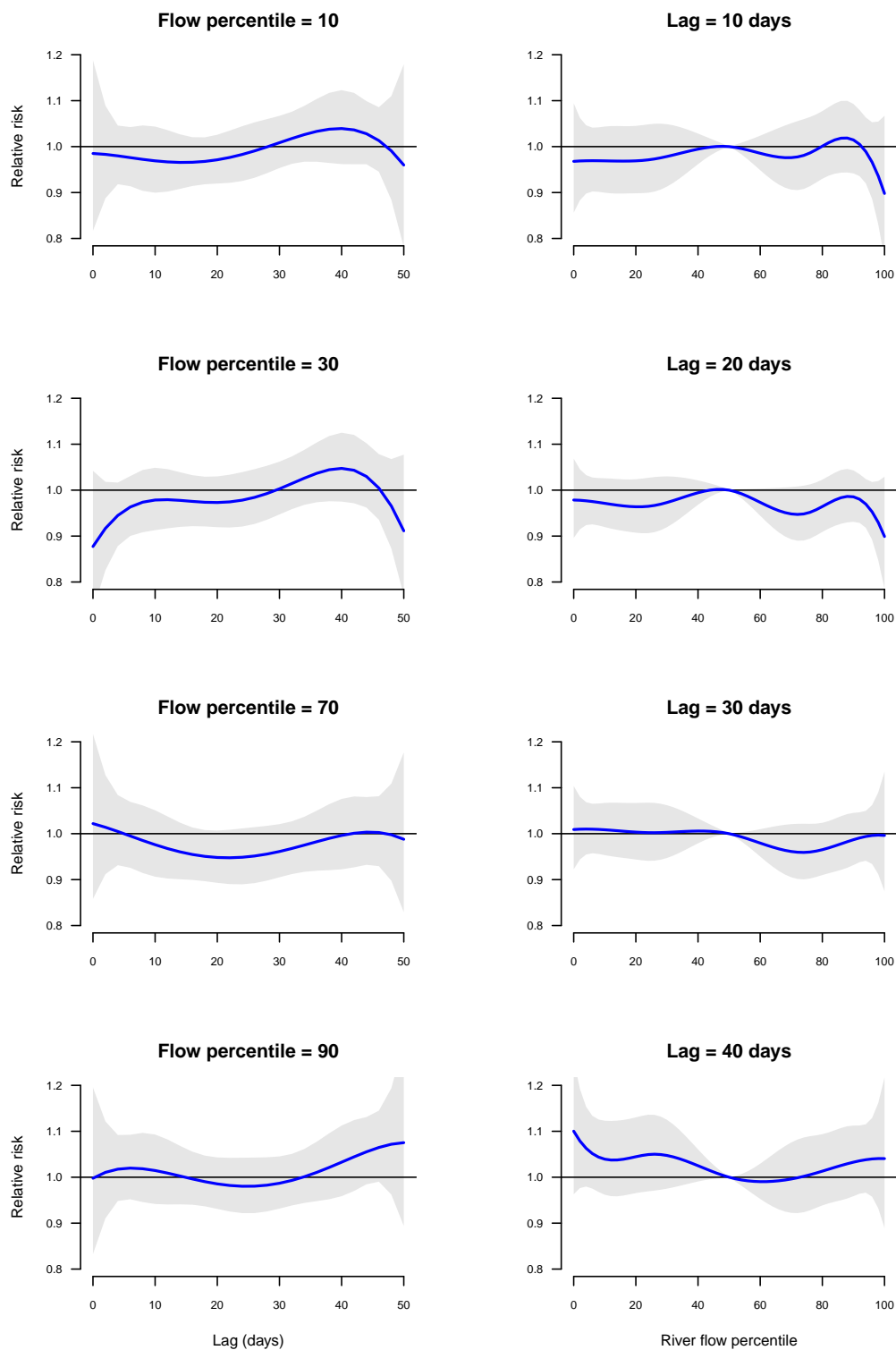
**Figure A.2:** Bubble plots of gastrointestinal illness cases for the four study diseases during the ten-year period 1997–2006, New Zealand. The size of the circles correspond to the incidence rates i.e. the larger the circle the higher the incidence rate. The electronic version of this document also shows the incidence rates for each year.



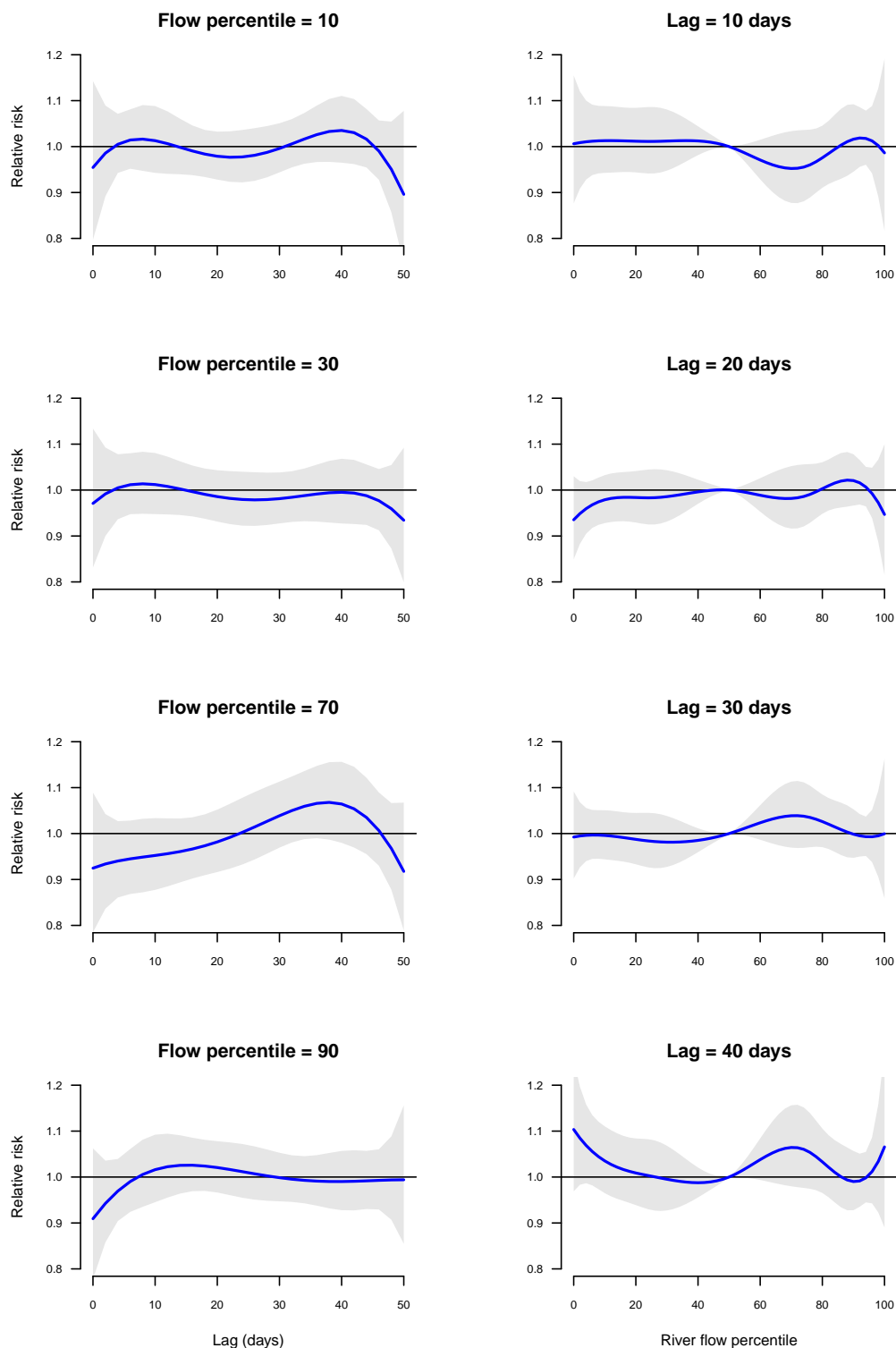
**Figure A.3: Waikato River (S00041)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95% confidence interval. The distributed lag non-linear model (DLNM) was fitted to data from the drinking water abstraction site located on the Waikato River, for the period 1997–2006, New Zealand.



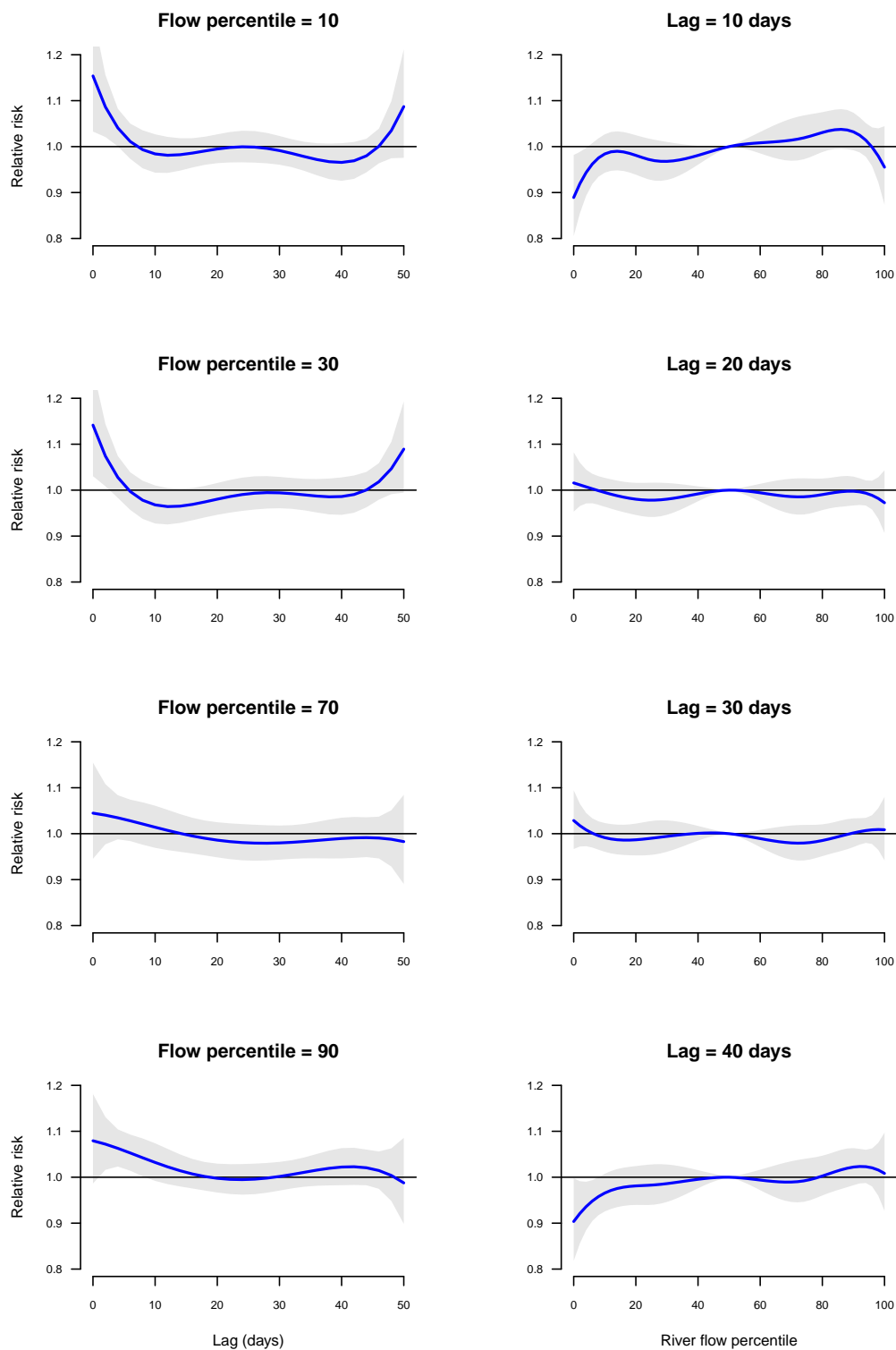
**Figure A.4: Turitea Dam (S00082)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Turitea River, for the period 1997–2006, New Zealand.



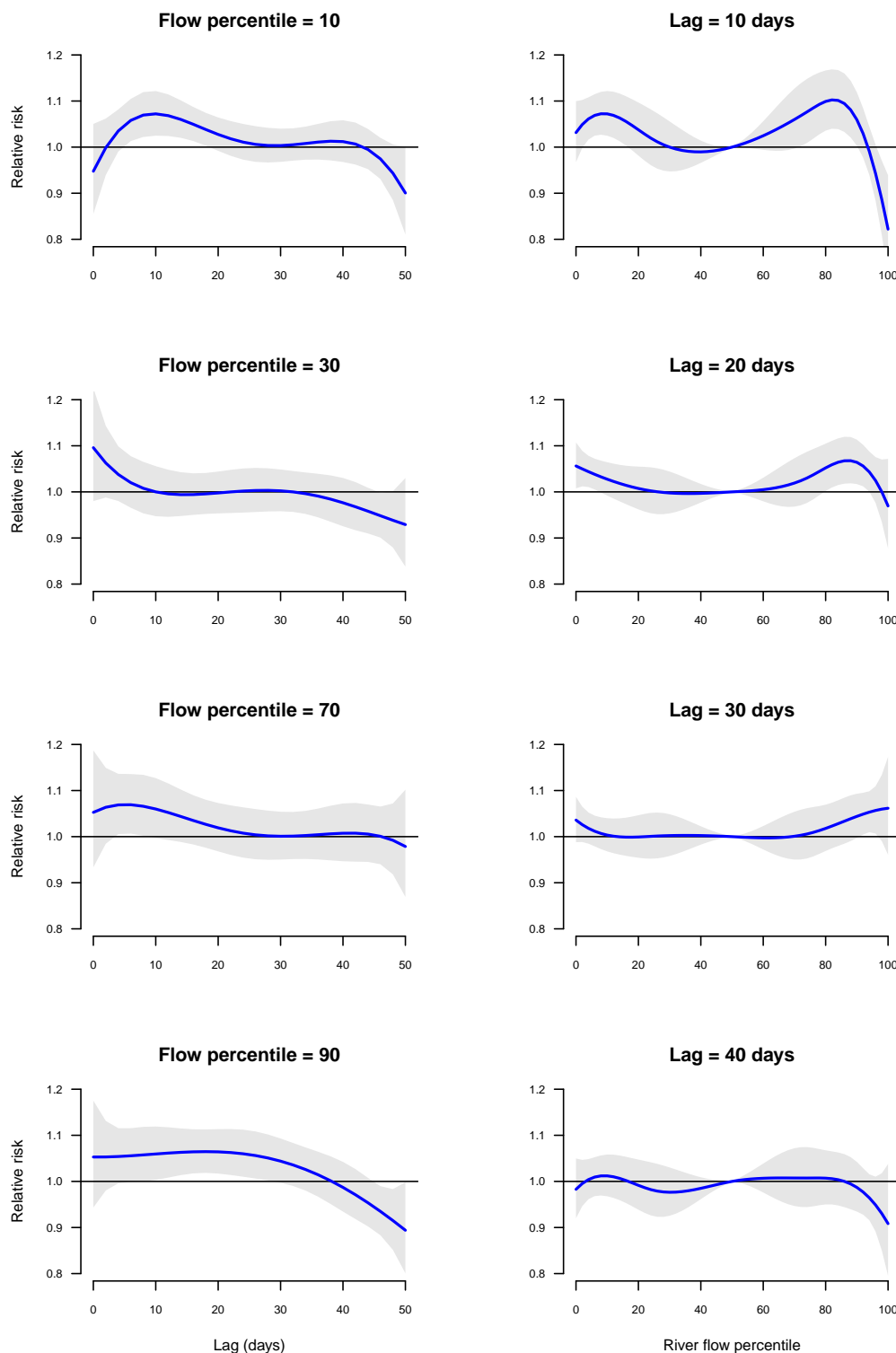
**Figure A.5: Te Arai River (S00106)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Te Arai River, for the period 1997–2006, New Zealand.



**Figure A.6: Waikanae River (S00123)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95% confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Waikanae River, for the period 1997–2006, New Zealand.

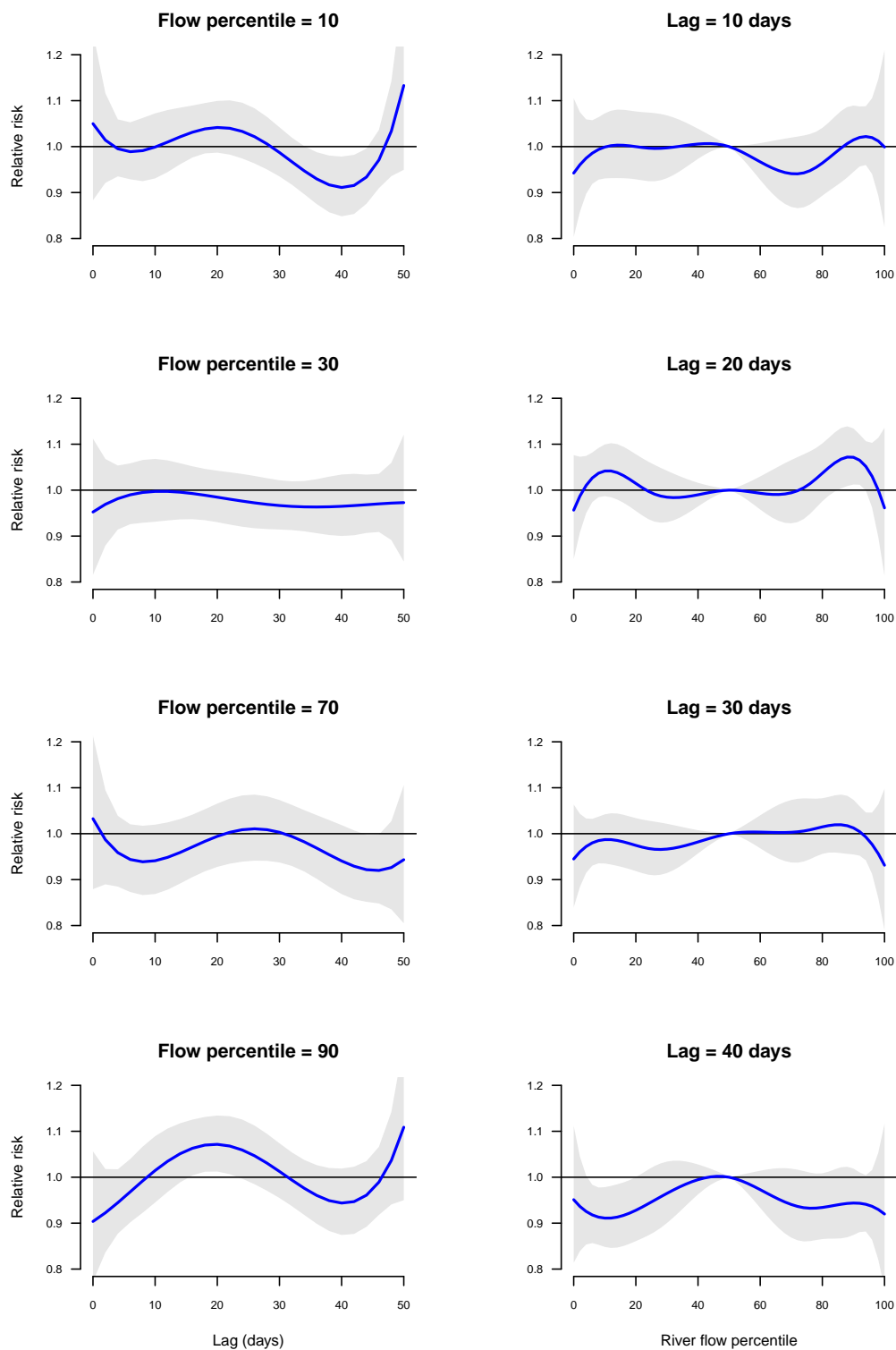


**Figure A.7: Pareora River (S00200)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Pareora River, for the period 1997–2006, New Zealand.

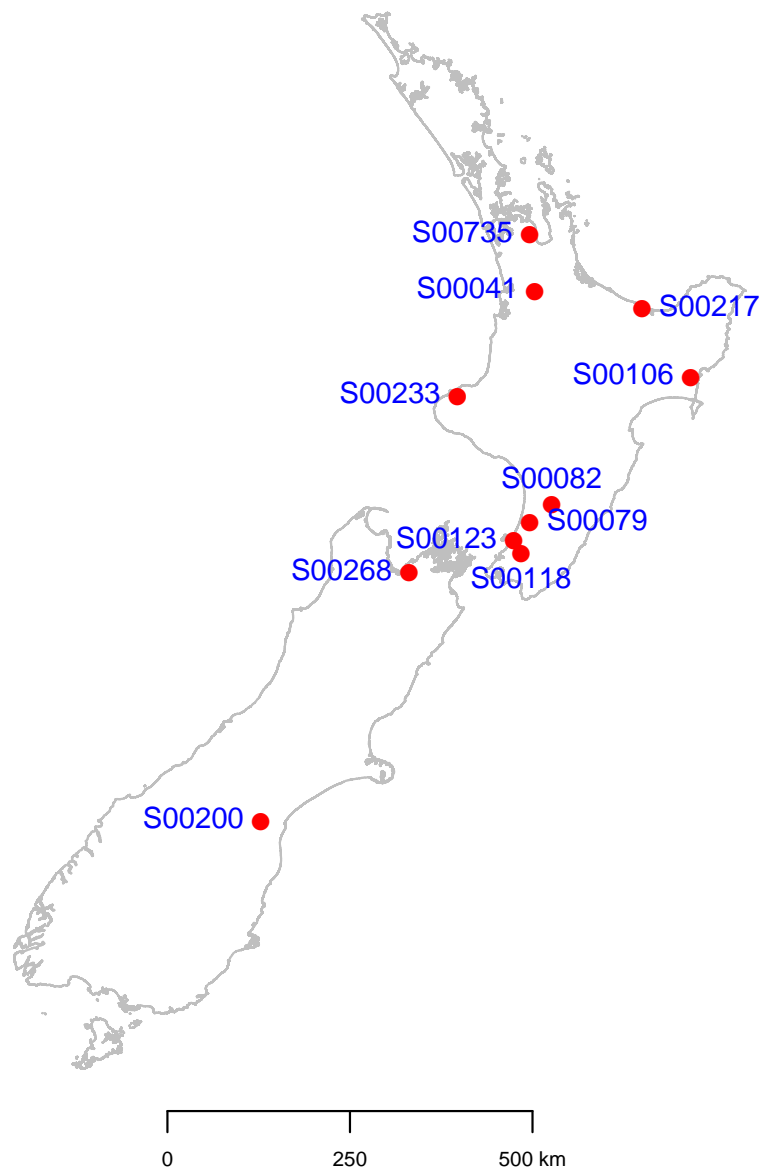


**Figure A.8: Waiwhakaiho River (S00233)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Waiwhakaiho River, for the period 1997–2006, New Zealand.





**Figure A.9: Maitai South Branch River (S00268)** — The relative risk of gastrointestinal illness reports attributable to the distributed lag river flow after adjusting for month, with the 50th river flow percentile being the reference. The relative risk is shown at different lags and different river flow percentiles with the shaded area representing the 95 % confidence interval. The DLNM was fitted to data from the drinking water abstraction site located on the Maitai South Branch River, for the period 1997–2006, New Zealand.



**Figure A.10:** Location of drinking water abstraction sites used in the distributed lag analysis, for the period 1997–2006, New Zealand.

**Table A.1:** A description of shapefiles used for geospatial data and their sources.

Shapefile	Description	Source	Website
nz-rivers-and-streams-centrelines	Rivers and streams	LINZ	<a href="http://koordinates.com/">http://koordinates.com/</a>
lcdb-v30-land-cover-datab	Land cover	LCR	<a href="http://lris.scinfo.org.nz/">http://lris.scinfo.org.nz/</a>
fsl-salinity	Soil salinity	LCR	<a href="http://lris.scinfo.org.nz/">http://lris.scinfo.org.nz/</a>
fsl-soil-temperature-regi	Soil temperature	LCR	<a href="http://lris.scinfo.org.nz/">http://lris.scinfo.org.nz/</a>
south-island-soilscapes	Soilscape for South Island	LCR	<a href="http://lris.scinfo.org.nz/">http://lris.scinfo.org.nz/</a>
north-island-soilscapes	Soilscape for North Island	LCR	<a href="http://lris.scinfo.org.nz/">http://lris.scinfo.org.nz/</a>
marlborough_wsp	Watershed for Marlborough Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
waikato_wsp	Watershed for Waikato Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
westcoast_wsp	Watershed for Westcoast Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
gisborne_wsp	Watershed for Gisborne Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
hawkesbay_wsp	Watershed for Hawkes Bay Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
wellington_wsp	Watershed for Wellington Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
tasman_wsp	Watershed for Tasman Region	MfE	<a href="http://koordinates.com/">http://koordinates.com/</a>
AgriBase_Apr11	Farming activities	AssureQuality	

**Table A.2:** Geospatial data for the sixteen surface water sources monitored for *Campylobacter*, *E. coli*, *Cryptosporidium* and *Giardia* between September 2009 and March 2014, New Zealand.

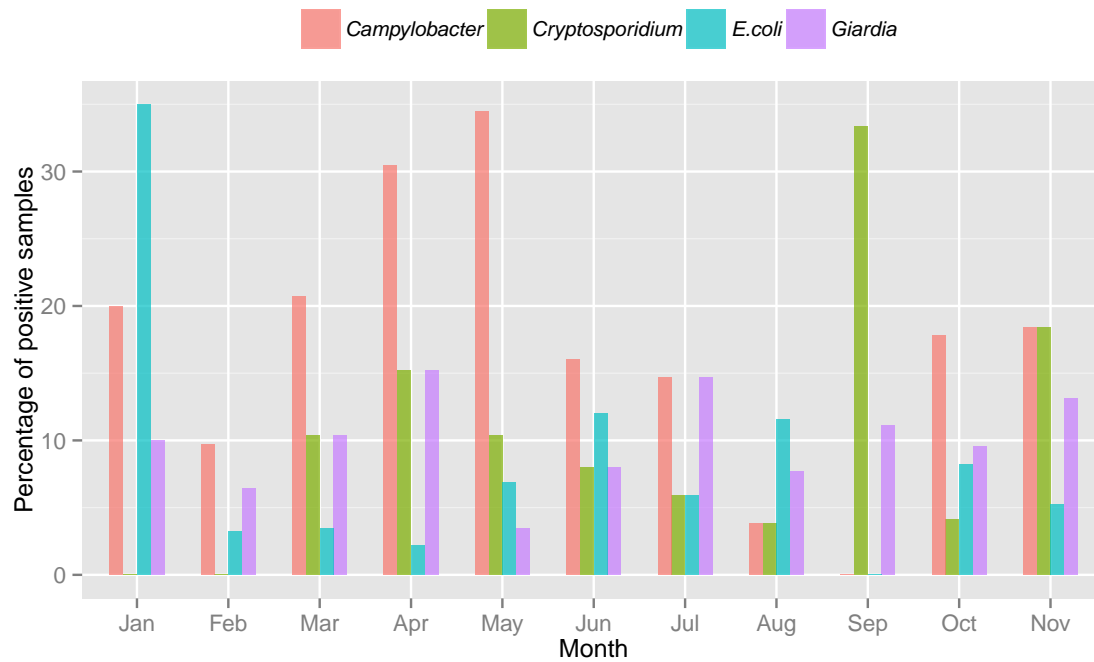
Variable	S00092	S00099	S00865	S00041	S00009	S00298	S00299	S00082
Catchment_size (sq. km)	22.4	8.2	14064.3	8289.3	7841.2	29.6	16.9	23.4
<b>Domestic ruminant densities (number of animals per km<sup>2</sup>)</b>								
Beef	0.0	0.0	31.8	17.8	15.7	103.6	48.1	25.9
Dairy	0.0	0.0	91.9	69.4	64.0	67.8	42.8	18.8
Deer	0.0	0.0	7.9	8.7	8.8	1.4	124.9	0.0
Sheep	0.0	0.0	115.0	92.1	90.9	141.8	184.1	664.2
<b>Soil temperature (areal proportion covered by temperature range)</b>								
Cool_Mesic	0.0	0.0	0.3	0.5	0.5	0.0	0.0	0.0
Cold_Mesic	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Mild_Mesic	0.0	0.0	0.2	0.3	0.3	0.8	0.8	1.0
Thermic	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Warm_Mesic	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
<b>Lithology (areal proportion covered by soil type)</b>								
Alluvium	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Greywacke	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.6
Igneous_Volcanics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Loess	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
Rhyolite	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Sedimentary_Rock	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tephra_Lapilli	0.0	0.0	0.6	0.6	0.7	0.9	1.0	0.0
Weathered_Mafic	0.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0
Weathered_Soft_Rocks	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0
<b>Land use (areal proportion covered by type of land usage)</b>								
Alpine	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Cropland	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
Forest	1.0	1.0	0.4	0.5	0.5	0.5	0.8	0.8
Grassland	0.0	0.0	0.5	0.4	0.4	0.4	0.2	0.2
Gravel	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Settlement	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Shrubland	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Wetland	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0

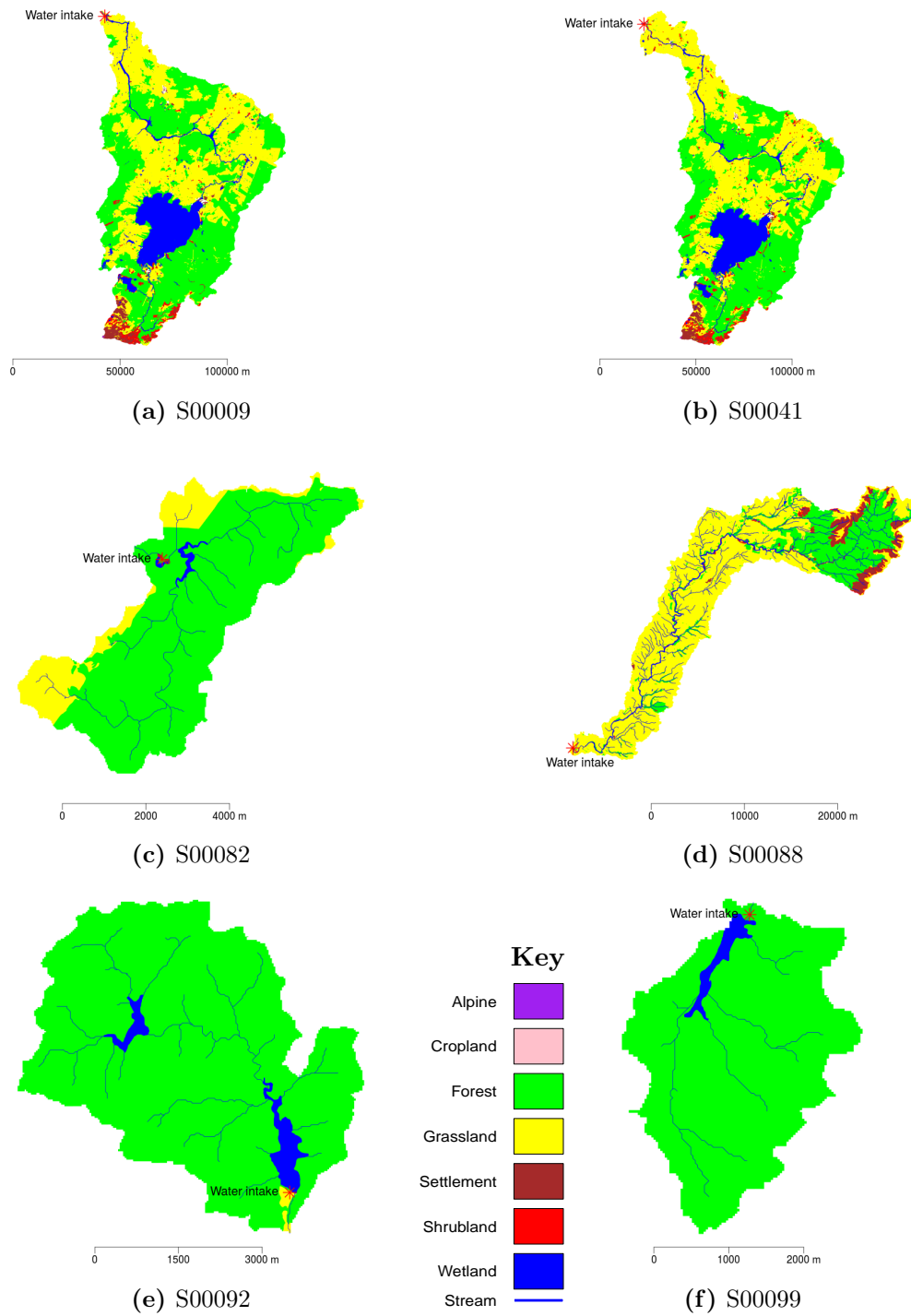
Variable	S00088	S00383	S00120	S00121	S00434	S00118	S00124	S00200
Catchment_size (sq. km)	291.7	78.4	27.5	21.0	30.3	86.9	11.5	64.5
<b>Domestic ruminant densities (number of animals per km<sup>2</sup>)</b>								
Beef	50.8	8.4	3.1	17.0	18.9	0.0	81.6	41.5
Dairy	39.0	0.0	0.0	0.0	0.0	0.0	99.3	0.0
Deer	18.7	2.8	0.0	0.0	0.0	0.0	68.3	4.6
Sheep	542.7	76.5	22.2	221.7	282.5	0.0	433.2	350.7
<b>Soil temperature (areal proportion covered by temperature range)</b>								
Cool_Mesic	0.3	0.3	0.8	0.2	0.1	0.3	0.0	0.0
Cold_Mesic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mild_Mesic	0.6	0.7	0.2	0.8	0.9	0.7	1.0	1.0
Thermic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Warm_Mesic	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>Lithology (areal proportion covered by soil type)</b>								
Alluvium	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0
Greywacke	0.3	1.0	0.9	0.9	0.9	1.0	0.0	0.9
Igneous_Volcanics	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Loess	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Rhyolite	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sedimentary_Rock	0.4	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Tephra_Lapilli	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weathered_Mafic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Weathered_Soft_Rocks	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
<b>Land use (areal proportion covered by type of land usage)</b>								
Alpine	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Cropland	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Forest	0.3	0.6	1.0	1.0	1.0	1.0	0.0	0.1
Grassland	0.6	0.2	0.0	0.0	0.0	0.0	0.7	0.9
Gravel	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Settlement	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
Shrubland	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Wetland	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

**Table A.3:** Geospatial data for the four groundwater sources monitored for *Campylobacter*, *E. coli*, *Cryptosporidium* and *Giardia* between September 2009 and March 2014, New Zealand.

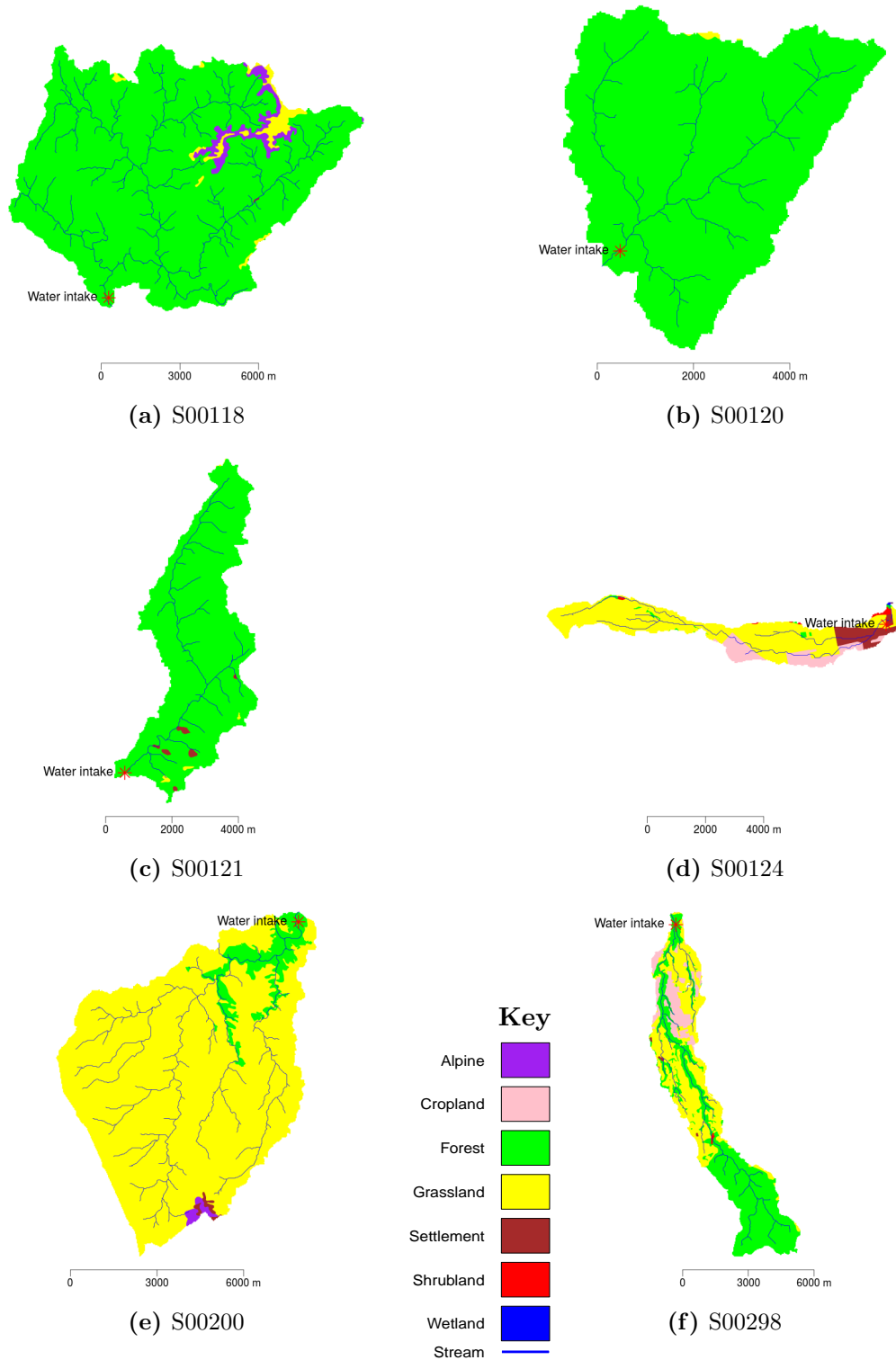
Variable	G00122	G00183	G00197	G01679
<b>Catchment size (km<sup>2</sup>)</b>				
Catchment_size	3.1	2.1	4.1	2.7
<b>Domestic ruminant densities (number of animals per km<sup>2</sup>)</b>				
Beef_density	0.0	498.9	132.1	177.6
Dairy_density	0.0	594.2	277.4	330.0
Deer_density	0.0	0.0	0.0	37.5
Sheep_density	0.0	3085.8	102.9	760.3
<b>Soil temperature (areal proportion covered by temperature range)</b>				
Cool_Mesic_prop		0.1	0.0	
Mild_Mesic_prop		0.9	1.0	
<b>Lithology (areal proportion covered by soil type)</b>				
Alluvium_prop		0.0	1.0	0.0
Igneous_Volcanics_prop		0.0	0.0	0.0
Loess_prop		0.0	0.0	0.0
Tephra_Lapilli_prop		1.0	0.0	1.0
<b>Land use (areal proportion covered by type of land usage)</b>				
Cropland_prop	0.0	0.0	0.4	0.0
Forest_prop	0.0	0.5	0.0	0.2
Grassland_prop	0.0	0.5	0.6	0.8
Settlement_prop	1.0	0.0	0.0	0.0



**Figure A.11:** Percentage of positive samples for the four study pathogens in calendar months between September 2009 and March 2014, New Zealand. No samples were collected during the month of December.

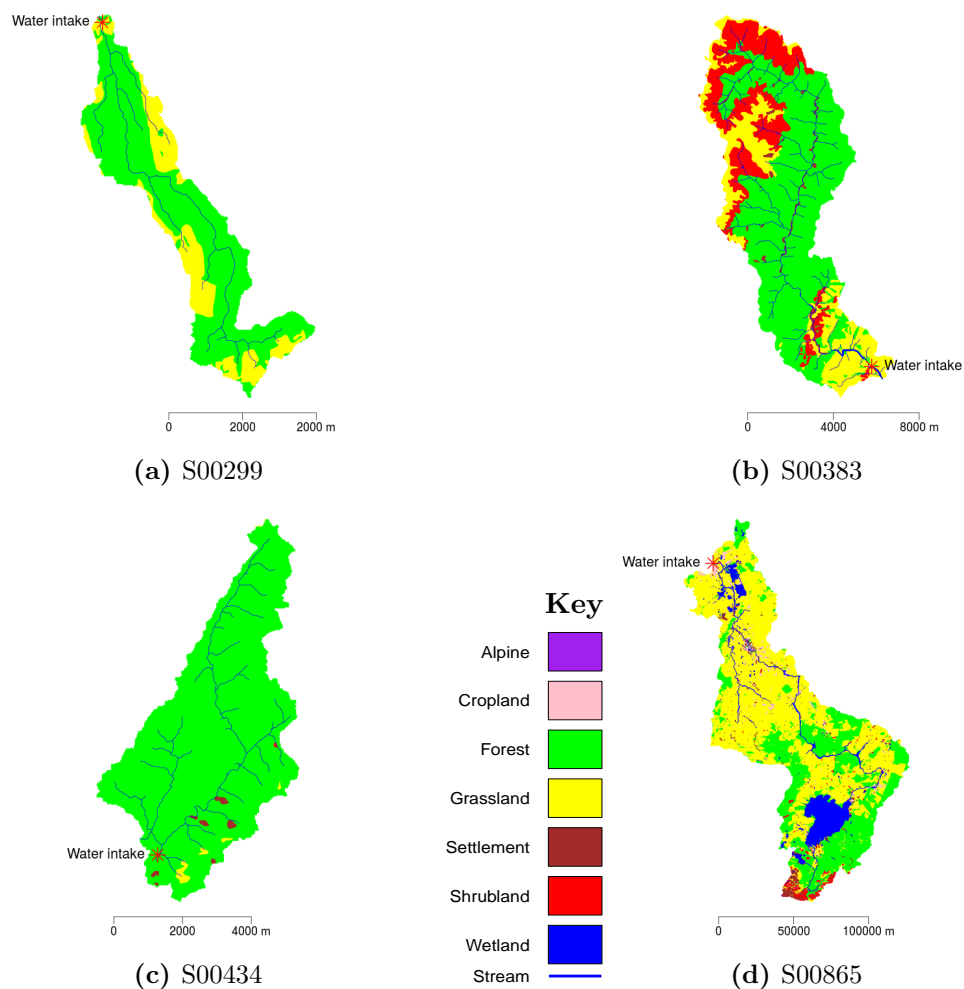


**Figure A.12:** Land cover for the first six study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.

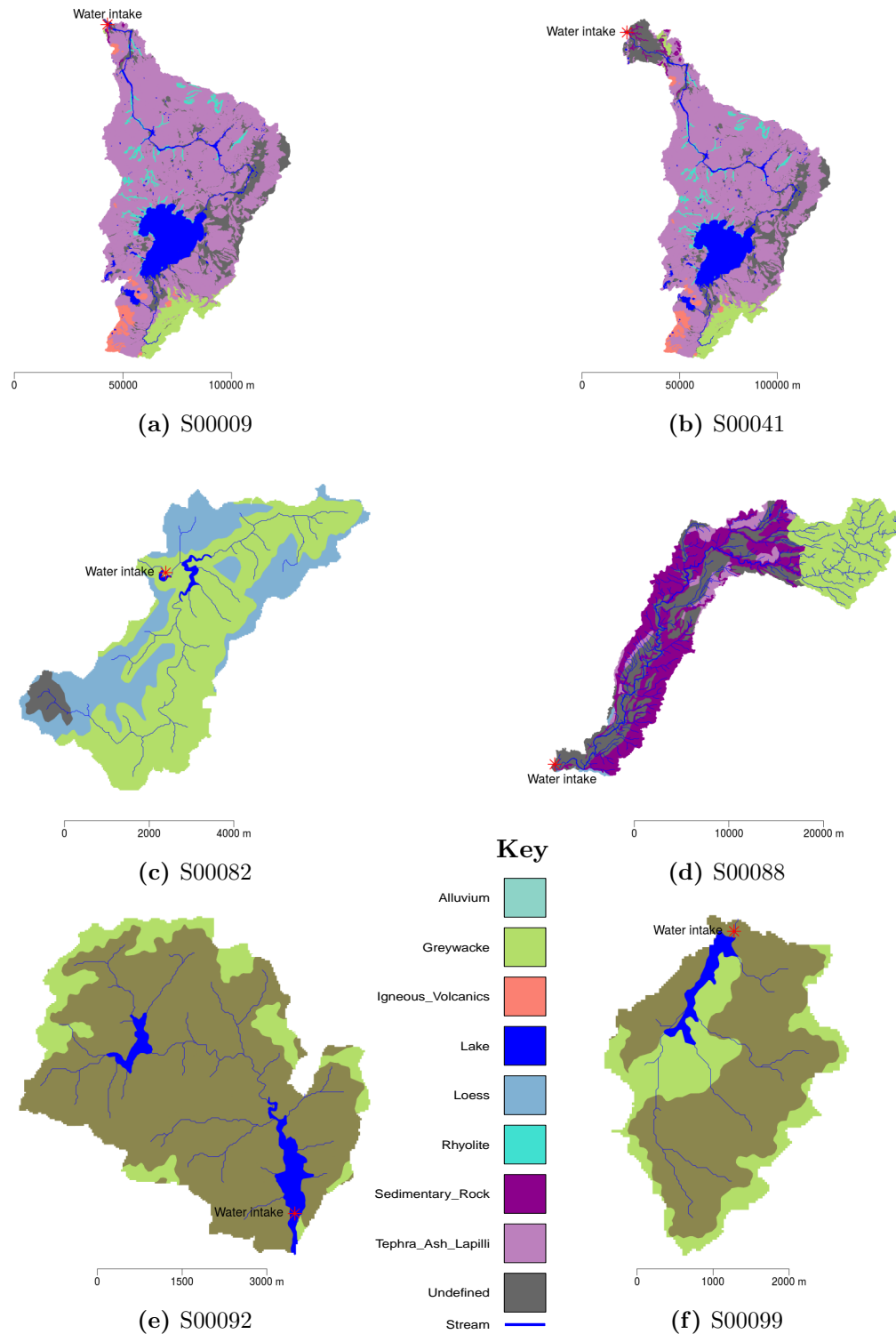


**Figure A.13:** Land cover for the second six study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.

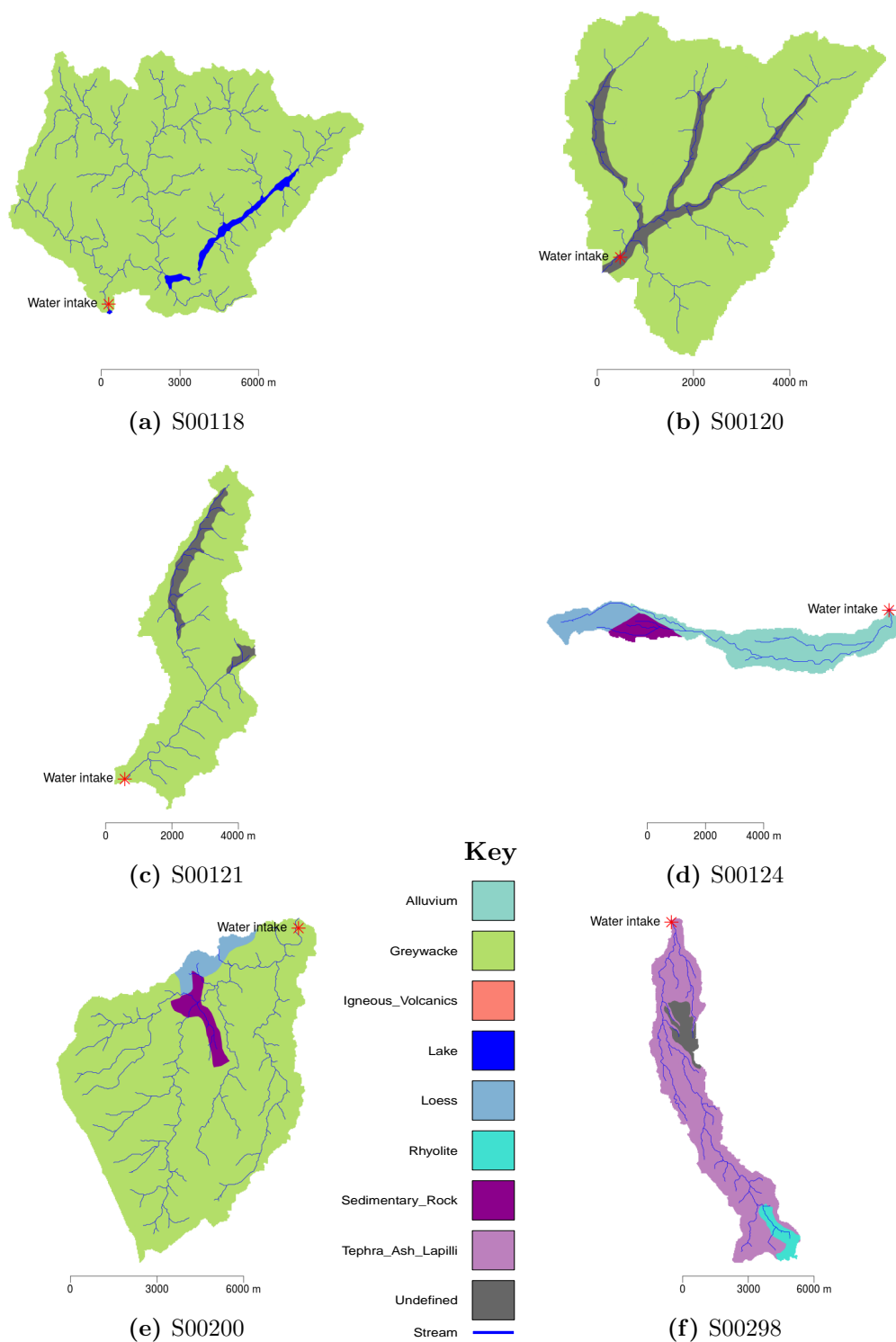




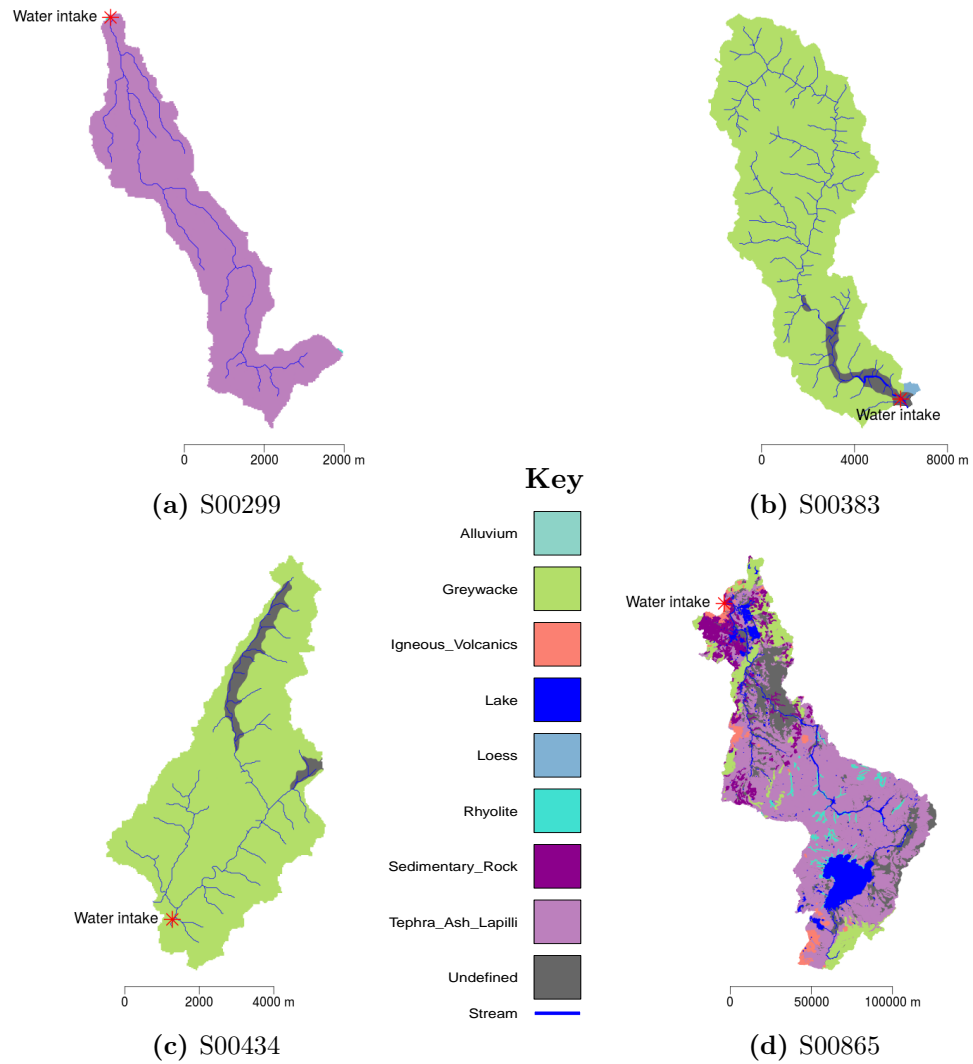
**Figure A.14:** Land cover for the last four study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.



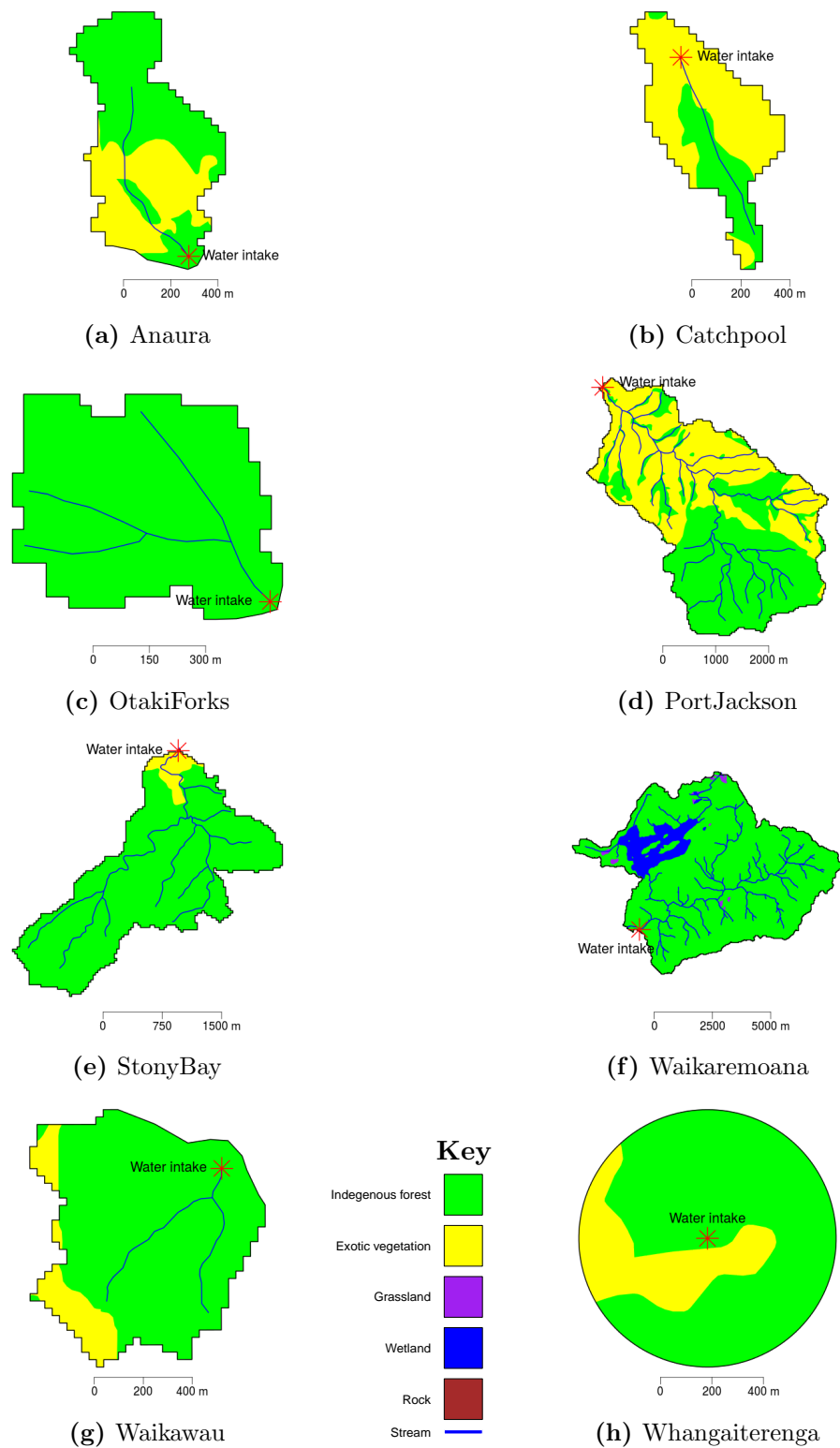
**Figure A.15:** Lithology for the first six study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.



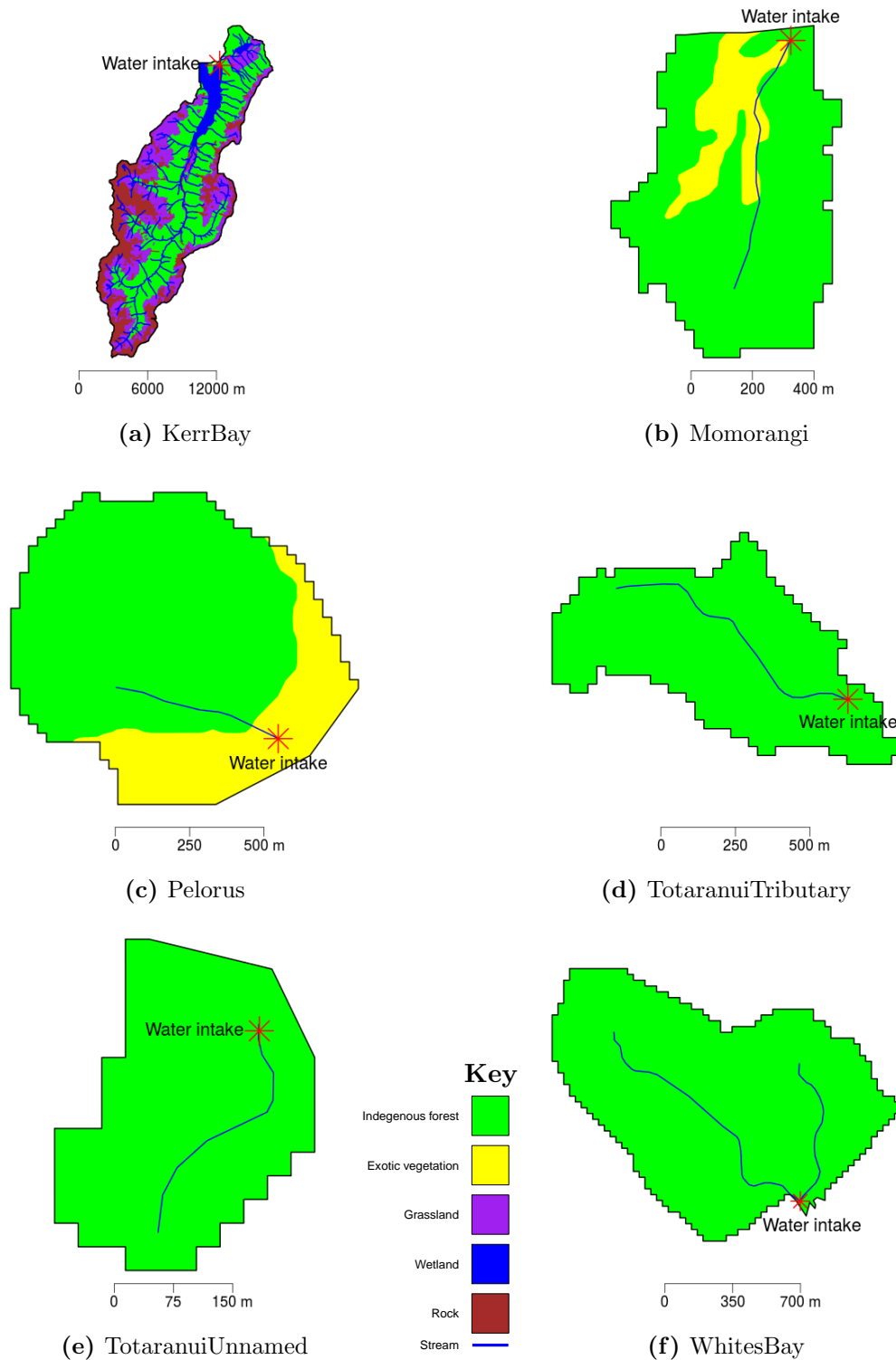
**Figure A.16:** Lithology for the second six study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.



**Figure A.17:** Lithology for the last four study catchments supplying surface raw water monitored for microbes associated with drinking water quality, September 2009–March 2014, New Zealand.



**Figure A.18:** Land cover for study campground catchments operated the Department of Conservation located in the North Island, surveyed during 2011/2012 and 2012/2013 summer seasons, New Zealand.



**Figure A.19:** Land cover for study campground catchments operated the Department of Conservation located in the South Island, surveyed during 2011/2012 and 2012/2013 summer seasons, New Zealand.

**List 1:** Preparation of crude DNA for *Campylobacter* PCR and MLST

1. Labelled one microcentrifuge (Eppendorf®) tube 'P/M' and left another unmarked.
2. Aseptically added 1000 µL of 2 % Chelex® to tube labelled 'P/M'.
3. Added loopful of pure culture, equal to two large colonies, to tube containing Chelex®.
4. Vortexed thoroughly then pierced tube cap with sterile needle.
5. Incubated tube at 100 °C (in heating block) for 10 min.
6. Removed tube from heating block and cooled to room temperature.
7. Centrifuged tube for 3 min at 13 000 *g*.
8. Transferred 400–800 µL into the second (unmarked) tube.
9. Quantified amount of DNA in the sample using Nanodrop®.

**List 2:** Preparation of glycerol cultures for long-term storage of *Campylobacter*

1. Scraped all the growth from a 48-hour pure *Campylobacter* culture, grown on a horse blood agar plate, using a sterile swab.
2. Suspended the growth in 3 mL of 15 % glycerol broth.
3. Aseptically transferred approximately 1.8 mL of broth into a 2 mL cryovial.
4. Stored cryovial at –80 °C.

**List 3:** Metagenomic DNA extraction protocol for water using the Epicentre™ Kit

1. Using forceps and scissors presoaked in 70 % ethanol, removed membrane from filter apparatus, cut into four pieces and placed along the side (near the bottom) of a 50 mL sterile conical tube.
2. Added 1 mL of Filter Wash Buffer containing 0.2 % Tween™ 20 to filter pieces in the tube to wash off microbes trapped on the membrane.
3. Vortexed tube at a low setting to rewet the filter pieces; then increased setting to the highest speed (14,000*g*) for 2 min with intermittent breaks.
4. Transferred cell suspension to a clean microcentrifuge tube, then centrifuged tube at 14,000*g* for 2 min to pellet the cells. Discarded supernatant.
5. Resuspended cell pellet in 300 µL of TE Buffer<sup>1</sup>, then added 2 µL of Ready-Lyse Lysozyme Solution and 1 µL of RNase A to cell suspension. Mixed by vortexing.
6. Incubated at 37 °C for 30 min.
7. Added 300 µL of Meta-Lysis Solution (2X) and 1 µL of Proteinase K and mixed by vortexing.
8. Incubated at 65 °C for 15 min.
9. Cooled to room temperature, then placed on ice for 3–5 min.
10. Added 350 µL of MPC Protein Precipitation Reagent and vortexed vigorously for 10 s.
11. Pelleted debris by centrifugation for 10 min at 14,000*g* in a microcentrifuge at 4 °C.
12. Transferred supernatant to a clean 1.7 mL microcentrifuge tube and discarded pellet.
13. Added 570 µL of isopropanol to supernatant. Mixed by inverting tube multiple times.
14. Pelleted DNA by centrifugation for 10 min at 14,000*g* in a microcentrifuge at 4 °C.
15. Used a pipet tip to remove isopropanol without dislodging DNA pellet. Briefly pulse-centrifuged the sample and removed any residual liquid with a pipet tip, without disturbing pellet.
16. Added 500 µL of 70 % ethanol without disturbing pellet. Then centrifuged for 10 min at 14,000*g* in a microcentrifuge at 4 °C.
17. Used a pipet tip to remove ethanol without dislodging DNA pellet. Briefly pulse-centrifuged sample and removed any residual liquid with a pipet tip, without disturbing pellet.
18. Air-dried pellet for 8 min at room temperature.
19. Resuspended DNA pellet in 50 µL of TE Buffer.
20. Validated size and concentration of the isolated DNA by comparing to Fosmid Control DNA (40 kb; 100 ng/µL) provided in the kit, via gel electrophoresis on a 2 % agarose gel. Used 2 µL of isolated DNA preparation for this analysis.

---

<sup>1</sup>Is composed of Tris and EDTA; Tris is an abbreviation for tris(hydroxymethyl)aminomethane while EDTA is an abbreviation for ethylenediaminetetraacetic acid



### List 4: Metagenomic DNA extraction protocol using NucleoSpin® kit

1. Sample preparation
  - (a) Added sample into a NucleoSpin® Bead Tube containing ceramic beads:
    - i. **Feaces** — pea-size faecal sample.
    - ii. **Water** — filter membrane cut into small pieces using forceps and scissors presoaked in 70 % ethanol.
  - (b) Added 700 µL Buffer SL1.
2. Lysis condition adjustment
  - (a) Added 150 µL Enhancer SX and closed cap.
3. Sample lysis
  - (a) Attached NucleoSpin® Bead tubes horizontally to a vortexer and vortexed samples at full speed at room temperature (18–25 °C) for 5 min.
4. Contaminant precipitation
  - (a) Centrifuged for 2 min at 11 000 *g* to eliminate foam caused by detergent.
  - (b) Transferred up to 700 µL of supernatant into lidded microcentrifuge tube.
  - (c) Added 150 µL Buffer SL3 and vortexed for 5 s.
  - (d) Incubated for 5 min at 0–4 °C.
  - (e) Centrifuged for 1 min at 11 000 *g*.
5. Lysate filtration
  - (a) Placed a NucleoSpin® Inhibitor Removal Column (red ring) into collection tube, cut off lid and kept lid.
  - (b) Loaded up to 700 µL clear supernatant of step 4 onto filter.
  - (c) Centrifuged for 1 min at 11 000 *g*.
  - (d) Discarded NucleoSpin® Inhibitor Removal Column.
6. Binding condition adjustment
  - (a) Added 250 µL Buffer SB and closed lid.
  - (b) Vortexed for 5 s.
7. DNA binding
  - (a) Placed NucleoSpin® Soil Column (green ring) in Collection Tube (2 mL) with no lid.
  - (b) Loaded 550 µL sample onto column (taking care that the lid is unattached).
  - (c) Centrifuged for 1 min at 11 000 *g*.
  - (d) Discarded flow-through and placed column back into collection tube.
  - (e) Loaded remaining sample onto column with lid on.
  - (f) Centrifuged for 1 min at 11 000 *g*.

- (g) Discarded flow-through and placed column back into collection tube.
- 8. Silica membrane washing and drying
  - (a) Added 500  $\mu$ L Buffer SB to NucleoSpin<sup>®</sup> Soil Column.
  - (b) Centrifuged for 30 s at 11 000 *g*.
  - (c) Discarded flow-through and placed column back into collection tube.
  - (d) Added 550  $\mu$ L Buffer SW1 to NucleoSpin<sup>®</sup> Soil Column.
  - (e) Centrifuged for 30 s at 11 000 *g*.
  - (f) Discarded flow-through and placed column back into collection tube.
  - (g) Added 700  $\mu$ L Buffer SW2 to NucleoSpin<sup>®</sup> Soil Column.
  - (h) Closed lid and vortex for 2 s.
  - (i) Centrifuged for 30 s at 11 000 *g*. Discarded flow-through and placed column back into collection tube.
  - (j) Repeated steps 8g to 8i.
- 9. Silica membrane drying
  - (a) Centrifuged for 2 min at 11 000 *g*.
- 10. DNA elution
  - (a) Placed NucleoSpin<sup>®</sup> Soil Column into new microcentrifuge tube (not provided in kit).
  - (b) Added 50  $\mu$ L of Buffer SE and incubated at room temperature for 1 min without closing lid.
  - (c) Centrifuged at 11 000 *g* for 30 s with lid closed.

**List 5:** Protocol for fluorescence microscopy of *Cryptosporidium* oocyst and *Giardia* cyst in faecal samples

1. Added pea-size faecal sample to a microcentrifuge tube containing 700  $\mu$ L phosphate buffered saline (PBS) and mixed thoroughly.
2. Transferred 50  $\mu$ L of supernatant onto a microscope slide and incubated at 37 °C for 30–40 min.
3. Fixed slide using 50  $\mu$ L methanol and re-incubate at 37 °C for 10 min.
4. Placed slide in a humidity chamber, added 50  $\mu$ L diluted Aqua-Glo® stain and incubated at 37 °C for a further 30–60 min.
5. Removed excess stain by gently tilting slide to one side on a paper towel.
6. Washed slide by adding 50  $\mu$ L PBS and tilting as in step 5.
7. Pipetted off excess fluid and air-dried slide for ~2 min.
8. Added one drop of mounting media and covered with a cover slip; secured cover slip with drop of nail polish on corners.
9. Examined slide for *Cryptosporidium* oocysts and *Giardia* cysts under a BX 60 fluorescence microscope.

**List 6:** Protocol for fluorescence microscopy of *Cryptosporidium* oocyst and *Giardia* cyst in water samples

1. Placed filter module, along with residual fluid, in a Stomacher<sup>®</sup> 3500 bag.
2. Dismantled filter module within Stomacher<sup>®</sup> 3500 bag to recover foam disks; set aside filter housing.
3. Added 500 mL PBS and homogenised using a Stomacher<sup>®</sup> 3500 (Seward, West Sussex, UK) for 10 min on **normal** paddling setting.
4. Transferred eluent into a 2 L container, after wringing filter disks *in-situ* to recover as much of eluent as possible.
5. Transferred eluent into a 500 mL conical centrifuge tube and centrifuged at 3000 *g* for 15 min at 10 °C using a Sorvall RT7 Benchtop centrifuge.
6. Aspirated off the top 450 mL supernatant using a venturi vacuum unit.
7. Vortexed remaining fluid to resuspend pellet collected at the bottom of centrifuge tube.
8. Transferred mixture into a 50 mL centrifuge tube and centrifuged as in step 5.
9. Aspirated supernatant as in step 6, leaving 10 mL in which pellet was resuspended.
10. Transferred mixture into a glass tube.
11. Added 1000 µL SL Buffer A, 1000 µL SL Buffer B, 100 µL anti-*Giardia* and 100 µL anti-*Cryptosporidium* magnetic beads. Mixed thoroughly.
12. Incubated tube at room temperature for 1 h while gently mixing on a tube shaker, Barnstead-/ThermoLyne Labquake<sup>®</sup> (Thermo Scientific, Massachusetts, USA).
13. Placed tube in a magnetic holder and gently inverted tube, with fluid flowing over magnet, for 2 min. Discarded fluid without disturbing beads attracted to magnet.
14. Added 50 µL PBS, inverted tube and discarded fluid as in step 13.
15. Added 1000 µL water and removed tube from magnetic holder.
16. Transferred mixture into a microcentrifuge tube using a pipette.
17. Placed microcentrifuge tube in a small magnetic holder and repeated steps 13 and 14.
18. Added 50 µL water to the microcentrifuge tube and vortexed.
19. Incubated at 80 °C in heating block for 10 min.
20. Placed microcentrifuge tube back into small magnetic holder.
21. Pipetted liquid onto a microscope slide fluorescence microscopy slide, incubated at 37 °C for 30 min or until dry.
22. Fixed slide by adding 50 µL methanol and incubated at 37 °C until dry (~5 min).
23. Placed slide into humid chamber, added 50 µL diluted Aqua-Glo<sup>®</sup> stain and incubated at 37 °C for 30–40 min.
24. Removed liquid off slide by gently tilting slide on a paper towel.

25. Returned slide to horizontal position, added 50  $\mu$ L water and tilted slide as in step 24.
26. Air-dried slide for  $\sim$ 2 min.
27. Added one drop of mounting media and covered with a cover slip; secured cover slip with drop of nail polish on corners.
28. Examined slide for *Cryptosporidium* oocysts and *Giardia* cysts under a BX 60 fluorescence microscope.

**List 7:** Multilocus sequence typing (MLST) step 1: *Campylobacter* species confirmation

1. Inspected wells for dehydration; added 25  $\mu$ L water to completely dehydrated wells or 12  $\mu$ L water to those partially dehydrated
2. Added 25  $\mu$ L PEG to all wells using a tray.
3. Covered plate and mixed.
4. Incubated at 37 °C for 15 min.
5. Span at 3000 rpm for 30 min.
6. Inverted plate onto 4–5 sheets of tissue paper and span upside down at 300 rpm for 2 min.
7. Added 150  $\mu$ L of 80 % ethanol.
8. Span at 2500 rpm for 10 min.
9. Inverted plate onto more tissue and spin upside down at 300 rpm for 2 min.
10. Air-dried plate in cupboard overnight.

**Table A.4:** Constituents of the *Campylobacter* and *Giardia* polymerase chain reaction master mixes

Details	<i>Campylobacter</i>					<i>Giardia</i>	
	<i>C. sp. nova</i> I	<i>C. coli</i>	<i>C. jejuni</i>	Others	MLST	Inner	Outer
<b>Master Mix</b>							
<b>Buffer (10X)</b>							
Volume	2.0	2.0	2.0	2.0	2.0	2.0	2.0
<b>dNTP (2.0 mM)</b>							
Volume (μL)	0.4	0.6	0.6	1.0	1.0	1.0	
<b>MgCl<sub>2</sub> (50.0 mM)</b>							
Volume (μL)	1.0	1.0	1.0	1.0	0.6	0.6	0.6
<b>Forward Primer</b>							
Concentration (pmol/μL)	2.0	2.0	2.0	2.0	2.0		
Volume (μL)	2.0	2.0	2.0	2.0	2.0	2.0	2.0
<b>Reverse Primer</b>							
Concentration (pmol/μL)	2.0	2.0	2.0	2.0	2.0		
Volume (μL)	2.0	2.0	2.0	2.0	2.0	2.0	2.0
<b>Platinum<sup>®</sup> Taq (2 units/μL)</b>							
Volume (μL)	0.2	0.2	0.2	0.2	0.2	0.2	0.2
<b>Water</b>							
Volume (μL)	10.4	10.2	10.2	9.8	10.2	5.2	7.2
<b>BSA (2.0 mg/mL)</b>							
Concentration						2.0	2.0
Volume (μL)						2.0	2.0
<b>DMSO</b>							
Volume (μL)						1.0	1.0
<b>Addition to Master Mix</b>							
<b>DNA (2.0–25 ng/μL)</b>							
Volume (μL)	2.0	2.0	2.0	2.0	2.0	4.0	2.0

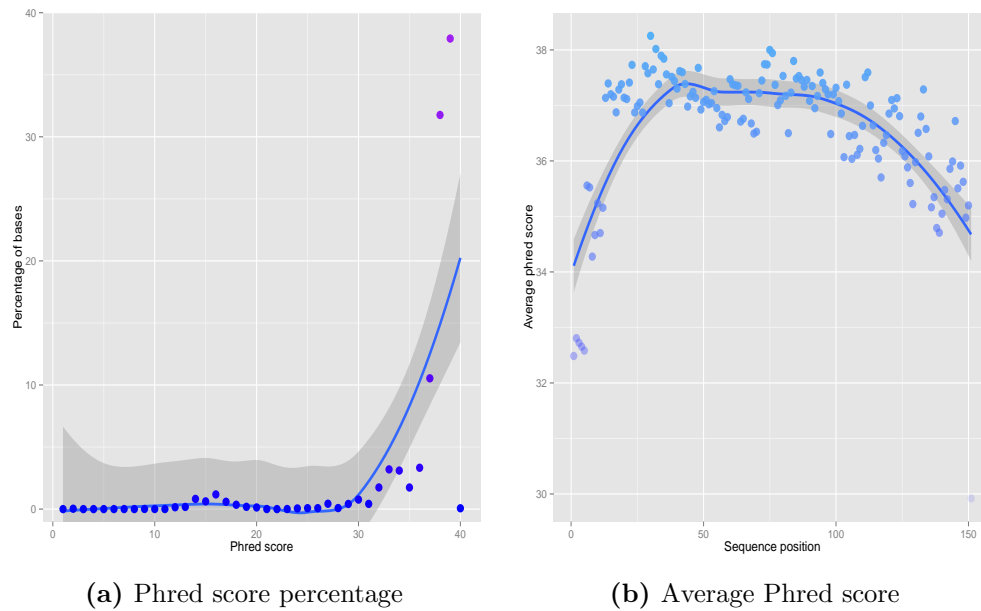
**Table A.5:** PCR conditions for selected *Campylobacter* species, *Campylobacter* genus, *Campylobacter* multilocus sequence typing and *Giardia*

Details	<i>C. sp. nova</i> I	<i>C. coli</i>	<i>C. jejuni</i>	<i>Campy</i> genus	MLST	<i>Giardia</i>
<b>Initial denaturation</b>						
Cycles	1	1	1	1	1	1
Temperature	96	95	95	95	15	95
Duration	120	120	120	120	900	800
<b>Denaturation</b>						
Cycles	40	35	40	40	35	35
Temperature	94	96	94	94	94	94
Duration	20	30	15	30	30	60
<b>Annealing</b>						
Cycles	40	35	40	40	35	35
Temperature	55	58	60	56	50	60
Duration	20	30	20	30	30	90
<b>Elongation</b>						
Cycles	40	35	40	40	35	2
Temperature	72	72	72	72	72	72
Duration	10	30	30	30	90	120
<b>Final elongation</b>						
Cycles	1	1	1	1	1	1
Temperature	72	72	72	72	72	72
Duration	120	120	120	120	420	800
<b>Holding step</b>						
Temperature	10	10	10	10	10	10
<b>Expected product size</b>						
Base pairs	106	462	603	816		

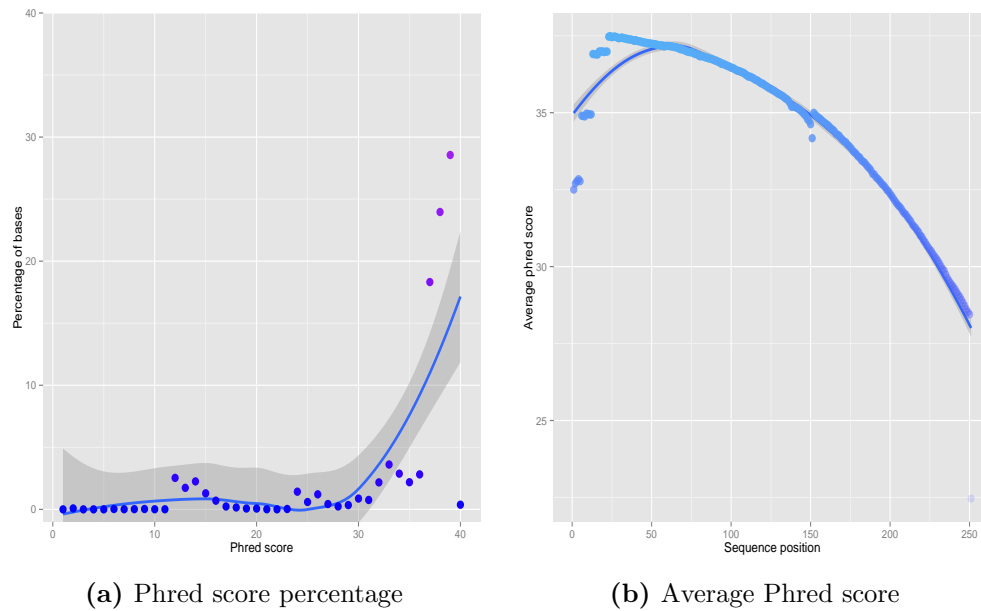


**Table A.6:** Table showing the encoding for the four bases (A, C, T, G) and encoding for ambiguous positions in a DNA-sequence.

Code	Meaning	Etymology	Complement	Opposite
A	A	Adenosine	T	B
T (or U)	T	Thymidine (or Uridine)	A	V
G	G	Guanine	C	H
C	C	Cytosine	G	D
K	G or T	Keto	M	M
M	A or C	Amino	K	K
R	A or G	Purine	Y	Y
Y	C or T	Pyrimidine	R	R
S	C or G	Strong	S	W
W	A or T	Weak	W	S
B	C or G or T	not A (B comes after A)	V	A
V	A or C or G	not T (or U) (V comes after U)	B	T/U
H	A or C or T	not G (H comes after G)	D	G
D	A or G or T	not C (D comes after C)	H	C
X/N	G or A or T or C	any	N	.
.	not G or A or T or C		.	N
-	gap of indeterminate length			



**Figure A.20:** Phred scores for 16S sequences overlaid with a smoothed estimate (solid line) and the 95% confidence interval (shaded area). The percentage of bases for each Phred score **(a)** and the average Phred score at each position on the sequence **(b)** show that the sequences were of high quality.



**Figure A.21:** Phred scores for WGS sequences overlaid with a smoothed estimate (solid line) and the 95% confidence interval (shaded area). The percentage of bases for each Phred score **(a)** and the average Phred score at each position on the sequence **(b)** show that the sequences were of high quality.

**Table A.7:** The 1-proportional similarity index values (bottom off-diagonals), with 95% confidence intervals (top off-diagonals), used for assessing the divergence of taxa related to the Family *Campylobacteraceae* in the 16S metagenome sampled from campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand.

	Passerine	Pukeko	Possum	Rabbit	Tap	Duck	Stream	Cattle	Lake	Sheep
Passerine		0.10-0.13	0.09-0.09	0.09-0.09	1.00-1.00	0.96-0.97	1.00-1.00	0.92-0.96	1.00-1.00	0.95-0.96
Pukeko	0.11		0.13-0.16	0.10-0.16	1.00-1.00	0.96-0.98	1.00-1.00	0.98-1.00	1.00-1.00	1.00-1.00
Possum	0.09	0.14		0.00-0.04	1.00-1.00	0.99-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
Rabbit	0.09	0.13	0.01		1.00-1.00	0.99-1.00	1.00-1.00	1.00-1.00	1.00-1.00	1.00-1.00
Tap	0.99	0.99	1.00	1.00		0.98-0.99	0.06-0.08	1.00-1.00	0.09-0.67	1.00-1.00
Duck	0.96	0.96	0.99	0.99	0.98		0.98-1.00	0.97-1.00	0.41-0.91	1.00-1.00
Stream	1.00	1.00	1.00	1.00	0.07	0.99		1.00-1.00	0.13-0.67	1.00-1.00
Cattle	0.94	0.99	1.00	1.00	1.00	0.99	1.00		1.00-1.00	0.02-0.05
Lake	1.00	1.00	1.00	1.00	0.33	0.66	0.38	1.00		1.00-1.00
Sheep	0.95	1.00	1.00	1.00	1.00	1.00	1.00	0.02	1.00	

**Table A.8:** The 1-proportional similarity index values (bottom off-diagonals), with 95% confidence intervals (top off-diagonals), used for assessing the divergence of taxa related to WHO-recognised pathogen extracted from 16S metagenome sampled from campgrounds operated by the Department of Conservation during the 2011/12 and 2012/13 summer seasons, New Zealand.

	Cattle	Duck	Intake	Passerine	Possum	Pukeko	Rabbit	Sheep	Tap
Cattle		0.03-0.49	0.47-0.87	0.08-0.52	0.14-0.60	0.08-0.51	0.04-0.70	0.05-0.68	0.52-0.91
Duck	0.02		0.45-0.85	0.07-0.47	0.14-0.57	0.07-0.43	0.03-0.71	0.04-0.61	0.53-0.88
Intake	0.67	0.65		0.31-0.73	0.30-0.74	0.34-0.76	0.57-0.94	0.57-0.90	0.11-0.41
Passerine	0.16	0.15	0.51		0.10-0.51	0.07-0.42	0.14-0.68	0.12-0.64	0.38-0.76
Possum	0.25	0.24	0.45	0.16		0.10-0.52	0.20-0.80	0.20-0.72	0.36-0.76
Pukeko	0.15	0.13	0.52	0.08	0.18		0.15-0.74	0.13-0.67	0.40-0.79
Rabbit	0.25	0.26	0.73	0.30	0.38	0.33		0.00-0.56	0.66-0.94
Sheep	0.23	0.24	0.73	0.27	0.36	0.31	0.02		0.64-0.94
Tap	0.72	0.70	0.09	0.56	0.52	0.57	0.78	0.78	





# References

1. Acevedo, M. F. (2013). *Data analysis and statistics for geography, environmental science, and Engineering*. CRC Press. ISBN: 978-1-4e98-8501-7. URL: <http://www.crcpress.com> (see pp. 80, 81).
2. Ahmed, W. et al. (2012). “Fecal indicators and zoonotic pathogens in household drinking water taps fed from rainwater tanks in Southeast Queensland, Australia”. English. In: *Applied and Environmental Microbiology* 78.1, pp. 219–226. ISSN: 00992240. DOI: [10.1128/AEM.06554-11](https://doi.org/10.1128/AEM.06554-11) (see p. 95).
3. Al Mawly, J et al. (2014). “Prevalence of endemic enteropathogens of calves in New Zealand dairy farms”. In: *New Zealand veterinary journal* just-accepted, pp. 1–18 (see p. 71).
4. Alimentarius, C. (2003). *Hazard analysis and critical control point (HACCP) system and guidelines for its application* (see p. 4).
5. Allen, H. K. et al. (2013). “Estimation of viral richness from shotgun metagenomes using a frequency count approach”. In: *Microbiome* 1.1, p. 5 (see p. 156).
6. Almon, S. (1965). “The distributed lag between capital appropriations and expenditures”. In: *Econometrica: Journal of the Econometric Society*, pp. 178–196 (see pp. 85, 95).
7. Altschul, S. et al. (1990). “Basic local alignment search tool”. English. In: *Journal of Molecular Biology* 215.3, pp. 403–410. ISSN: 00222836. DOI: [10.1006/jmbi.1990.9999](https://doi.org/10.1006/jmbi.1990.9999) (see p. 161).
8. Amann, R. I., W. Ludwig, and K. H. Schleifer (1995). “Phylogenetic identification and in situ detection of individual microbial cells without cultivation”. In: *Microbiological Reviews* 59.1, pp. 143–169 (see p. 34).
9. Amar, C. et al. (2002). “Sensitive PCR-restriction fragment length polymorphism assay for detection and genotyping of *Giardia duodenalis* in human feces”. English. In: *Journal of Clinical Microbiology* 40.2, pp. 446–452. ISSN: 00951137. DOI: [10.1128/JCM.40.2.446-452.2002](https://doi.org/10.1128/JCM.40.2.446-452.2002) (see p. 129).
10. Ames, S. et al. (2013). “Scalable metagenomic taxonomy classification using a reference genome database”. English. In: *Bioinformatics* 29.18, pp. 2253–2260. ISSN: 13674803. DOI: [10.1093/bioinformatics/btt389](https://doi.org/10.1093/bioinformatics/btt389) (see p. 29).
11. Anderson, I., M. Rhodes, and H. Kator (1979). “Sublethal stress in *Escherichia coli*: A function of salinity”. English. In: *Applied and Environmental Microbiology* 38.6, pp. 1147–1152. ISSN: 00992240 (see p. 30).
12. Anderson, K., J. Whitlock, and V. Harwood (2005). “Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments”. English. In: *Applied and Environmental Microbiology* 71.6, pp. 3041–3048. ISSN: 00992240. DOI: [10.1128/AEM.71.6.3041-3048.2005](https://doi.org/10.1128/AEM.71.6.3041-3048.2005) (see p. 30).
13. Anderson, M. and I. Schrijver (2010). “Next generation DNA sequencing and the future of genomic medicine”. English. In: *Genes* 1.1, pp. 38–69. ISSN: 20734425. DOI: [10.3390/genes1010038](https://doi.org/10.3390/genes1010038) (see p. 7).
14. Applied Maths NV (2014). *BioNumerics manual version 7.0*. Sint-Martens-Latem, Belgium. URL: <http://www.applied-maths.com/> (see p. 161).
15. Armstrong, B. (2006). “Models for the relationship between ambient temperature and daily mortality”. English. In: *Epidemiology* 17.6, pp. 624–631. ISSN: 10443983. DOI: [10.1097/01.ede.0000239732.50999.8f](https://doi.org/10.1097/01.ede.0000239732.50999.8f) (see p. 95).

16. Arnell, N. (2004). "Climate change and global water resources: SRES emissions and socio-economic scenarios". English. In: *Global Environmental Change* 14.1, pp. 31–52. ISSN: 09593780. DOI: [10.1016/j.gloenvcha.2003.10.006](#) (see p. 13).
17. Arvelo, W. et al. (2012). "Norovirus outbreak of probable waterborne transmission with high attack rate in a Guatemalan resort". English. In: *Journal of Clinical Virology* 55.1, pp. 8–11. ISSN: 13866532. DOI: [10.1016/j.jcv.2012.02.018](#) (see pp. 106, 129).
18. Ashbolt, N. J., W. O. Grabow, and M. Snozzi (2001). *Water quality: Guidelines, standards and health*. Ed. by L. Fewtrell and J. Bartram. IWA Publishing, pp. 289–316 (see p. 23).
19. Auld, H., D. MacIver, and J. Klaassen (2004). "Heavy rainfall and waterborne disease outbreaks: The Walkerton example". In: *Journal of Toxicology and Environmental Health-Part a-Current Issues* 67.20-22, pp. 1879–1887. ISSN: 1528-7394 (see pp. 14, 75).
20. Australia National Notifiable Disease Surveillance System (2014). Department of Health. URL: <http://www9.health.gov.au/cda/source/cda-index.cfm> (see p. 75).
21. Babic, M., A. Hujer, and R. Bonomo (2006). "What's new in antibiotic resistance? Focus on beta-lactamases". English. In: *Drug Resistance Updates* 9.3, pp. 142–156. ISSN: 13687646. DOI: [10.1016/j.drug.2006.05.005](#) (see pp. 43, 157).
22. Baca-Garcia, E. et al. (2007). "Variables associated with familial suicide attempts in a sample of suicide attempters". English. In: *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 31.6, pp. 1312–1316. ISSN: 02785846. DOI: [10.1016/j.pnpb.2007.05.019](#) (see pp. 69, 70).
23. Bai, Y. et al. (2013). "Integrated metagenomic and physiochemical analyses to evaluate the potential role of microbes in the sand filter of a drinking water treatment system". English. In: *PLoS ONE* 8.4. ISSN: 19326203. DOI: [10.1371/journal.pone.0061011](#) (see pp. 36, 39, 135, 137).
24. Baker, M. G. et al. (2012). "Increasing incidence of serious infectious diseases and inequalities in New Zealand: A national epidemiological study". In: *Lancet* 379.9821, pp. 1112–1119. ISSN: 0140-6736 (see pp. 16, 45).
25. Ball, A. (2007). *Estimation of the burden of water-borne disease in New Zealand: Preliminary report*. <http://www.health.govt.nz/publication/estimation-burden-water-borne-disease-new-zealand-preliminary-report> (see p. 15).
26. Ballantine, D. and R. Davies-Colley (2014). "Water quality trends in New Zealand rivers: 1989-2009". English. In: *Environmental Monitoring and Assessment* 186.3, pp. 1939–1950. ISSN: 01676369. DOI: [10.1007/s10661-013-3508-5](#) (see p. 1).
27. Baltrus, D. et al. (2009). "The complete genome sequence of *Helicobacter pylori* strain G27". English. In: *Journal of Bacteriology* 91.1, pp. 447–448. ISSN: 00219193. DOI: [10.1128/JB.01416-08](#) (see p. 136).
28. Bates, D. et al. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. URL: <http://CRAN.R-project.org/package=lme4> (see pp. 57, 120).
29. Baudart, J. and P. Lebaron (2010). "Rapid detection of *Escherichia coli* in waters using fluorescent in situ hybridization, direct viable counting and solid phase cytometry". English. In: *Journal of Applied Microbiology* 109.4, pp. 1253–1264. ISSN: 13645072. DOI: [10.1111/j.1365-2672.2010.04752.x](#) (see p. 25).
30. Baudart, J. et al. (2002). "Rapid and sensitive enumeration of viable diluted cells of members of the Family *Enterobacteriaceae* in freshwater and drinking water". English. In: *Applied and Environmental Microbiology* 68.10, pp. 5057–5063. ISSN: 00992240. DOI: [10.1128/AEM.68.10.5057-5063.2002](#) (see p. 25).
31. Beaudeau, P. et al. (2014). "A time series study of gastroenteritis and tap water quality in the Nantes area, France, 2002-2007". English. In: *Journal of Exposure Science and Environmental Epidemiology* 24.2, pp. 192–199. ISSN: 15590631. DOI: [10.1038/jes.2013.5](#) (see pp. 76, 77, 94).

32. Bej, A. et al. (1990). "Detection of coliform bacteria in water by polymerase chain reaction and gene probes". English. In: *Applied and Environmental Microbiology* 56.2, pp. 307–314. ISSN: 00992240 (see p. 26).
33. Bengraïne, K. and T. Marhaba (2003). "Using principal component analysis to monitor spatial and temporal changes in water quality". English. In: *Journal of Hazardous Materials* 100.1-3, pp. 179–195. ISSN: 03043894. DOI: [10.1016/S0304-3894\(03\)00104-3](https://doi.org/10.1016/S0304-3894(03)00104-3) (see p. 46).
34. Bentley, D. et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". English. In: *Nature* 456.7218, pp. 53–59. ISSN: 00280836. DOI: [10.1038/nature07517](https://doi.org/10.1038/nature07517) (see p. 8).
35. Berk, R. A. (2003). Thousand Oaks, CA : Sage Publications. ISBN: 0761929045 (see p. 117).
36. Berke, O. (2004). "Exploratory disease mapping: Kriging the spatial risk function from regional count data". English. In: *International Journal of Health Geographics* 3. ISSN: 1476072X. DOI: [10.1186/1476-072X-3-18](https://doi.org/10.1186/1476-072X-3-18) (see p. 95).
37. Betancourt, W. and J. Rose (2004). "Drinking water treatment processes for removal of *Cryptosporidium* and *Giardia*". English. In: *Veterinary Parasitology* 126.1-2 SPEC.ISS. Pp. 219–234. ISSN: 03044017. DOI: [10.1016/j.vetpar.2004.09.002](https://doi.org/10.1016/j.vetpar.2004.09.002) (see pp. 15, 20, 21).
38. Biggs, P. et al. (2011). "Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence type reveals multiple loci of different ancestral lineage". English. In: *PLoS ONE* 6.11. ISSN: 19326203. DOI: [10.1371/journal.pone.0027121](https://doi.org/10.1371/journal.pone.0027121) (see p. 136).
39. Biryukov, A. et al. (2005). "New physical methods of disinfection of water". English. In: *Journal of Russian Laser Research* 26.1, pp. 13–25. ISSN: 10712836. DOI: [10.1007/s10946-005-0002-8](https://doi.org/10.1007/s10946-005-0002-8) (see p. 21).
40. Blaser, M. (2008). "Understanding microbe-induced cancers". English. In: *Cancer Prevention Research* 1.1, pp. 15–20. ISSN: 19406207. DOI: [10.1158/1940-6207.CAPR-08-0024](https://doi.org/10.1158/1940-6207.CAPR-08-0024) (see p. 135).
41. Blattner, F. et al. (1997). "The complete genome sequence of *Escherichia coli* K-12". English. In: *Science* 277.5331, pp. 1453–1462. ISSN: 00368075. DOI: [10.1126/science.277.5331.1453](https://doi.org/10.1126/science.277.5331.1453) (see p. 136).
42. Boccia, D. et al. (2002). "Waterborne outbreak of norwalk-like virus gastroenteritis at a tourist resort, Italy". English. In: *Emerging Infectious Diseases* 8.6, pp. 563–568. ISSN: 10806040 (see pp. 106, 129).
43. Bohmer, P. (1997). "Outbreak of campylobacteriosis at a school camp linked to water supply". English. In: *New Zealand Public Health Report* 4.8, pp. 58–59. ISSN: 11730250 (see pp. 31, 106, 129, 135).
44. Bonadonna, L. et al. (2002). "Occurrence of *Cryptosporidium* oocysts in sewage effluents and correlation with microbial, chemical and physical water variables". English. In: *Environmental Monitoring and Assessment* 75.3, pp. 241–252. ISSN: 01676369. DOI: [10.1023/A:1014852201424](https://doi.org/10.1023/A:1014852201424) (see p. 156).
45. Borrego, J. et al. (1987). "Coliphages as an indicator of faecal pollution in water. Its relationship with indicator and pathogenic microorganisms". English. In: *Water Research* 21.12, pp. 1473–1480. ISSN: 00431354 (see p. 22).
46. Boulware, D. (2004). "Influence of hygiene on gastrointestinal illness among wilderness backpackers". English. In: *Journal of Travel Medicine* 11.1, pp. 27–33. ISSN: 11951982 (see p. 106).
47. Bouwer, E. J. and P. B. Crowe (1988). "Biological processes in drinking water treatment". In: *Journal (American Water Works Association)* 80.9, pp. 82–93. URL: <http://www.jstor.org/stable/41292287> (see p. 15).
48. Breiman, L. (2001). "Random forests". English. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 08856125. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324) (see p. 52).
49. Breiman, L. et al. (1984). *Classification and regression trees*. The Wadsworth Statistics/Probability Series. Wadsworth International Group (see p. 52).



50. Bridle, H. (2013). *Waterborne pathogens: Detection methods and applications*. English. Elsevier B.V., pp. 1–401. ISBN: 9780444595430 (see p. 3).
51. Brieseman, M. (1987). “Town water supply as the cause of an outbreak of *Campylobacter* infection”. English. In: *New Zealand Medical Journal* 100.821, pp. 212–213. ISSN: 00288446 (see p. 31).
52. Britton, E. et al. (2010a). “Positive association between ambient temperature and salmonellosis notifications in New Zealand, 1965–2006”. English. In: *Australian and New Zealand Journal of Public Health* 34.2, pp. 126–129. ISSN: 13260200. DOI: [10.1111/j.1753-6405.2010.00495.x](https://doi.org/10.1111/j.1753-6405.2010.00495.x) (see p. 94).
53. — (2010b). “The impact of climate variability and change on cryptosporidiosis and giardiasis rates in New Zealand”. English. In: *Journal of Water and Health* 8.3, pp. 561–571. ISSN: 14778920. DOI: [10.2166/wh.2010.049](https://doi.org/10.2166/wh.2010.049) (see pp. 33, 94).
54. Brown, T. J. et al. (1992). “Presence and distribution of *Giardia* cysts in New Zealand waters”. In: (see p. 14).
55. Brunkard, J. et al. (2011). “Surveillance for waterborne disease outbreaks associated with drinking water - United States, 2007–2008”. English. In: *Morbidity and Mortality Weekly Report* 60.SS-12, pp. 38–68. ISSN: 01492195 (see p. 16).
56. Brunstein, J. (2013). “PCR: The basics of the polymerase chain reaction.” English. In: *MLO: medical laboratory observer* 45.4, pp. 32, 34–35. ISSN: 05807247 (see p. 25).
57. Bryant, D. and V. Moulton (2004). “Neighbor-Net: An agglomerative method for the construction of phylogenetic networks”. English. In: *Molecular Biology and Evolution* 21.2, pp. 255–265. ISSN: 07374038. DOI: [10.1093/molbev/msh018](https://doi.org/10.1093/molbev/msh018) (see p. 142).
58. Calderon, R. (2000). “The epidemiology of chemical contaminants of drinking water”. English. In: *Food and Chemical Toxicology* 38.SUPPL.1, S13–S20. ISSN: 02786915 (see p. 13).
59. Caporaso, J. et al. (2010). “QIIME allows analysis of high-throughput community sequencing data”. English. In: *Nature Methods* 7.5, pp. 335–336. ISSN: 15487091. DOI: [10.1038/nmeth.f.303](https://doi.org/10.1038/nmeth.f.303) (see pp. 161, 166).
60. Carpi, G. et al. (2011). “Metagenomic profile of the bacterial communities associated with *Ixodes ricinus* ticks”. English. In: *PLoS ONE* 6.10. ISSN: 19326203. DOI: [10.1371/journal.pone.0025604](https://doi.org/10.1371/journal.pone.0025604) (see p. 155).
61. Chakravorty, S. et al. (2007). “A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria”. English. In: *Journal of Microbiological Methods* 69.2, pp. 330–339. ISSN: 01677012. DOI: [10.1016/j.mimet.2007.02.005](https://doi.org/10.1016/j.mimet.2007.02.005) (see p. 35).
62. Chao, A. (1984). “Nonparametric estimation of the number of classes in a population”. In: *Scandinavian Journal of statistics*, pp. 265–270 (see p. 142).
63. Chao, Y. et al. (2013). “Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment”. English. In: *Scientific Reports* 3. ISSN: 20452322. DOI: [10.1038/srep03550](https://doi.org/10.1038/srep03550) (see pp. 36, 39, 137).
64. Chatterjee, S. and J. Simonoff (2013). *Handbook of regression analysis*. English. John Wiley and Sons. ISBN: 9780470887165. DOI: [10.1002/9781118532843](https://doi.org/10.1002/9781118532843) (see p. 118).
65. Cheng, M. et al. (1989). “Mitochondrial heat-shock protein *hsp60* is essential for assembly of proteins imported into yeast mitochondria”. English. In: *Nature* 337.6208, pp. 620–625. ISSN: 00280836 (see p. 35).
66. Christiansen, T. et al. (2012). *Programming Perl*. 4th Edition. O’Reilly Media. ISBN: 978-0-596-00492-7 (see p. 161).
67. Cinque, K. and N. Jayasuriya (2010). “Catchment process affecting drinking water quality, including the significance of rainfall events, using factor analysis and event mean concentrations”. English. In: *Journal of Water and Health* 8.4, pp. 751–763. ISSN: 14778920. DOI: [10.2166/wh.2010.162](https://doi.org/10.2166/wh.2010.162) (see pp. 45, 95).

68. Close, M. et al. (2008). "Microbial groundwater quality and its health implications for a border-strip irrigated dairy farm catchment, South Island, New Zealand". English. In: *Journal of Water and Health* 6.1, pp. 83–98. ISSN: 14778920. DOI: [10.2166/wh.2007.020](#) (see pp. 33, 70, 72).
69. Close, M. et al. (2010). "Microbial transport from dairying under two spray-irrigation systems in Canterbury, New Zealand". English. In: *Journal of Environmental Quality* 39.3, pp. 824–833. ISSN: 00472425. DOI: [10.2134/jeq2009.0208](#) (see p. 70).
70. Cole, S. et al. (1998). "Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence". English. In: *Nature* 393.6685, pp. 537–544. ISSN: 00280836. DOI: [10.1038/31159](#) (see p. 136).
71. Cornwell, D. et al. (2003). "Demonstrating *Cryptosporidium* removal using spore monitoring at lime-softening plants". English. In: *Journal / American Water Works Association* 95.5, pp. 124–133. ISSN: 0003150X (see p. 20).
72. Costet, N. et al. (2011). "Water disinfection by-products and bladder cancer: Is there a European specificity? A pooled and meta-analysis of European case-control studies". English. In: *Occupational and Environmental Medicine* 68.5, pp. 379–385. ISSN: 13510711. DOI: [10.1136/oem.2010.062703](#) (see p. 2).
73. Cowie, G. and A. Bell (2013). "A retrospective review of notified human cryptosporidiosis cases in the Waikato region of New Zealand, 2004 to 2011". English. In: *New Zealand Medical Journal* 126.1383. ISSN: 00288446 (see p. 33).
74. Cox, M., D. Peterson, and P. Biggs (2010). "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data". English. In: *BMC Bioinformatics* 11. ISSN: 14712105. DOI: [10.1186/1471-2105-11-485](#) (see pp. 141, 161).
75. Craun, G. F. and R. L. Calderon (2006). "Observational epidemiologic studies of endemic waterborne risks: Cohort, case-control, time-series, and ecologic studies". In: *Journal of Water and Health* 4.SUPPL. 2, pp. 101–120 (see p. 15).
76. Craun, M. et al. (2006). "Waterborne outbreaks reported in the United States". English. In: *Journal of Water and Health* 4.SUPPL. 2, pp. 19–30. ISSN: 14778920. DOI: [10.2166/wh.2006.016](#) (see p. 45).
77. Cressie, N. (1990). "The origins of kriging". English. In: *Mathematical Geology* 22.3, pp. 239–252. ISSN: 08828121. DOI: [10.1007/BF00889887](#) (see p. 81).
78. Crim, S. et al. (2014). "Incidence and trends of infection with pathogens transmitted commonly through food — foodborne diseases active surveillance network, 10 U.S. Sites, 2006–2013". English. In: *Morbidity and Mortality Weekly Report* 63.15, pp. 328–332. ISSN: 1545861X (see p. 75).
79. Croce, O. et al. (2014). "Draft genome sequence of *Mycobacterium austroafricanum* DSM 44191". In: *Genome announcements* 2.2, e00317–14 (see pp. 149, 151).
80. Şen, Z. (2009). *Spatial modeling principles in earth sciences*. English. Springer Netherlands, pp. 1–351. ISBN: 9781402096716. DOI: [10.1007/978-1-4020-9672-3](#) (see pp. 81, 82).
81. Curko, J. et al. (2013). "Treatment of spent filter backwash water from drinking water treatment with immersed ultrafiltration membranes". English. In: *Desalination and Water Treatment* 51.25–27, pp. 4901–4906. ISSN: 19443994. DOI: [10.1080/19443994.2013.774142](#) (see p. 20).
82. Curriero, F. C. et al. (2001). "The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994". In: *American Journal of Public Health* 91.8, pp. 1194–1199. ISSN: 0090-0036 (see p. 75).
83. Daniel, R. (2005). "The metagenomics of soil". English. In: *Nature Reviews Microbiology* 3.6, pp. 470–478. ISSN: 17401526. DOI: [10.1038/nrmicro1160](#) (see p. 135).
84. Davis, J. C. and R. J. Sampson (2002). *Statistics and data analysis in geology*. Vol. 3. Wiley New York (see pp. 79, 80, 82).

85. De La Rica, R. and M. Stevens (2012). “Plasmonic ELISA for the ultrasensitive detection of disease biomarkers with the naked eye”. English. In: *Nature Nanotechnology* 7.12, pp. 821–824. ISSN: 17483387. DOI: [10.1038/nnano.2012.186](#) (see p. 25).
86. Dechesne, M. and E. Soyeux (2007). “Assessment of source water pathogen contamination”. English. In: *Journal of Water and Health* 5.SUPPL. 1, pp. 39–50. ISSN: 14778920. DOI: [10.2166/wh.2006.055](#) (see pp. 69, 70, 128).
87. Delafont, V. et al. (2013). “Microbiome of free-living amoebae isolated from drinking water”. English. In: *Water Research*. ISSN: 00431354. DOI: [10.1016/j.watres.2013.07.047](#) (see pp. 36, 39, 137).
88. Delpla, I. et al. (2009). “Impacts of climate change on surface water quality in relation to drinking water production”. English. In: *Environment International* 35.8, pp. 1225–1233. ISSN: 01604120. DOI: [10.1016/j.envint.2009.07.001](#) (see p. 13).
89. Denis, M. et al. (2001). “*Campylobacter* contamination in French chicken production from farm to consumers. Use of a PCR assay for detection and identification of *Campylobacter jejuni* and *C. coli*”. English. In: *Journal of Applied Microbiology* 91.2, pp. 255–267. ISSN: 13645072. DOI: [10.1046/j.1365-2672.2001.01380.x](#) (see p. 113).
90. Derrien, M. et al. (2012). “Origin of fecal contamination in waters from contrasted areas: Stanols as microbial source tracking markers”. English. In: *Water Research* 46.13, pp. 4009–4016. ISSN: 00431354. DOI: [10.1016/j.watres.2012.05.003](#) (see p. 30).
91. Devane, M. et al. (2007). “A PCR marker for detection in surface waters of faecal pollution derived from ducks”. English. In: *Water Research* 41.16, pp. 3553–3560. ISSN: 00431354. DOI: [10.1016/j.watres.2007.06.043](#) (see p. 30).
92. Diggle, P. and P. J. Ribeiro (2007). *Model-based geostatistics*. Springer. ISBN: 10: 0-387-32907-2 (see p. 83).
93. Dingle, K. et al. (2001). “Multilocus sequence typing system for *Campylobacter jejuni*”. English. In: *Journal of Clinical Microbiology* 39.1, pp. 14–23. ISSN: 00951137. DOI: [10.1128/JCM.39.1.14-23.2001](#) (see p. 110).
94. Dinsdale, E. et al. (2013). “Multivariate analysis of functional metagenomes”. English. In: *Frontiers in Genetics* 4.APR. ISSN: 16648021. DOI: [10.3389/fgene.2013.00041](#) (see p. 69).
95. Donnison, A., C. Ross, and B. Thorrold (2004). “Impact of land use on the faecal microbial quality of hill-country streams”. English. In: *New Zealand Journal of Marine and Freshwater Research* 38.5, pp. 845–855. ISSN: 00288330 (see pp. 33, 71).
96. Drawz, S. and R. Bonomo (2010). “Three decades of  $\beta$ -lactamase inhibitors”. English. In: *Clinical Microbiology Reviews* 23.1, pp. 160–201. ISSN: 08938512. DOI: [10.1128/CMR.00037-09](#) (see p. 43).
97. DuBois, P. (2008). *MySQL*. Fourth. New Jersey: Addison-Wesley. ISBN: 978-0-672-32916-6 (see pp. 78, 161).
98. Duncanson, M. et al. (2000). “Rates of notified cryptosporidiosis and quality of drinking water supplies in Aotearoa, New Zealand”. In: *Water Research* 34.15, pp. 3804–3812. ISSN: 0043-1354 (see p. 33).
99. Eberhart-Phillips, J. et al. (1997). “Campylobacteriosis in New Zealand: Results of a case-control study”. English. In: *Journal of Epidemiology and Community Health* 51.6, pp. 686–691. ISSN: 0143005X (see pp. 32, 71, 95, 97).
100. Edzwald, J. (1993). “Coagulation in drinking water treatment: Particles, organics and coagulants”. English. In: *Water Science and Technology* 27.11, pp. 21–35. ISSN: 02731223 (see p. 20).
101. — (1995). “Principles and applications of dissolved air flotation”. English. In: *Water Science and Technology* 31.3-4, pp. 1–23. ISSN: 02731223. DOI: [10.1016/0273-1223\(95\)00200-7](#) (see p. 20).
102. — (2007). “Developments of high rate dissolved air flotation for drinking water treatment”. English. In: *Journal of Water Supply: Research and Technology - AQUA* 56.6-7, pp. 399–409. ISSN: 00037214. DOI: [10.2166/aqua.2007.013](#) (see p. 20).

103. — (2010). “Dissolved air flotation and me”. English. In: *Water Research* 44.7, pp. 2077–2106. ISSN: 00431354. DOI: [10.1016/j.watres.2009.12.040](https://doi.org/10.1016/j.watres.2009.12.040) (see p. 20).
104. Efron, B. and R. Tibshirani (1986). “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy”. In: *Statistical science*, pp. 54–75 (see p. 143).
105. Eid, J. et al. (2009). “Real-time DNA sequencing from single polymerase molecules”. English. In: *Science* 323.5910, pp. 133–138. ISSN: 00368075. DOI: [10.1126/science.1162986](https://doi.org/10.1126/science.1162986) (see p. 9).
106. Ellis, R. et al. (2013). “Comparison of the distal gut microbiota from people and animals in Africa”. English. In: *PLoS ONE* 8.1. ISSN: 19326203. DOI: [10.1371/journal.pone.0054783](https://doi.org/10.1371/journal.pone.0054783) (see p. 34).
107. *New Zealand public health surveillance report* (2014) 12.2. ISSN: 1178-8313. URL: <https://surv.esr.cri.nz/surveillance/NZPHSR.php> (see pp. 16, 69, 75, 142).
108. Environmental Systems Resource Institute (2010). *ArcMap 10.0*. ESRI. Redlands, California. (see p. 161).
109. Fair, G. and J. Morris (1949). “Behavior of chlorine as a water disinfectant.” English. In: *Water & sewage works* 96.5, pp. 101–104. ISSN: 00431125 (see p. 21).
110. Feinsinger, P., E. E. Spears, and R. W. Poole (1981). “A simple measure of niche breadth”. In: *Ecology*, pp. 27–32 (see p. 143).
111. Feng, Y. and L. Xiao (2011). “Zoonotic potential and molecular epidemiology of *Giardia* species and giardiasis”. English. In: *Clinical Microbiology Reviews* 24.1, pp. 110–140. ISSN: 08938512. DOI: [10.1128/CMR.00033-10](https://doi.org/10.1128/CMR.00033-10) (see p. 129).
112. Ferguson, A. et al. (2012). “Comparison of fecal indicators with pathogenic bacteria and rotavirus in groundwater”. English. In: *Science of the Total Environment* 431, pp. 314–322. ISSN: 00489697. DOI: [10.1016/j.scitotenv.2012.05.060](https://doi.org/10.1016/j.scitotenv.2012.05.060) (see p. 22).
113. Ferguson, C. et al. (2007). “Development of a process-based model to predict pathogen budgets for the Sydney drinking water catchment”. English. In: *Journal of Water and Health* 5.2, pp. 187–208. ISSN: 14778920. DOI: [10.2166/wh.2007.013](https://doi.org/10.2166/wh.2007.013) (see pp. 46, 69, 128).
114. Ferrarini, M. et al. (2013). “An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome”. English. In: *BMC Genomics* 14.1. ISSN: 14712164. DOI: [10.1186/1471-2164-14-670](https://doi.org/10.1186/1471-2164-14-670) (see p. 9).
115. Field, K. and M. Samadpour (2007). “Fecal source tracking, the indicator paradigm, and managing water quality”. English. In: *Water Research* 41.16, pp. 3517–3538. ISSN: 00431354. DOI: [10.1016/j.watres.2007.06.056](https://doi.org/10.1016/j.watres.2007.06.056) (see pp. 21, 29).
116. Field, K. et al. (2003). “A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking.” English. In: *J Water Health* 1.4, pp. 181–194. ISSN: 14778920 (see p. 29).
117. Fiers, W. et al. (1976). “Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene”. English. In: *Nature* 260.5551, pp. 500–507. ISSN: 00280836. DOI: [10.1038/260500a0](https://doi.org/10.1038/260500a0) (see p. 136).
118. French, N. et al. (2014). “*Campylobacter* ecology and evolution”. In: ed. by S. Sheppard. Caister Academic Press. Chap. Evolution of *Campylobacter* species in New Zealand, pp. 221–240 (see p. 112).
119. Fuller, C. et al. (2009). “The challenges of sequencing by synthesis”. English. In: *Nature Biotechnology* 27.11, pp. 1013–1023. ISSN: 10870156. DOI: [10.1038/nbt.1585](https://doi.org/10.1038/nbt.1585) (see p. 7).
120. Gabriel, K. (1971). “The biplot graphic display of matrices with application to principal component analysis”. English. In: *Biometrika* 58.3, pp. 453–467. ISSN: 00063444. DOI: [10.1093/biomet/58.3.453](https://doi.org/10.1093/biomet/58.3.453) (see p. 80).

121. Gall, J. and M. Pardue (1969). "Formation and detection of RNA-DNA hybrid molecules in cytological preparations." English. In: *Proceedings of the National Academy of Sciences of the United States of America* 63.2, pp. 378–383. ISSN: 00278424 (see p. 25).
122. Gallas-Lindemann, C. et al. (2013). "Detection of *Toxoplasma gondii* oocysts in different water resources by loop mediated isothermal amplification (LAMP)". English. In: *Acta Tropica* 125.2, pp. 231–236. ISSN: 0001706X. DOI: [10.1016/j.actatropica.2012.10.007](https://doi.org/10.1016/j.actatropica.2012.10.007) (see p. 26).
123. Gan, S. and K. Patel (2013). "Enzyme immunoassay and enzyme-linked immunosorbent assay". English. In: *Journal of Investigative Dermatology* 133.9, pp. 1–3. ISSN: 0022202X. DOI: [10.1038/jid.2013.287](https://doi.org/10.1038/jid.2013.287) (see pp. 24, 25).
124. Gao, B., H. Hahn, and E. Hoffmann (2002). "Evaluation of aluminum-silicate polymer composite as a coagulant for water treatment". English. In: *Water Research* 36.14, pp. 3573–3581. ISSN: 00431354. DOI: [10.1016/S0043-1354\(02\)00054-4](https://doi.org/10.1016/S0043-1354(02)00054-4) (see p. 20).
125. Garcia Armisen, T. and P. Servais (2004). "Combining direct viable count (DVC) and fluorescent in situ hybridisation (FISH) to enumerate viable *E. coli* in rivers and wastewaters". English. In: *Water Science and Technology* 50.1, pp. 271–275. ISSN: 02731223 (see p. 25).
126. Garrod, L. (1960). "The relative antibacterial activity of four penicillins." English. In: *British medical journal* 2.5214, pp. 1695–1696. ISSN: 00071447 (see p. 135).
127. Gasparrini, A. (2011). "Distributed lag linear and non-linear models in R: The package dlnm". In: *Journal of Statistical Software* 43.8, pp. 1–20. URL: <http://www.jstatsoft.org/v43/i08/> (see p. 85).
128. — (2014). "Modeling exposure-lag-response associations with distributed lag non-linear models". English. In: *Statistics in Medicine* 33.5, pp. 881–899. ISSN: 02776715. DOI: [10.1002/sim.5963](https://doi.org/10.1002/sim.5963) (see p. 85).
129. Gasparrini, A. and B. Armstrong (2013). "Reducing and meta-analysing estimates from distributed lag non-linear models". English. In: *BMC Medical Research Methodology* 13.1. ISSN: 14712288. DOI: [10.1186/1471-2288-13-1](https://doi.org/10.1186/1471-2288-13-1) (see p. 85).
130. Gasparrini, A., B. Armstrong, and M. Kenward (2010). "Distributed lag non-linear models". English. In: *Statistics in Medicine* 29.21, pp. 2224–2234. ISSN: 02776715. DOI: [10.1002/sim.3940](https://doi.org/10.1002/sim.3940) (see p. 85).
131. — (2012). "Multivariate meta-analysis for non-linear and other multi-parameter associations". English. In: *Statistics in Medicine* 31.29, pp. 3821–3839. ISSN: 02776715. DOI: [10.1002/sim.5471](https://doi.org/10.1002/sim.5471) (see p. 97).
132. Gatnar, E. (2008). "Fusion of multiple statistical classifiers". English. In: Freiburg, pp. 19–27. ISBN: 9783540782391 (see p. 54).
133. Geldreich, E. and B. Kenner (1969). "Concepts of fecal streptococci in stream pollution." English. In: *Journal of the Water Pollution Control Federation* 41.8, Suppl:R336+. ISSN: 00431303 (see p. 29).
134. Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press. ISBN: 978-0-521-86706-1 (see pp. 86, 97, 117, 119).
135. Ghai, R. et al. (2011). "Metagenomics of the water column in the pristine upper course of the Amazon River". In: *Plos One* 6.8. ISSN: 1932-6203 (see pp. 34, 135).
136. Gianoulis, T. et al. (2009). "Quantifying environmental adaptation of metabolic pathways in metagenomics". English. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.5, pp. 1374–1379. ISSN: 00278424. DOI: [10.1073/pnas.0808022106](https://doi.org/10.1073/pnas.0808022106) (see p. 155).
137. Gomez-Alvarez, V., R. Revetta, and J. Domingo (2012). "Metagenomic analyses of drinking water receiving different disinfection treatments". English. In: *Applied and Environmental Microbiology* 78.17, pp. 6095–6102. ISSN: 00992240. DOI: [10.1128/AEM.01018-12](https://doi.org/10.1128/AEM.01018-12) (see pp. 36, 39, 40, 43, 135, 137).

138. Gomi, R. et al. (2014). “Fecal source tracking in water by next-generation sequencing technologies using host-specific *Escherichia coli* genetic markers”. In: *Environmental science & technology* (see pp. 156, 167).
139. Goodman, P., D. Dockery, and L. Clancy (2004). “Cause-specific mortality and the extended effects of particulate pollution and temperature exposure”. English. In: *Environmental Health Perspectives* 112.2, pp. 179–185. ISSN: 00916765 (see p. 85).
140. Goovaerts, P. (1992). “Factorial kriging analysis: A useful tool for exploring the structure of multivariate spatial soil information”. English. In: *Journal of Soil Science* 43.4, pp. 597–619. ISSN: 00224588 (see p. 82).
141. Göransson, G., M. Larson, and D. Bendz (2013). “Variation in turbidity with precipitation and flow in a regulated river system-river Göta Älv, SW Sweden”. English. In: *Hydrology and Earth System Sciences* 17.7, pp. 2529–2542. ISSN: 10275606. DOI: [10.5194/hess-17-2529-2013](https://doi.org/10.5194/hess-17-2529-2013) (see pp. 76, 77, 94).
142. Gorkiewicz, G. et al. (2003). “Species-specific identification of campylobacters by partial 16S rRNA gene sequencing”. English. In: *Journal of Clinical Microbiology* 41.6, pp. 2537–2546. ISSN: 00951137 (see p. 35).
143. Gormley, F. et al. (2008). “Has retail chicken played a role in the decline of human campylobacteriosis?” English. In: *Applied and Environmental Microbiology* 74.2, pp. 383–390. ISSN: 00992240. DOI: [10.1128/AEM.01455-07](https://doi.org/10.1128/AEM.01455-07) (see p. 97).
144. Gray, N. (1994). *Drinking water quality. Problems and solutions*. English. Wiley. ISBN: 0471948179; 0471948187 (see p. 2).
145. Greer, A., S. Drews, and D. Fisman (2009). “Why ”Winter” vomiting disease? Seasonality, hydrology, and norovirus epidemiology in Toronto, Canada”. English. In: *EcoHealth* 6.2, pp. 192–199. ISSN: 16129202. DOI: [10.1007/s10393-009-0247-8](https://doi.org/10.1007/s10393-009-0247-8) (see p. 76).
146. Guarner, F. and J.-R. Malagelada (2003). “Gut flora in health and disease”. English. In: *Lancet* 361.9356, pp. 512–519. ISSN: 01406736. DOI: [10.1016/S0140-6736\(03\)12489-0](https://doi.org/10.1016/S0140-6736(03)12489-0) (see p. 135).
147. Gunther, N. et al. (2011). “*GyrB* versus 16S rRNA sequencing for the identification of *Campylobacter jejuni*, *Campylobacter coli*, and *Campylobacter lari*”. English. In: *Journal of Nucleic Acids Investigation* 2.1, pp. 39–42. ISSN: 20356005. DOI: [10.4081/jnai.2011.e7](https://doi.org/10.4081/jnai.2011.e7) (see p. 35).
148. Guo, F. et al. (2013). “Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment”. English. In: *PLoS ONE* 8.10. ISSN: 19326203. DOI: [10.1371/journal.pone.0076185](https://doi.org/10.1371/journal.pone.0076185) (see p. 35).
149. Gyurek, L., G. Finch, and M. Belosevic (1997). “Modeling chlorine inactivation requirements of *Cryptosporidium parvum* oocysts”. English. In: *Journal of Environmental Engineering* 123.9, pp. 865–875. ISSN: 07339372 (see p. 21).
150. Hadley, G. (1988). *Linear algebra*. Narosa Publishing House. ISBN: 81-85015-81-3 (see p. 85).
151. Hand, D. et al. (2013). “Pyrosequencing the canine faecal microbiota: Breadth and depth of biodiversity”. English. In: *PLoS ONE* 8.1. ISSN: 19326203. DOI: [10.1371/journal.pone.0053115](https://doi.org/10.1371/journal.pone.0053115) (see pp. 39, 135).
152. Harwood, V. et al. (2005). “Validity of the indicator organism paradigm for pathogen reduction in reclaimed water and public health protection”. English. In: *Applied and Environmental Microbiology* 71.6, pp. 3163–3170. ISSN: 00992240. DOI: [10.1128/AEM.71.6.3163-3170.2005](https://doi.org/10.1128/AEM.71.6.3163-3170.2005) (see pp. 22, 156).
153. Harwood, V. et al. (2014). “Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships between pathogens and human health outcomes”. English. In: *FEMS Microbiology Reviews* 38.1, pp. 1–40. ISSN: 01686445. DOI: [10.1111/1574-6976.12031](https://doi.org/10.1111/1574-6976.12031) (see pp. 22, 156).
154. Hastie, T. et al. (2009). *The elements of statistical learning*. Vol. 2. 1. Springer (see pp. 53–55).



155. Hengl, T. (2009). *A practical guide to geostatistical mapping*. Amsterdam. ISBN: 978-90-9024981-0. URL: <http://home.medewerker.uva.nl/t.hengl/> (see pp. 81, 83).
156. Hewitt, J. et al. (2007). "Gastroenteritis outbreak caused by waterborne norovirus at a New Zealand Ski Resort". English. In: *Applied and Environmental Microbiology* 73.24, pp. 7853–7857. ISSN: 00992240. DOI: [10.1128/AEM.00718-07](https://doi.org/10.1128/AEM.00718-07) (see pp. 31, 106, 129).
157. Hewitt, J. et al. (2013). "Evaluation of human adenovirus and human polyomavirus as indicators of human sewage contamination in the aquatic environment". English. In: *Water Research* 47.17, pp. 6750–6761. ISSN: 00431354. DOI: [10.1016/j.watres.2013.09.001](https://doi.org/10.1016/j.watres.2013.09.001) (see p. 22).
158. Hiemstra, P. et al. (2008). "Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network". In: *Computers & Geosciences* (see p. 84).
159. Hill, J., J. Town, and S. Hemmingsen (2006). "Improved template representation in *cpn* 60 polymerase chain reaction (PCR) product libraries generated from complex templates by application of a specific mixture of PCR primers". English. In: *Environmental Microbiology* 8.4, pp. 741–746. ISSN: 14622912. DOI: [10.1111/j.1462-2920.2005.00944.x](https://doi.org/10.1111/j.1462-2920.2005.00944.x) (see p. 35).
160. Hlavsa, M. et al. (2011). "Surveillance for waterborne disease outbreaks and other health events associated with recreational water — United States, 2007-2008". English. In: *Morbidity and Mortality Weekly Report* 60.SS-12, pp. 1–32. ISSN: 01492195 (see p. 16).
161. Ho, A. et al. (2011). "Sequencing by ligation variation with endonuclease V digestion and deoxynosine-containing query oligonucleotides". English. In: *BMC Genomics* 12. ISSN: 14712164. DOI: [10.1186/1471-2164-12-598](https://doi.org/10.1186/1471-2164-12-598) (see p. 7).
162. Hoadley, A. and B. Dutka (1977). "Bacterial indicators — health hazards associated with water". In: *Indicators of recreational water quality*. Cabelli, V. J. American Society for Testing and Materials, p. 132. ISBN: 0463500016 (see p. 22).
163. Hoque, E. M. et al. (2002). "Risk of giardiasis in Aucklanders: A case-control study". English. In: *International Journal of Infectious Diseases* 6.3, pp. 191–197. ISSN: 12019712. DOI: [10.1016/S1201-9712\(02\)90110-4](https://doi.org/10.1016/S1201-9712(02)90110-4) (see p. 32).
164. Hoque, M. et al. (2004). "A descriptive epidemiology of giardiasis in New Zealand and gaps in surveillance data". English. In: *New Zealand Medical Journal* 117.1205. ISSN: 11758716 (see p. 33).
165. Horn, B. and R. Lake (2013). "Incubation period for campylobacteriosis and its importance in the estimation of incidence related to travel". English. In: *Eurosurveillance* 18.40. ISSN: 1025496X (see p. 96).
166. Hosmer Jr, D. W. and S. Lemeshow (2000). *Applied logistic regression*. John Wiley & Sons. ISBN: 0-471-35632-8 (see p. 55).
167. Hothorn, T. et al. (2006a). "Survival ensembles". English. In: *Biostatistics* 7.3, pp. 355–373. ISSN: 14654644. DOI: [10.1093/biostatistics/kxj011](https://doi.org/10.1093/biostatistics/kxj011) (see pp. 53, 55).
168. Hothorn, T., K. Hornik, and A. Zeileis (2006b). "Unbiased recursive partitioning: A conditional inference framework". English. In: *Journal of Computational and Graphical Statistics* 15.3, pp. 651–674. ISSN: 10618600. DOI: [10.1198/106186006X133933](https://doi.org/10.1198/106186006X133933) (see p. 53).
169. Hrudey, S., E. Hrudey, and S. Pollard (2006). "Risk management for assuring safe drinking water". English. In: *Environment International* 32.8, pp. 948–957. ISSN: 01604120. DOI: [10.1016/j.envint.2006.06.004](https://doi.org/10.1016/j.envint.2006.06.004) (see p. 4).
170. Hübner, I. et al. (1992). "Rapid determination of members of the family *Enterobacteriaceae* in drinking water by an immunological assay using a monoclonal antibody against enterobacterial common antigen". English. In: *Applied and Environmental Microbiology* 58.9, pp. 3187–3191. ISSN: 00992240 (see p. 25).

171. Hughes, A. and J. Quinn (2014). “Before and after integrated catchment management in a headwater catchment: Changes in water quality”. English. In: *Environmental Management* 54.6, pp. 1288–1305. ISSN: 0364152X. DOI: [10.1007/s00267-014-0369-9](https://doi.org/10.1007/s00267-014-0369-9) (see p. 15).
172. Huson, D. H. et al. (2011). “Integrative analysis of environmental sequences using MEGAN4”. In: *Genome Research* 21.9, pp. 1552–1560 (see p. 161).
173. Huson, D. H. and C. Xie (2013). “A poor man’s BLASTX-high-throughput metagenomic protein database search using PAUDA”. In: *Bioinformatics* (see pp. 143, 161).
174. Huson, D. and D. Bryant (2006). “Application of phylogenetic networks in evolutionary studies”. English. In: *Molecular Biology and Evolution* 23.2, pp. 254–267. ISSN: 07374038. DOI: [10.1093/molbev/msj030](https://doi.org/10.1093/molbev/msj030) (see p. 161).
175. Huson, D. et al. (2007). “MEGAN analysis of metagenomic data”. English. In: *Genome Research* 17.3, pp. 377–386. ISSN: 10889051. DOI: [10.1101/gr.5969107](https://doi.org/10.1101/gr.5969107) (see p. 143).
176. Huttenhower, C. et al. (2012). “Structure, function and diversity of the healthy human microbiome”. English. In: *Nature* 486.7402, pp. 207–214. ISSN: 00280836. DOI: [10.1038/nature11234](https://doi.org/10.1038/nature11234) (see p. 135).
177. Ikram, R. et al. (1994). “A case control study to determine risk factors for *Campylobacter* infection in Christchurch in the summer of 1992–3.” English. In: *New Zealand Medical Journal* 107.988, pp. 430–432. ISSN: 00288446 (see p. 32).
178. Illumina Inc. (2011). *TruSeq DNA sample preparation v2 guide*. URL: [http://support.illumina.com/sequencing/sequencing\\_kits/truseq\\_rna\\_sample\\_prep\\_kit\\_v2/documentation.html](http://support.illumina.com/sequencing/sequencing_kits/truseq_rna_sample_prep_kit_v2/documentation.html) (see p. 139).
179. — (2014). *Understanding Illumina quality scores*. URL: [http://res.illumina.com/documents/products/technotes/technote\\_understanding\\_quality\\_scores.pdf](http://res.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf) (see p. 141).
180. Institute of Environmental Science and Research Ltd (2014). *Notifiable and other diseases in New Zealand: Annual report 2013*. Tech. rep. Porirua, New Zealand. URL: [www.surv.esr.cri.nz](http://www.surv.esr.cri.nz) (see p. 71).
181. Intergovernmental Panel on Climate Change (2013). *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Tech. rep. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp. URL: <https://www.ipcc.ch/report/ar5/wg1/> (see pp. 13, 14).
182. International Agency for Research on Cancer (2004). “Some drinking-water disinfectants and contaminants, including arsenic. Monographs on chloramine, chloral and chloral hydrate, dichloroacetic acid, trichloroacetic acid and 3-chloro-4-(dichloromethyl)-5-hydroxy-2(5H)-furanone.” English. In: *IARC monographs on the evaluation of carcinogenic risks to humans / World Health Organization, International Agency for Research on Cancer* 84, pp. 269–477. ISSN: 10171606. URL: <http://monographs.iarc.fr/ENG/Monographs/vol84/> (see p. 2).
183. JabRef Development Team (2014). *JabRef*. URL: <http://jabref.sf.net> (see p. 161).
184. Jagai, J. et al. (2012). “Seasonal patterns of gastrointestinal illness and streamflow along the Ohio River”. English. In: *International Journal of Environmental Research and Public Health* 9.5, pp. 1771–1790. ISSN: 16604601. DOI: [10.3390/ijerph9051771](https://doi.org/10.3390/ijerph9051771) (see pp. 76, 94).
185. Jardé, E., G. Gruau, and L. Mansuy-Huault (2007). “Detection of manure-derived organic compounds in rivers draining agricultural areas of intensive manure spreading”. English. In: *Applied Geochemistry* 22.8 SPEC. ISS. Pp. 1814–1824. ISSN: 08832927. DOI: [10.1016/j.apgeochem.2007.03.037](https://doi.org/10.1016/j.apgeochem.2007.03.037) (see p. 30).
186. Jaros, P. et al. (2013). “A prospective case-control and molecular epidemiological study of human cases of shiga toxin-producing *Escherichia coli* in New Zealand”. English. In: *BMC Infectious Diseases* 13.1. ISSN: 14712334. DOI: [10.1186/1471-2334-13-450](https://doi.org/10.1186/1471-2334-13-450) (see p. 32).
187. Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library (see p. 79).



188. Jolliffe, I. and B. Morgan (1992). "Principal component analysis and exploratory factor analysis." English. In: *Statistical methods in medical research* 1.1, pp. 69–95. ISSN: 09622802 (see p. 79).
189. Jose Figueras, M. and J. Borrego (2010). "New perspectives in monitoring drinking water microbial quality". English. In: *International Journal of Environmental Research and Public Health* 7.12, pp. 4179–4202. ISSN: 16604601. DOI: [10.3390/ijerph7124179](https://doi.org/10.3390/ijerph7124179) (see pp. 69, 95).
190. Karanis, P. et al. (2007). "Development and preliminary evaluation of a loop-mediated isothermal amplification procedure for sensitive detection of *Cryptosporidium* oocysts in fecal and water samples". English. In: *Applied and Environmental Microbiology* 73.17, pp. 5660–5662. ISSN: 00992240. DOI: [10.1128/AEM.01152-07](https://doi.org/10.1128/AEM.01152-07) (see p. 26).
191. Karnes, J. and R. Usatine (2014). "Management of external genital warts". English. In: *American Family Physician* 90.5, pp. 312–318. ISSN: 0002838X (see p. 135).
192. Kerkhof, L. and R. Goodman (2009). "Ocean microbial metagenomics". English. In: *Deep-Sea Research Part II: Topical Studies in Oceanography* 56.19-20, pp. 1824–1829. ISSN: 09670645. DOI: [10.1016/j.dsr2.2009.05.005](https://doi.org/10.1016/j.dsr2.2009.05.005) (see p. 135).
193. Khan, R. et al. (2007). "Environmental health indicators in New Zealand: Drinking water — A case study". English. In: *EcoHealth* 4.1, pp. 63–71. ISSN: 16129202. DOI: [10.1007/s10393-007-0089-1](https://doi.org/10.1007/s10393-007-0089-1) (see pp. 33, 94).
194. Kildare, B. et al. (2007). "16S rRNA-based assays for quantitative detection of universal, human-, cow-, and dog-specific fecal *Bacteroidales*: A Bayesian approach". English. In: *Water Research* 41.16, pp. 3701–3715. ISSN: 00431354. DOI: [10.1016/j.watres.2007.06.037](https://doi.org/10.1016/j.watres.2007.06.037) (see pp. 29, 156, 167).
195. King, N., R. Lake, and D. Campbell (2011). "Source attribution of nontyphoid salmonellosis in New Zealand using outbreak surveillance data". English. In: *Journal of Food Protection* 74.3, pp. 438–445. ISSN: 0362028X. DOI: [10.4315/0362-028X.JFP-10-323](https://doi.org/10.4315/0362-028X.JFP-10-323) (see p. 32).
196. Kolbert, C. and D. Persing (1999). "Ribosomal DNA sequencing as a tool for identification of bacterial pathogens". English. In: *Current Opinion in Microbiology* 2.3, pp. 299–305. ISSN: 13695274. DOI: [10.1016/S1369-5274\(99\)80052-6](https://doi.org/10.1016/S1369-5274(99)80052-6) (see p. 35).
197. Korich, D. et al. (1990). "Effects of ozone, chlorine dioxide, chlorine, and monochloramine on *Cryptosporidium parvum* oocyst viability". English. In: *Applied and Environmental Microbiology* 56.5, pp. 1423–1428. ISSN: 00992240 (see p. 21).
198. Kovalchik, S. (2014). *RISmed: Download content from NCBI databases*. URL: <http://CRAN.R-project.org/package=RISmed> (see p. 36).
199. Krentz, C., N. Prystajacky, and J. Isaac-Renton (2013). "Identification of fecal contamination sources in water using host-associated markers". English. In: *Canadian Journal of Microbiology* 59.3, pp. 210–220. ISSN: 00084166. DOI: [10.1139/cjm-2012-0618](https://doi.org/10.1139/cjm-2012-0618) (see p. 30).
200. Kuo, J.-T. et al. (2010). "A rapid method for the detection of representative coliforms in water samples: Polymerase chain reaction-enzyme-linked immunosorbent assay (PCR-ELISA)". English. In: *Journal of Industrial Microbiology and Biotechnology* 37.3, pp. 237–244. ISSN: 13675435. DOI: [10.1007/s10295-009-0666-0](https://doi.org/10.1007/s10295-009-0666-0) (see p. 26).
201. Kwon, S. et al. (2011). "Pyrosequencing demonstrated complex microbial communities in a membrane filtration system for a drinking water treatment plant". English. In: *Microbes and Environments* 26.2, pp. 149–155. ISSN: 13426311. DOI: [10.1264/jsme2.ME10205](https://doi.org/10.1264/jsme2.ME10205) (see pp. 36, 39, 137).
202. Lake, I. et al. (2005). "Effects of weather and river flow on cryptosporidiosis." English. In: *Journal of water and health*. 3.4, pp. 469–474. ISSN: 14778920 (see p. 76).
203. Lake, R. J. et al. (2010). "The disease pyramid for acute gastrointestinal illness in New Zealand". In: *Epidemiology and infection* 138.10, pp. 1468–1471 (see pp. 16, 96).

204. Lake, R., B. Adlam, and S. Perera (2009). *Acute gastrointestinal illness (AGI) Study: Final study report*. [http://www.foodsafety.govt.nz/elibrary/industry/acute-illness-study-gastrointestinal-report/Final\\_Report.pdf](http://www.foodsafety.govt.nz/elibrary/industry/acute-illness-study-gastrointestinal-report/Final_Report.pdf) (see pp. 16, 96).
205. Lalle, M. et al. (2005). “Genetic heterogeneity at the  $\beta$ -giardin locus among human and animal isolates of *Giardia duodenalis* and identification of potentially zoonotic subgenotypes”. English. In: *International Journal for Parasitology* 35.2, pp. 207–213. ISSN: 00207519. DOI: [10.1016/j.ijpara.2004.10.022](https://doi.org/10.1016/j.ijpara.2004.10.022) (see p. 129).
206. Lander, E. et al. (2001). “Initial sequencing and analysis of the human genome”. English. In: *Nature* 409.6822, pp. 860–921. ISSN: 00280836. DOI: [10.1038/35057062](https://doi.org/10.1038/35057062) (see p. 136).
207. Langer-Safer, P., M. Levine, and D. Ward (1982). “Immunological methods for mapping genes on *Drosophila polytene* chromosomes”. English. In: *Proceedings of the National Academy of Sciences of the United States of America* 79.14 I, pp. 4381–4385. ISSN: 00278424 (see p. 25).
208. Langeveld, S. et al. (1978). “Nucleotide sequence of the origin of replication in bacteriophage phiX174 RF DNA”. English. In: *Nature* 271.5644, pp. 417–420. ISSN: 00280836. DOI: [10.1038/271417a0](https://doi.org/10.1038/271417a0) (see p. 136).
209. Lawler, D. et al. (2006). “Turbidity dynamics during spring storm events in an urban headwater river system: The Upper Tame, West Midlands, UK”. English. In: *Science of the Total Environment* 360.1-3, pp. 109–126. ISSN: 00489697. DOI: [10.1016/j.scitotenv.2005.08.032](https://doi.org/10.1016/j.scitotenv.2005.08.032) (see p. 94).
210. Lax, P. D. (2007). *Linear algebra and its applications*. Wiley New York. ISBN: 978-0-471-375156-4 (cloth) (see p. 85).
211. Learmonth, J. et al. (2003). “Identification and genetic characterisation of *Giardia* and *Cryptosporidium* strains in humans and dairy cattle in the Waikato Region of New Zealand”. English. In: *Water Science and Technology* 47.3, pp. 21–26. ISSN: 02731223 (see p. 129).
212. Leboffe, M. J. and B. E. Pierce (2011). *A photographic atlas for the microbiology laboratory*. 4th Edition. Morton Publishing Company, Englewood, Colorado, USA (see p. 23).
213. Lee, C. et al. (2012). “Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing”. English. In: *PLoS ONE* 7.9. ISSN: 19326203. DOI: [10.1371/journal.pone.0044224](https://doi.org/10.1371/journal.pone.0044224) (see p. 36).
214. Leeming, R. et al. (1996). “Using faecal sterols from humans and animals to distinguish faecal pollution in receiving waters”. English. In: *Water Research* 30.12, pp. 2893–2900. ISSN: 00431354. DOI: [10.1016/S0043-1354\(96\)00011-5](https://doi.org/10.1016/S0043-1354(96)00011-5) (see p. 29).
215. Lehtola, M. et al. (2006). “Fluorescence in situ hybridization using peptide nucleic acid probes for rapid detection of *Mycobacterium avium* subsp. *avium* and *Mycobacterium avium* subsp. *paratuberculosis* in potable-water biofilms”. English. In: *Applied and Environmental Microbiology* 72.1, pp. 848–853. ISSN: 00992240. DOI: [10.1128/AEM.72.1.848-853.2006](https://doi.org/10.1128/AEM.72.1.848-853.2006) (see p. 25).
216. Li, J. and A. D. Heap (2008). *A review of spatial interpolation methods for environmental scientists*, p. 137 (see p. 82).
217. Liang, Y. et al. (2011). “Functional gene diversity of soil microbial communities from five oil-contaminated fields in China”. English. In: *ISME Journal* 5.3, pp. 403–413. ISSN: 17517362. DOI: [10.1038/ismej.2010.142](https://doi.org/10.1038/ismej.2010.142) (see p. 155).
218. Ling, L. et al. (2015). “A new antibiotic kills pathogens without detectable resistance”. English. In: *Nature* 517.7535, pp. 455–459. ISSN: 00280836. DOI: [10.1038/nature14098](https://doi.org/10.1038/nature14098) (see p. 34).
219. Linton, D., R. Owen, and J. Stanley (1996). “Rapid identification by PCR of the genus *Campylobacter* and of five *Campylobacter* species enteropathogenic for man and animals”. English. In: *Research in Microbiology* 147.9, pp. 707–718. ISSN: 09232508. DOI: [10.1016/S0923-2508\(97\)85118-2](https://doi.org/10.1016/S0923-2508(97)85118-2) (see pp. 112, 113).

220. Liu, L. et al. (2012). "Comparison of next-generation sequencing systems". English. In: *Journal of Biomedicine and Biotechnology* 2012. ISSN: 11107243. DOI: [10.1155/2012/251364](#) (see pp. 6, 8, 9).
221. Locas, A. et al. (2008). "Groundwater microbiological quality in Canadian drinking water municipal wells". English. In: *Canadian Journal of Microbiology* 54.6, pp. 472–478. ISSN: 00084166. DOI: [10.1139/W08-028](#) (see p. 26).
222. Logan, J. et al. (2000). "*Campylobacter lanienae* sp. nov., a new species isolated from workers in an abattoir". English. In: *International Journal of Systematic and Evolutionary Microbiology* 50.2, pp. 865–872. ISSN: 14665026 (see p. 151).
223. Lopman, B. et al. (2009). "Host, weather and virological factors drive *Norovirus* epidemiology: Time-series analysis of laboratory surveillance data in England and Wales". English. In: *PLoS ONE* 4.8. ISSN: 19326203. DOI: [10.1371/journal.pone.0006671](#) (see p. 75).
224. Lu, J., J. Santo Domingo, and O. Shanks (2007). "Identification of chicken-specific fecal microbial sequences using a metagenomic approach". English. In: *Water Research* 41.16, pp. 3561–3574. ISSN: 00431354. DOI: [10.1016/j.watres.2007.05.033](#) (see pp. 29, 30).
225. Lund, V. (1996). "Evaluation of *E. coli* as an indicator for the presence of *Campylobacter jejuni* and *Yersinia enterocolitica* in chlorinated and untreated oligotrophic lake water". English. In: *Water Research* 30.6, pp. 1528–1534. ISSN: 00431354. DOI: [10.1016/0043-1354\(96\)00034-6](#) (see p. 156).
226. Lütkepohl, H. (1981). "A model for non-negative and non-positive distributed lag functions". English. In: *Journal of Econometrics* 16.2, pp. 211–219. ISSN: 03044076 (see p. 85).
227. Lyra, A. et al. (2009). "Diarrhoea-predominant irritable bowel syndrome distinguishable by 16S rRNA gene phylotype quantification". English. In: *World Journal of Gastroenterology* 15.47, pp. 5936–5945. ISSN: 10079327. DOI: [10.3748/wjg.15.5936](#) (see p. 95).
228. MacKelprang, R. et al. (2011). "Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw". English. In: *Nature* 480.7377, pp. 368–371. ISSN: 00280836. DOI: [10.1038/nature10576](#) (see p. 135).
229. Mackenzie, W. R. et al. (1994). "A massive outbreak in Milwaukee of *Cryptosporidium* infection transmitted through the public water supply". In: *New England Journal of Medicine* 331.3, pp. 161–167. ISSN: 0028-4793 (see pp. 14, 75).
230. Magoč, T. and S. Salzberg (2011). "FLASH: Fast length adjustment of short reads to improve genome assemblies". English. In: *Bioinformatics* 27.21, pp. 2957–2963. ISSN: 13674803. DOI: [10.1093/bioinformatics/btr507](#) (see pp. 141, 161).
231. Maiden, M. et al. (1998). "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms". English. In: *Proceedings of the National Academy of Sciences of the United States of America* 95.6, pp. 3140–3145. ISSN: 00278424. DOI: [10.1073/pnas.95.6.3140](#) (see p. 110).
232. Manz, W. et al. (1993). "In situ identification of bacteria in drinking water and adjoining biofilms by hybridization with 16S and 23S rRNA-directed fluorescent oligonucleotide probes". English. In: *Applied and Environmental Microbiology* 59.7, pp. 2293–2298. ISSN: 00992240 (see p. 25).
233. Mardis, E. (2008). "Next-generation DNA sequencing methods". English. In: *Annual Review of Genomics and Human Genetics* 9, pp. 387–402. ISSN: 15278204. DOI: [10.1146/annurev.genom.9.081307.164359](#) (see pp. 7, 8).
234. — (2011). "A decade's perspective on DNA sequencing technology". English. In: *Nature* 470.7333, pp. 198–203. ISSN: 00280836. DOI: [10.1038/nature09796](#) (see pp. 10, 36).
235. Marshall, H. et al. (2011). "*Mycobacterium lentiflavum* in drinking water supplies, Australia". English. In: *Emerging Infectious Diseases* 17.3, pp. 395–402. ISSN: 10806040. DOI: [10.3201/eid1703.090948](#) (see pp. 149, 151).

236. Martínez, J. (2008). “Antibiotics and antibiotic resistance genes in natural environments”. English. In: *Science* 321.5887, pp. 365–367. ISSN: 00368075. DOI: [10.1126/science.1159483](https://doi.org/10.1126/science.1159483) (see pp. 43, 157).
237. Masip, L., K. Veeravalli, and G. Georgiou (2006). “The many faces of glutathione in bacteria”. English. In: *Antioxidants and Redox Signaling* 8.5-6, pp. 753–762. ISSN: 15230864. DOI: [10.1089/ars.2006.8.753](https://doi.org/10.1089/ars.2006.8.753) (see p. 40).
238. Matheson, A. (1992). “Structure, function and evolution of the archaeal ribosome.” English. In: *Biochemical Society Symposia* 58, pp. 89–98. ISSN: 00678694 (see p. 34).
239. Matilainen, A., M. Vepsäläinen, and M. Sillanpää (2010). “Natural organic matter removal by coagulation during drinking water treatment: A review”. English. In: *Advances in Colloid and Interface Science* 159.2, pp. 189–197. ISSN: 00018686. DOI: [10.1016/j.cis.2010.06.007](https://doi.org/10.1016/j.cis.2010.06.007) (see p. 20).
240. Maxam, A. and W. Gilbert (1977). “A new method for sequencing DNA”. English. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.2, pp. 560–564. ISSN: 00278424 (see p. 5).
241. McBride, G., A. Tait, and D. Slaney (2014). “Projected changes in reported campylobacteriosis and cryptosporidiosis rates as a function of climate change: A New Zealand study”. English. In: *Stochastic Environmental Research and Risk Assessment* 28.8, pp. 2133–2147. ISSN: 14363240. DOI: [10.1007/s00477-014-0920-5](https://doi.org/10.1007/s00477-014-0920-5) (see p. 14).
242. McBride, G., J. McWhirter, and M. Dalgety (2003). “Uncertainty in most probable number calculations for microbiological assays”. English. In: *Journal of AOAC International* 86.5, pp. 1084–1088. ISSN: 10603271 (see p. 23).
243. McDonald, D. et al. (2012). “The biological observation matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome”. In: *GigaScience* 1.1, p. 7 (see p. 142).
244. McMurdie, P. and S. Holmes (2013). “Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data”. English. In: *PLoS ONE* 8.4, e61217. ISSN: 19326203. DOI: [10.1371/journal.pone.0061217](https://doi.org/10.1371/journal.pone.0061217) (see p. 142).
245. Meays, C. et al. (2004). “Source tracking fecal bacteria in water: A critical review of current methods”. In: *Journal of Environmental Management* 73.1, pp. 71–79. ISSN: 0301-4797 (see p. 29).
246. Medema, G. J. et al. (2002). *Catchment characterisation and source water quality* (see p. 14).
247. Metzker, M. (2009). “Sequencing in real time”. English. In: *Nature Biotechnology* 27.2, pp. 150–151. ISSN: 10870156. DOI: [10.1038/nbt0209-150](https://doi.org/10.1038/nbt0209-150) (see p. 9).
248. — (2010). “Sequencing technologies the next generation”. English. In: *Nature Reviews Genetics* 11.1, pp. 31–46. ISSN: 14710056. DOI: [10.1038/nrg2626](https://doi.org/10.1038/nrg2626) (see p. 7).
249. Microsoft, E. (2012). *Microsoft*. URL: <http://www.microsoft.com/en-us/default.aspx> (see p. 161).
250. Milly, P., K. Dunne, and A. Vecchia (2005). “Global pattern of trends in streamflow and water availability in a changing climate”. English. In: *Nature* 438.7066, pp. 347–350. ISSN: 00280836. DOI: [10.1038/nature04312](https://doi.org/10.1038/nature04312) (see p. 13).
251. Mitchell, P., P. Graham, and M. Brieseman (1993). “Giardiasis in Canterbury: The first nine months reported cases.” English. In: *New Zealand Medical Journal* 106.962, pp. 350–352. ISSN: 00288446 (see p. 32).
252. Mitra, S. et al. (2010). “Comparison of multiple metagenomes using phylogenetic networks based on ecological indices”. English. In: *ISME Journal* 4.10, pp. 1236–1242. ISSN: 17517362. DOI: [10.1038/ismej.2010.51](https://doi.org/10.1038/ismej.2010.51) (see p. 156).
253. Mittelbach, F. et al. (2004). *The LATEX companion*. Addison-Wesley Professional (see p. 161).
254. Mizrahi-Man, O., E. Davenport, and Y. Gilad (2013). “Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: Evaluation of effective study designs”. English. In: *PLoS ONE* 8.1. ISSN: 19326203. DOI: [10.1371/journal.pone.0053608](https://doi.org/10.1371/journal.pone.0053608) (see p. 35).

255. Molles, M. C. (2013). *Ecology: Concepts and applications*. 6th Edition. WCB/McGraw-Hill Dubuque, IA. ISBN: 978-0-07-353243-3 (see p. 142).
256. Montgomery, D. C., E. A. Peck, and G. G. Vining (2012). *Introduction to linear regression analysis*. Hoboken, N.J. : Wiley. ISBN: 9780470542811 (see pp. 118, 119).
257. Moore, D. et al. (2010). *Cost benefit analysis of raising the quality of New Zealand networked drinking water*. Tech. rep. LEGG (see p. 16).
258. Morales-Sánchez, A. and E. Fuentes-Panana (2014). “Human viruses and cancer”. English. In: *Viruses* 6.10, pp. 4047–4079. ISSN: 19994915. DOI: [10.3390/v6104047](https://doi.org/10.3390/v6104047) (see p. 135).
259. Morens, D. et al. (1979). “A waterborne outbreak of gastroenteritis with secondary person-to-person spread. Association with a viral agent”. English. In: *Lancet* 1.8123, pp. 964–966. ISSN: 01406736. DOI: [10.1016/S0140-6736\(79\)91734-3](https://doi.org/10.1016/S0140-6736(79)91734-3) (see p. 106).
260. Moriarty, E. and B. Gilpin (2009). *Faecal source tracking in the Avon River, Christchurch March - May 2009*. Tech. rep. Institute of Environmental Science and Research Limited. URL: <http://ecan.govt.nz/publications/Reports/technical-report-faecal-source-tracking-avon-river-march-may-2009-screen-000809.pdf> (see p. 30).
261. Muellner, P. et al. (2013). “Molecular-based surveillance of campylobacteriosis in New Zealand - From source attribution to genomic epidemiology”. English. In: *Eurosurveillance* 18.3. ISSN: 1025496X (see pp. 41, 71, 97, 135).
262. Mullis, K. and F. Faloon (1987). “Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction”. English. In: *Methods in Enzymology* 155.C, pp. 335–350. ISSN: 00766879. DOI: [10.1016/0076-6879\(87\)55023-6](https://doi.org/10.1016/0076-6879(87)55023-6) (see p. 25).
263. Müllner, P. et al. (2009). “Source attribution of food-borne zoonoses in New Zealand: A modified hald model”. English. In: *Risk Analysis* 29.7, pp. 970–984. ISSN: 02724332. DOI: [10.1111/j.1539-6924.2009.01224.x](https://doi.org/10.1111/j.1539-6924.2009.01224.x) (see p. 15).
264. Myers, D. N. et al. (2007). *Fecal indicator bacteria: United States geological survey techniques of water-resources investigations*. Version 2.0. Book 9. Chap. A7. URL: <http://water.usgs.gov/owq/FieldManual/Chapter7/index.html> (see p. 23).
265. Nalbantoglu, O. et al. (2011). “RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles”. English. In: *BMC Bioinformatics* 12. ISSN: 14712105. DOI: [10.1186/1471-2105-12-41](https://doi.org/10.1186/1471-2105-12-41) (see p. 156).
266. Neish, A. (2009). “Microbes in gastrointestinal health and disease”. English. In: *Gastroenterology* 136.1, pp. 65–80. ISSN: 00165085. DOI: [10.1053/j.gastro.2008.10.080](https://doi.org/10.1053/j.gastro.2008.10.080) (see p. 135).
267. Nelson, W. (2010). “Campylobacteriosis in New Zealand”. English. In: *Epidemiology and Infection* 138.12, pp. 1762–1763. ISSN: 09502688. DOI: [10.1017/S0950268810001172](https://doi.org/10.1017/S0950268810001172) (see p. 135).
268. New Zealand Ministry for the Environment (2003). *Microbiological water quality guidelines for marine and freshwater recreational areas*. Tech. rep. Wellington. URL: [www.mfe.govt.nz](http://www.mfe.govt.nz) (see pp. 1, 31, 115, 120, 128).
269. New Zealand Ministry of Business, Innovation and Employment (2009). *Tourist activity: Nature-based tourism*. URL: [www.tourismresearch.govt.nz](http://www.tourismresearch.govt.nz) (see p. 105).
270. — (2013). *New Zealand sectors report 2013: Tourism*. URL: <http://www.mbie.govt.nz/what-we-do/business-growth-agenda/sectors-reports-series> (see p. 105).
271. New Zealand Ministry of Health (2007). *Health (Drinking Water) Amendment Act 2007*. <http://www.drinkingwater.esr.cri.nz/general/drinkingwateract.asp> (see p. 17).
272. — (2008). *Drinking-Water Standards for New Zealand*. <http://www.drinkingwater.esr.cri.nz/general/standards.asp> (see pp. 1, 17, 21, 22, 28, 29, 31, 69, 115, 128).



273. — (2011). *Register of Community Drinking-Water Supplies*. <http://www.health.govt.nz/publication/register-community-drinking-water-supplies-new-zealand-2011-edition> (see p. 17).
274. — (2014). *Annual report on drinking-water quality 2012–2013*. URL: <http://www.health.govt.nz/publication/annual-report-drinking-water-quality-2012-2013> (see p. 13).
275. Nichols, D. et al. (2010). “Use of ichip for high-throughput in situ cultivation of ”uncultivable microbial species”. English. In: *Applied and Environmental Microbiology* 76.8, pp. 2445–2450. ISSN: 00992240. DOI: [10.1128/AEM.01754-09](https://doi.org/10.1128/AEM.01754-09) (see p. 34).
276. Nichols, G. et al. (2009). “Rainfall and outbreaks of drinking water related disease and in England and Wales”. English. In: *Journal of Water and Health* 7.1, pp. 1–8. ISSN: 14778920. DOI: [10.2166/wh.2009.143](https://doi.org/10.2166/wh.2009.143) (see p. 75).
277. Niessen, L. et al. (2013). “The application of loop-mediated isothermal amplification (LAMP) in food testing for bacterial pathogens and fungal contaminants”. English. In: *Food Microbiology* 36.2, pp. 191–206. ISSN: 07400020. DOI: [10.1016/j.fm.2013.04.017](https://doi.org/10.1016/j.fm.2013.04.017) (see p. 26).
278. Nikaido, H. (2009). “Multidrug resistance in bacteria”. English. In: *Annual Review of Biochemistry* 78, pp. 119–146. ISSN: 00664154. DOI: [10.1146/annurev.biochem.78.082907.145923](https://doi.org/10.1146/annurev.biochem.78.082907.145923) (see pp. 43, 157).
279. Notomi, T. et al. (2000). “Loop-mediated isothermal amplification of DNA.” English. In: *Nucleic acids research* 28.12, E63. ISSN: 13624962 (see p. 26).
280. Nygård, K. et al. (2004). “Waterborne outbreak of gastroenteritis in a religious summer camp in Norway, 2002”. English. In: *Epidemiology and Infection* 132.2, pp. 223–229. ISSN: 09502688. DOI: [10.1017/S0950268803001894](https://doi.org/10.1017/S0950268803001894) (see pp. 106, 129).
281. Oh, S. et al. (2011). “Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem”. English. In: *Applied and Environmental Microbiology* 77.17, pp. 6000–6011. ISSN: 00992240. DOI: [10.1128/AEM.00107-11](https://doi.org/10.1128/AEM.00107-11) (see pp. 36, 39, 137).
282. Oikonomou, G. et al. (2013). “Fecal microbial diversity in pre-weaned dairy calves as described by pyrosequencing of metagenomic 16S rDNA. Associations of *Faecalibacterium* species with health and growth”. English. In: *PLoS ONE* 8.4. ISSN: 19326203. DOI: [10.1371/journal.pone.0063157](https://doi.org/10.1371/journal.pone.0063157) (see pp. 39, 135).
283. OMOE (2006). *Procedure for disinfection of drinking water in Ontario*. Tech. rep. Ontario Ministry of the Environment (see p. 21).
284. Omre, H. (1987). “Bayesian kriging—Merging observations and qualified guesses in kriging”. English. In: *Mathematical Geology* 19.1, pp. 25–39. ISSN: 08828121. DOI: [10.1007/BF01275432](https://doi.org/10.1007/BF01275432) (see p. 83).
285. Omre, H. and K. Halvorsen (1989). “The Bayesian bridge between simple and universal kriging”. English. In: *Mathematical Geology* 21.7, pp. 767–786. ISSN: 08828121. DOI: [10.1007/BF00893321](https://doi.org/10.1007/BF00893321) (see p. 83).
286. Onozuka, D. and M. Hashizume (2011). “Weather variability and paediatric infectious gastroenteritis”. English. In: *Epidemiology and Infection* 139.9, pp. 1369–1378. ISSN: 09502688. DOI: [10.1017/S0950268810002451](https://doi.org/10.1017/S0950268810002451) (see p. 75).
287. Onozuka, D., M. Hashizume, and A. Hagihara (2010). “Effects of weather variability on infectious gastroenteritis”. English. In: *Epidemiology and Infection* 138.2, pp. 236–243. ISSN: 09502688. DOI: [10.1017/S0950268809990574](https://doi.org/10.1017/S0950268809990574) (see p. 75).
288. Overbeek, R. et al. (2005). “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes”. English. In: *Nucleic Acids Research* 33.17, pp. 5691–5702. ISSN: 03051048. DOI: [10.1093/nar/gki866](https://doi.org/10.1093/nar/gki866) (see pp. 39, 143).

289. Pace, N. (1997). "A molecular view of microbial diversity and the biosphere". English. In: *Science* 276.5313, pp. 734–740. ISSN: 00368075. DOI: [10.1126/science.276.5313.734](https://doi.org/10.1126/science.276.5313.734) (see p. 136).
290. Pardhan-Ali, A. et al. (2012). "A spatial and temporal analysis of notifiable gastrointestinal illness in the Northwest Territories, Canada, 1991-2008". English. In: *International Journal of Health Geographics* 11. ISSN: 1476072X. DOI: [10.1186/1476-072X-11-17](https://doi.org/10.1186/1476-072X-11-17) (see p. 95).
291. Parliamentary Commissioner for the Environment (2004). *Growing for good: Intensive farming, sustainability and New Zealand's environment*. Wellington: Parliamentary Commissioner for the Environment. (see p. 2).
292. Patin, N. et al. (2013). "Effects of OTU clustering and PCR artifacts on microbial diversity estimates". English. In: *Microbial Ecology* 65.3, pp. 709–719. ISSN: 00953628. DOI: [10.1007/s00248-012-0145-4](https://doi.org/10.1007/s00248-012-0145-4) (see p. 35).
293. Payment, P., E. Franco, and J. Siemiatycki (1993). "Absence of relationship between health effects due to tap water consumption and drinking water quality parameters". English. In: *Water Science and Technology* 27.3-4, pp. 137–143. ISSN: 02731223 (see p. 22).
294. Payment, P. et al. (1997). "A prospective epidemiological study of gastrointestinal health effects due to the consumption of drinking water". English. In: *International Journal of Environmental Health Research* 7.1, pp. 5–31. ISSN: 09603123. DOI: [10.1080/09603129773977](https://doi.org/10.1080/09603129773977) (see p. 45).
295. Peeters, K. and A. Willems (2011). "The *gyrB* gene is a useful phylogenetic marker for exploring the diversity of *Flavobacterium* strains isolated from terrestrial and aquatic habitats in Antarctica". English. In: *FEMS Microbiology Letters* 321.2, pp. 130–140. ISSN: 03781097. DOI: [10.1111/j.1574-6968.2011.02326.x](https://doi.org/10.1111/j.1574-6968.2011.02326.x) (see p. 35).
296. Pereira, F. et al. (2010). "Identification of species by multiplex analysis of variable-length sequences". English. In: *Nucleic Acids Research* 38.22, e203. ISSN: 03051048. DOI: [10.1093/nar/gkq865](https://doi.org/10.1093/nar/gkq865) (see p. 35).
297. Pham-Duc, P. et al. (2014). "Diarrhoeal diseases among adult population in an agricultural community Hanam province, Vietnam, with high wastewater and excreta re-use". English. In: *BMC Public Health* 14.1. ISSN: 14712458. DOI: [10.1186/1471-2458-14-978](https://doi.org/10.1186/1471-2458-14-978) (see p. 95).
298. Pinto, A., C. Xi, and L. Raskin (2012). "Bacterial community structure in the drinking water microbiome is governed by filtration processes". English. In: *Environmental Science and Technology* 46.16, pp. 8851–8859. ISSN: 0013936X. DOI: [10.1021/es302042t](https://doi.org/10.1021/es302042t) (see pp. 36, 39, 137).
299. Plummer, R. et al. (2010). "The development of new environmental policies and processes in response to a crisis: the case of the multiple barrier approach for safe drinking water". English. In: *Environmental Science and Policy* 13.6, pp. 535–548. ISSN: 14629011. DOI: [10.1016/j.envsci.2010.05.004](https://doi.org/10.1016/j.envsci.2010.05.004) (see p. 4).
300. Plutzer, J., A. Törökne, and P. Karanis (2010). "Combination of ARAD microfibre filtration and LAMP methodology for simple, rapid and cost-effective detection of human pathogenic *Giardia duodenalis* and *Cryptosporidium* spp. in drinking water". English. In: *Letters in Applied Microbiology* 50.1, pp. 82–88. ISSN: 02668254. DOI: [10.1111/j.1472-765X.2009.02758.x](https://doi.org/10.1111/j.1472-765X.2009.02758.x) (see p. 26).
301. Prüss, A. et al. (2002). "Estimating the burden of disease from water, sanitation, and hygiene at a global level". English. In: *Environmental Health Perspectives* 110.5, pp. 537–542. ISSN: 00916765 (see pp. 14, 15, 45).
302. Python Software Foundation (2013). *Python language reference, version 2.7*. URL: <http://www.python.org> (see p. 161).
303. Qin, J. et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing". In: *Nature* 464.7285, pp. 59–65 (see pp. 34, 135).
304. Qin, X. et al. (2012). "Complete genome sequence of *Enterococcus faecium* strain TX16 and comparative genomic analysis of *Enterococcus faecium* genomes". English. In: *BMC Microbiology* 12. ISSN: 14712180. DOI: [10.1186/1471-2180-12-135](https://doi.org/10.1186/1471-2180-12-135) (see p. 136).

305. R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/> (see pp. 53, 80, 161).
306. Rakocinski, C., J. Lyczkowski-Shultz, and S. Richardson (1996). "Ichthyoplankton assemblage structure in Mississippi sound as revealed by canonical correspondence analysis". English. In: *Estuarine, Coastal and Shelf Science* 43.2, pp. 237–257. ISSN: 02727714. DOI: [10.1006/ecss.1996.0067](https://doi.org/10.1006/ecss.1996.0067) (see p. 155).
307. Rappé, M. and S. Giovannoni (2003). "The uncultured microbial majority". English. In: *Annual Review of Microbiology* 57, pp. 369–394. ISSN: 00664227. DOI: [10.1146/annurev.micro.57.030502.090759](https://doi.org/10.1146/annurev.micro.57.030502.090759) (see p. 136).
308. Read, C., P. Monis, and R. Thompson (2004). "Discrimination of all genotypes of *Giardia duodenalis* at the glutamate dehydrogenase locus using PCR-RFLP". English. In: *Infection, Genetics and Evolution* 4.2, pp. 125–130. ISSN: 15671348. DOI: [10.1016/j.meegid.2004.02.001](https://doi.org/10.1016/j.meegid.2004.02.001) (see p. 114).
309. Reece, R. and A. Maxwell (1991). "DNA gyrase: Structure and function". English. In: *Critical Reviews in Biochemistry and Molecular Biology* 26.3-4, pp. 335–375. ISSN: 10409238 (see p. 35).
310. Reischer, G. et al. (2007). "A quantitative real-time PCR assay for the highly sensitive and specific detection of human faecal influence in spring water from a large alpine catchment area". English. In: *Letters in Applied Microbiology* 44.4, pp. 351–356. ISSN: 02668254. DOI: [10.1111/j.1472-765X.2006.02094.x](https://doi.org/10.1111/j.1472-765X.2006.02094.x) (see p. 26).
311. Reissmann, F. and W. Uhl (2006). "Ultrafiltration for the reuse of spent filter backwash water from drinking water treatment". English. In: *Desalination* 198.1-3, pp. 225–235. ISSN: 00119164. DOI: [10.1016/j.desal.2006.03.517](https://doi.org/10.1016/j.desal.2006.03.517) (see p. 20).
312. Rompre, A. et al. (2002). "Detection and enumeration of coliforms in drinking water: Current methods and emerging approaches". English. In: *Journal of Microbiological Methods* 49.1, pp. 31–54. ISSN: 01677012. DOI: [10.1016/S0167-7012\(01\)00351-7](https://doi.org/10.1016/S0167-7012(01)00351-7) (see p. 23).
313. Rose, H. E. (2002). *Linear algebra: A pure mathematical approach*. Berlin: Springer. ISBN: 3-7643-6792-X (see p. 85).
314. Rossi, A. et al. (2005). "Determinants of once-only contact in a community-based psychiatric service". English. In: *Social Psychiatry and Psychiatric Epidemiology* 40.1, pp. 50–56. ISSN: 09337954. DOI: [10.1007/s00127-005-0845-x](https://doi.org/10.1007/s00127-005-0845-x) (see pp. 69, 70).
315. Rothberg, J. et al. (2011). "An integrated semiconductor device enabling non-optical genome sequencing". English. In: *Nature* 475.7356, pp. 348–352. ISSN: 00280836. DOI: [10.1038/nature10242](https://doi.org/10.1038/nature10242) (see pp. 8, 9).
316. Salter, S. J. et al. (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses". In: *BMC biology* 12.1, p. 87 (see pp. 157, 158).
317. Sanger, F. and A. Coulson (1975). "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". English. In: *Journal of Molecular Biology* 94.3, pp. 441–448. ISSN: 00222836 (see p. 5).
318. Sangwan, N. et al. (2012). "Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels". English. In: *PLoS ONE* 7.9. ISSN: 19326203. DOI: [10.1371/journal.pone.0046219](https://doi.org/10.1371/journal.pone.0046219) (see p. 135).
319. Savill, M. et al. (2001). "Enumeration of *Campylobacter* in New Zealand recreational and drinking waters". English. In: *Journal of Applied Microbiology* 91.1, pp. 38–46. ISSN: 13645072. DOI: [10.1046/j.1365-2672.2001.01337.x](https://doi.org/10.1046/j.1365-2672.2001.01337.x) (see pp. 33, 34, 70).
320. Schloss, P. and S. Westcott (2011). "Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis". English. In: *Applied and Environmental Microbiology* 77.10, pp. 3219–3226. ISSN: 00992240. DOI: [10.1128/AEM.02810-10](https://doi.org/10.1128/AEM.02810-10) (see pp. 35, 36).



321. Schmeisser, C. et al. (2003). "Metagenome survey of biofilms in drinking-water networks". In: *Applied and environmental microbiology* 69.12, pp. 7298–7309 (see pp. 14, 34, 36, 39, 40, 43, 137).
322. Schousboe, M., J. Lynds, and C. Ambrose (2013). "Increased incidence of *Escherichia coli* bacteremia post-Christchurch earthquake 2011: Possible associations". English. In: *Prehospital and Disaster Medicine* 28.3, pp. 202–209. ISSN: 1049023X. DOI: [10.1017/S1049023X13000137](https://doi.org/10.1017/S1049023X13000137) (see p. 33).
323. Schwartz, J. b. (2000). "The distributed lag between air pollution and daily deaths". English. In: *Epidemiology* 11.3, pp. 320–326. ISSN: 10443983. DOI: [10.1097/00001648-200005000-00016](https://doi.org/10.1097/00001648-200005000-00016) (see p. 95).
324. Scott, T. et al. (2002). "Microbial source tracking: Current methodology and future directions". English. In: *Applied and Environmental Microbiology* 68.12, pp. 5796–5803. ISSN: 00992240. DOI: [10.1128/AEM.68.12.5796-5803.2002](https://doi.org/10.1128/AEM.68.12.5796-5803.2002) (see p. 29).
325. Segonds, C. et al. (2009). "Microbiological and epidemiological features of clinical respiratory isolates of *Burkholderia gladioli*". English. In: *Journal of Clinical Microbiology* 47.5, pp. 1510–1516. ISSN: 00951137. DOI: [10.1128/JCM.02489-08](https://doi.org/10.1128/JCM.02489-08) (see pp. 149, 151).
326. Sen, A. and M. S. Srivastava (1990). *Regression analysis: Theory, methods, and applications*. Springer. ISBN: 3-540-97211-0 (see p. 118).
327. Shah, V. et al. (2007). "Evaluating potential applications of faecal sterols in distinguishing sources of faecal contamination from mixed faecal samples". English. In: *Water Research* 41.16, pp. 3691–3700. ISSN: 00431354. DOI: [10.1016/j.watres.2007.04.006](https://doi.org/10.1016/j.watres.2007.04.006) (see p. 29).
328. Shanks, O. et al. (2006). "Competitive metagenomic DNA hybridization identifies host-specific microbial genetic markers in cow fecal samples". English. In: *Applied and Environmental Microbiology* 72.6, pp. 4054–4060. ISSN: 00992240. DOI: [10.1128/AEM.00023-06](https://doi.org/10.1128/AEM.00023-06) (see p. 30).
329. Shanks, O. et al. (2007). "Identification of bacterial DNA markers for the detection of human fecal pollution in water". English. In: *Applied and Environmental Microbiology* 73.8, pp. 2416–2422. ISSN: 00992240. DOI: [10.1128/AEM.02474-06](https://doi.org/10.1128/AEM.02474-06) (see p. 30).
330. Shen, Y. et al. (2008). "Projection of future world water resources under SRES scenarios: Water withdrawal". English. In: *Hydrological Sciences Journal* 53.1, pp. 11–33. ISSN: 02626667. DOI: [10.1623/hysj.53.1.11](https://doi.org/10.1623/hysj.53.1.11) (see p. 13).
331. Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing". English. In: *Nature Biotechnology* 26.10, pp. 1135–1145. ISSN: 10870156. DOI: [10.1038/nbt1486](https://doi.org/10.1038/nbt1486) (see p. 7).
332. Sheppard, S. et al. (2009). "*Campylobacter* genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6". English. In: *International Journal of Food Microbiology* 134.1-2, pp. 96–103. ISSN: 01681605. DOI: [10.1016/j.ijfoodmicro.2009.02.010](https://doi.org/10.1016/j.ijfoodmicro.2009.02.010) (see pp. 97, 135).
333. Shi, P. et al. (2013). "Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water". English. In: *Water Research* 47.1, pp. 111–120. ISSN: 00431354. DOI: [10.1016/j.watres.2012.09.046](https://doi.org/10.1016/j.watres.2012.09.046) (see pp. 36, 40, 137, 157).
334. Shin, S. et al. (2013). "Advantages of single-molecule real-time sequencing in high-GC content genomes". English. In: *PLoS ONE* 8.7. ISSN: 19326203. DOI: [10.1371/journal.pone.0068824](https://doi.org/10.1371/journal.pone.0068824) (see p. 9).
335. Simmons, G. et al. (2001). "Contamination of potable roof-collected rainwater in Auckland, New Zealand". English. In: *Water Research* 35.6, pp. 1518–1524. ISSN: 00431354. DOI: [10.1016/S0043-1354\(00\)00420-6](https://doi.org/10.1016/S0043-1354(00)00420-6) (see p. 33).
336. Simpson, E. (1949). "Measurement of diversity". English. In: *Nature* 163.4148, p. 688. ISSN: 00280836 (see p. 142).
337. Sipos, R. et al. (2007). "Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis". English. In: *FEMS Microbiology Ecology* 60.2, pp. 341–350. ISSN: 01686496. DOI: [10.1111/j.1574-6941.2007.00283.x](https://doi.org/10.1111/j.1574-6941.2007.00283.x) (see p. 36).

338. Siqueira Jr., J., A. Fouad, and I. Rôças (2012). “Pyrosequencing as a tool for better understanding of human microbiomes”. English. In: *Journal of Oral Microbiology* 4.2012. ISSN: 20002297. DOI: [10.3402/jom.v4i0.10743](#) (see p. 7).
339. Smith, A. et al. (2006). “Outbreaks of waterborne infectious intestinal disease in England and Wales, 1992–2003”. English. In: *Epidemiology and Infection* 134.6, pp. 1141–1149. ISSN: 09502688. DOI: [10.1017/S0950268806006406](#) (see p. 16).
340. Smith Jr., J. and J. Perdek (2004). “Assessment and management of watershed microbial contaminants”. English. In: *Critical reviews in environmental science and technology* 34.2, pp. 109–139. ISSN: 10643389. DOI: [10.1080/10643380490430663](#) (see p. 13).
341. Sneath, P. H. and R. R. Sokal (1973). *Numerical taxonomy; the principles and practice of numerical classification*. W. H. Freeman, San Francisco (see p. 140).
342. Snel, S. J., M. G. Baker, and K. Venugopal (2009). “The epidemiology of giardiasis in New Zealand, 1997–2006”. In: *NZ Med J* 122.1290, pp. 62–75 (see p. 14).
343. Sokal, R. R. and P. H. Sneath (1963). *Principles of numerical taxonomy*. W. H. Freeman, San Francisco (see p. 140).
344. Somboonna, N. et al. (2012). “Metagenomic profiles of free-living archaea, bacteria and small eukaryotes in coastal areas of Sichang island, Thailand.” English. In: *BMC genomics* 13 Suppl 7. ISSN: 14712164 (see p. 156).
345. Southwood, T. R. E. and P. A. Henderson (2009). *Ecological methods*. John Wiley & Sons. ISBN: 0-632-05477-8 (see p. 142).
346. Speicher, M. and N. Carter (2005). “The new cytogenetics: Blurring the boundaries with molecular biology”. English. In: *Nature Reviews Genetics* 6.10, pp. 782–792. ISSN: 14710056. DOI: [10.1038/nrg1692](#) (see p. 25).
347. Statistics New Zealand (2013). *2013 census QuickStats about a place: New Zealand* (see p. 13).
348. Stehr-Green, J. et al. (1991). “Waterborne outbreak of *Campylobacter jejuni* in Christchurch: The importance of a combined epidemiologic and microbiologic investigation.” English. In: *New Zealand Medical Journal* 104.918, pp. 356–358. ISSN: 00288446 (see p. 31).
349. Stevens, C. and I. Hume (1998). “Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients”. English. In: *Physiological Reviews* 78.2, pp. 393–427. ISSN: 00319333 (see p. 135).
350. Stinear, T. et al. (2008). “Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*”. English. In: *Genome Research* 18.5, pp. 729–741. ISSN: 10889051. DOI: [10.1101/gr.075069.107](#) (see p. 136).
351. Strobl, C. et al. (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. English. In: *BMC Bioinformatics* 8. ISSN: 14712105. DOI: [10.1186/1471-2105-8-25](#) (see p. 55).
352. Strobl, C. et al. (2008). “Conditional variable importance for random forests”. English. In: *BMC Bioinformatics* 9. ISSN: 14712105. DOI: [10.1186/1471-2105-9-307](#) (see p. 55).
353. Strobl, C., J. Malley, and G. Tutz (2009). “An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests”. English. In: *Psychological Methods* 14.4, pp. 323–348. ISSN: 1082989X. DOI: [10.1037/a0016973](#) (see pp. 53, 54).
354. Stucki, U. et al. (1995). “Identification of *Campylobacter jejuni* on the basis of a species-specific gene that encodes a membrane protein”. English. In: *Journal of Clinical Microbiology* 33.4, pp. 855–859. ISSN: 00951137 (see pp. 50, 113).

355. Suffet, I. and P. Rosenfeld (2007). "The anatomy of odour wheels for odours of drinking water, wastewater, compost and the urban environment". English. In: *Water Science and Technology* 55.5. Ed. by B. G. R. J. Watson S.B. Brownlee B., pp. 335–344. ISSN: 02731223. DOI: [10.2166/wst.2007.196](https://doi.org/10.2166/wst.2007.196) (see p. 3).
356. Ter Braak, C. and P. Verdonschot (1995). "Canonical correspondence analysis and related multivariate methods in aquatic ecology". English. In: *Aquatic Sciences* 57.3, pp. 255–289. ISSN: 10151621 (see p. 155).
357. Conservation Act 1987. URL: <http://www.legislation.govt.nz/act/public/1987/0065/latest/DLM103610.html> (see p. 105).
358. National Parks Act 1980. URL: <http://www.legislation.govt.nz/act/public/1980/0066/latest/DLM36963.html> (see p. 105).
359. Reserves Act 1977. URL: <http://www.legislation.govt.nz/act/public/1977/0066/latest/whole.html> (see p. 105).
360. Thiruppathiraja, C. et al. (2011). "Development of electrochemical based sandwich enzyme linked immunosensor for *Cryptosporidium parvum* detection in drinking water". English. In: *Journal of Environmental Monitoring* 13.10, pp. 2782–2787. ISSN: 14640325. DOI: [10.1039/c1em10372e](https://doi.org/10.1039/c1em10372e) (see p. 25).
361. Thomas, M. et al. (2006). "A role of high impact weather events in waterborne disease outbreaks in Canada, 1975–2001". English. In: *International Journal of Environmental Health Research* 16.3, pp. 167–180. ISSN: 09603123. DOI: [10.1080/09603120600641326](https://doi.org/10.1080/09603120600641326) (see p. 76).
362. Thomas, V. et al. (2004). "Amoebae in domestic water systems: Resistance to disinfection treatments and implication in *Legionella* persistence". English. In: *Journal of Applied Microbiology* 97.5, pp. 950–963. ISSN: 13645072. DOI: [10.1111/j.1365-2672.2004.02391.x](https://doi.org/10.1111/j.1365-2672.2004.02391.x) (see p. 14).
363. Thompson, J. and P. Milos (2011). "The properties and applications of single-molecule DNA sequencing". English. In: *Genome Biology* 12.2. ISSN: 14747596. DOI: [10.1186/gb-2011-12-2-217](https://doi.org/10.1186/gb-2011-12-2-217) (see p. 9).
364. Thompson, J. and K. Steinmann (2010). "Single molecule sequencing with a HeliScope genetic analysis system". English. In: *Current Protocols in Molecular Biology* SUPPL. 92, pp. 7.10.1–7.10.14. ISSN: 19343639. DOI: [10.1002/0471142727.mb0710s92](https://doi.org/10.1002/0471142727.mb0710s92) (see pp. 7, 9).
365. Thornley, C., M. Baker, and C. Nicol (2002). "The rising incidence of *Salmonella* infection in New Zealand, 1995–2001". English. In: *New Zealand Public Health Report* 9.4, pp. 25–28. ISSN: 11730250 (see pp. 32, 33).
366. Till, D. et al. (2008). "Large-scale freshwater microbiological study: Rationale, results and risks". English. In: *Journal of Water and Health* 6.4, pp. 443–460. ISSN: 14778920. DOI: [10.2166/wh.2008.071](https://doi.org/10.2166/wh.2008.071) (see pp. 33, 70).
367. Tomb, J.-F. et al. (1997). "The complete genome sequence of the gastric pathogen *Helicobacter pylori*". English. In: *Nature* 388.6642, pp. 539–547. ISSN: 00280836. DOI: [10.1038/41483](https://doi.org/10.1038/41483) (see p. 136).
368. Tornevi, A., O. Bergstedt, and B. Forsberg (2014). "Precipitation effects on microbial pollution in a river: Lag structures and seasonal effect modification". English. In: *PLoS ONE* 9.5. ISSN: 19326203. DOI: [10.1371/journal.pone.0098546](https://doi.org/10.1371/journal.pone.0098546) (see pp. 76, 95).
369. Tsukamura, M., H. Van Der Meulen, and W. Grabow (1983). "Numerical taxonomy of rapidly growing, scotochromogenic mycobacteria of the *Mycobacterium parafortuitum* complex: *Mycobacterium austroafricanum* sp. nov. and *Mycobacterium diernhoferi* sp. nov., nom. rev." English. In: *International Journal of Systematic Bacteriology* 33.3, pp. 460–469. ISSN: 00207713 (see p. 149).
370. Tuomisto, H. (2010). "A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity". English. In: *Ecography* 33.1, pp. 2–22. ISSN: 09067590. DOI: [10.1111/j.1600-0587.2009.05880.x](https://doi.org/10.1111/j.1600-0587.2009.05880.x) (see p. 142).

371. Turnbaugh, P. et al. (2007). “The human microbiome project”. English. In: *Nature* 449.7164, pp. 804–810. ISSN: 00280836. DOI: [10.1038/nature06244](https://doi.org/10.1038/nature06244) (see p. 135).
372. Turner, D. (2011). “Next-generation DNA sequencing technologies”. In: *Encyclopedia of Analytical Chemistry*. ISSN: 0470027312 (see pp. 10, 36).
373. United Kingdom Centre for Infectious Disease Surveillance and Control (2013). *Gastrointestinal infections: 2011 and 2012 report*. Tech. rep. London, United Kingdom: Public Health England. URL: [http://www.hpa.org.uk/webw/HPAweb&HPAwebStandard/HPAweb\\_C/1317140455257](http://www.hpa.org.uk/webw/HPAweb&HPAwebStandard/HPAweb_C/1317140455257) (see p. 75).
374. United States Environmental Protection Agency (2000). *Improved enumeration methods for the recreational water quality indicators: Enterococci and Escherichia coli*. Tech. rep. EPA/821/R-97/004. United States Environmental Protection Agency. URL: [http://water.epa.gov/type/oceb/beaches/upload/2006\\_06\\_19\\_beaches\\_rvsdman.pdf](http://water.epa.gov/type/oceb/beaches/upload/2006_06_19_beaches_rvsdman.pdf) (see p. 31).
375. — (2002). *Method 1604: Total coliforms and Escherichia coli in water by membrane filtration using a simultaneous detection technique (MI medium)*. Tech. rep. EPA-821-R-02-024. U.S. EPA Office of Water, Office of Science and Technology, Northwest, Washington, D.C.: U.S. Environmental Protection Agency Office of Water (see p. 23).
376. — (2004). *Drinking water treatment*. Tech. rep. United States Environmental Protection Agency (see p. 20).
377. — (2012). *Method 1623.1: Cryptosporidium and Giardia in Water by Filtration/IMS/FA*. Tech. rep. United States Environmental Protection Agency. URL: <http://www.epa.gov/nerlcwww/online.html#protos> (see pp. 50, 114).
378. Unwin, M. (2014). *National scale mapping of rivers showing a threshold dissolved inorganic nitrogen concentration of 0.8 mg/L*. Tech. rep. Christchurch: National Institute of Water and Atmospheric Research Limited. URL: <http://www.interest.co.nz/sites/default/files/DIN%20NIWA.pdf> (see p. 46).
379. Urwin, R. and M. Maiden (2003). “Multi-locus sequence typing: A tool for global epidemiology”. English. In: *Trends in Microbiology* 11.10, pp. 479–487. ISSN: 0966842X. DOI: [10.1016/j.tim.2003.08.006](https://doi.org/10.1016/j.tim.2003.08.006) (see p. 112).
380. Van Ingen, J. et al. (2008). “Clinical relevance of *Mycobacterium simiae* in pulmonary samples”. English. In: *European Respiratory Journal* 31.1, pp. 106–109. ISSN: 09031936. DOI: [10.1183/09031936.00076107](https://doi.org/10.1183/09031936.00076107) (see pp. 149, 151).
381. Vanlaera, E. et al. (2008). “*Burkholderia latens* sp. nov., *Burkholderia diffusa* sp. nov., *Burkholderia arboris* sp. nov., *Burkholderia seminalis* sp. nov., and *Burkholderia metallica* sp. nov., novel species within the *Burkholderia cepacia* complex”. English. In: *International Journal of Systematic and Evolutionary Microbiology* 58.7, pp. 1580–1590. ISSN: 14665026. DOI: [10.1099/ijs.0.65634-0](https://doi.org/10.1099/ijs.0.65634-0) (see p. 151).
382. Venter, J. (2001). “The human genome”. English. In: *Journal of Biolaw and Business* 4.3, pp. 98–102. ISSN: 10955127 (see p. 136).
383. Venter, J. et al. (2004). “Environmental genome shotgun sequencing of the Sargasso Sea”. English. In: *Science* 304.5667, pp. 66–74. ISSN: 00368075. DOI: [10.1126/science.1093857](https://doi.org/10.1126/science.1093857) (see p. 135).
384. Villanueva, C. et al. (2014). “Assessing exposure and health consequences of chemicals in drinking water: Current state of knowledge and research needs”. English. In: *Environmental Health Perspectives* 122.3, pp. 213–221. ISSN: 00916765. DOI: [10.1289/ehp.1206229](https://doi.org/10.1289/ehp.1206229) (see p. 2).
385. Volk, C. et al. (2000). “Impact of enhanced and optimized coagulation on removal of organic matter and its biodegradable fraction in drinking water”. English. In: *Water Research* 34.12, pp. 3247–3257. ISSN: 00431354. DOI: [10.1016/S0043-1354\(00\)00033-6](https://doi.org/10.1016/S0043-1354(00)00033-6) (see p. 20).
386. Von Mering, C. et al. (2007). “Quantitative phylogenetic assessment of microbial communities in diverse environments”. In: *Science* 315.5815, pp. 1126–1130 (see p. 34).

387. Vujicic, M. and J. Sanderson (2008). *Linear algebra thoroughly explained*. Springer. ISBN: 978-3-540-74637-9 (see p. 85).
388. Waarbeek, H. ter et al. (2010). “Waterborne gastroenteritis outbreak at a scouting camp caused by two norovirus genogroups: GI and GII”. English. In: *Journal of Clinical Virology* 47.3, pp. 268–272. ISSN: 13866532. DOI: [10.1016/j.jcv.2009.12.002](https://doi.org/10.1016/j.jcv.2009.12.002) (see pp. 106, 107, 129).
389. Wagenaar, J., D. Mevius, and A. Havelaar (2006). “*Campylobacter* in primary animal production and control strategies to reduce the burden of human campylobacteriosis”. English. In: *OIE Revue Scientifique et Technique* 25.2, pp. 581–594. ISSN: 02531933 (see p. 97).
390. Wang, F., L. Jiang, and B. Ge (2012a). “Loop-mediated isothermal amplification assays for detecting Shiga toxin-producing *Escherichia coli* in ground beef and human stools”. English. In: *Journal of Clinical Microbiology* 50.1, pp. 91–97. ISSN: 00951137. DOI: [10.1128/JCM.05612-11](https://doi.org/10.1128/JCM.05612-11) (see p. 26).
391. Wang, Y. et al. (2012b). “Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of Illumina tags”. English. In: *Applied and Environmental Microbiology* 78.23, pp. 8264–8271. ISSN: 00992240. DOI: [10.1128/AEM.01821-12](https://doi.org/10.1128/AEM.01821-12) (see pp. 36, 137).
392. Ward, N., R. Wolfe, and B. Olson (1984). “Effect of pH, application technique, and chlorine-to-nitrogen ratio on disinfectant activity of inorganic chloramines with pure culture bacteria”. English. In: *Applied and Environmental Microbiology* 48.3, pp. 508–514. ISSN: 00992240 (see p. 21).
393. Webster, R. and M. Oliver (2008). *Geostatistics for environmental scientists: Second edition*. English. John Wiley & Sons, Ltd., pp. 1–315. ISBN: 9780470028582. DOI: [10.1002/9780470517277](https://doi.org/10.1002/9780470517277) (see pp. 81–83, 97).
394. Wetterstrand, K. (2014). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: <http://www.genome.gov/sequencingcosts/> (see pp. 10, 36).
395. Wheeler, D. et al. (2000). “Database resources of the National Center for Biotechnology Information”. English. In: *Nucleic Acids Research* 28.1, pp. 10–14. ISSN: 03051048 (see p. 143).
396. Wilke, M., A. Lovering, and N. Strynadka (2005). “ $\beta$ -Lactam antibiotic resistance: A current structural perspective”. English. In: *Current Opinion in Microbiology* 8.5, pp. 525–533. ISSN: 13695274. DOI: [10.1016/j.mib.2005.08.016](https://doi.org/10.1016/j.mib.2005.08.016) (see p. 43).
397. Williamson, W. et al. (2011). “Enteric viruses in New Zealand drinking-water sources”. English. In: *Water Science and Technology* 63.8, pp. 1744–1751. ISSN: 02731223. DOI: [10.2166/wst.2011.117](https://doi.org/10.2166/wst.2011.117) (see p. 46).
398. Winkworth, C. L. (2010). “Land-use change and emerging public health risks in New Zealand: assessing *Giardia* risks”. In: *Journal of the New Zealand Medical Association* 123.1322 (see p. 129).
399. Winkworth, C. et al. (2008). “Molecular characterization of *Giardia* isolates from calves and humans in a region in which dairy farming has recently intensified”. English. In: *Applied and Environmental Microbiology* 74.16, pp. 5100–5105. ISSN: 00992240. DOI: [10.1128/AEM.00232-08](https://doi.org/10.1128/AEM.00232-08) (see p. 129).
400. Woese, C. and G. Fox (1977). “Phylogenetic structure of the prokaryotic domain: The primary kingdoms”. English. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11, pp. 5088–5090. ISSN: 00278424 (see p. 34).
401. Wooley, J., A. Godzik, and I. Friedberg (2010). “A primer on metagenomics”. English. In: *PLoS Computational Biology* 6.2. ISSN: 1553734X. DOI: [10.1371/journal.pcbi.1000667](https://doi.org/10.1371/journal.pcbi.1000667) (see p. 136).
402. World Health Organization (1997). *Guidelines for drinking-water quality*. Tech. rep. URL: [http://www.who.int/water\\_sanitation\\_health/dwq/gdwqvol32ed.pdf](http://www.who.int/water_sanitation_health/dwq/gdwqvol32ed.pdf) (see p. 23).
403. — (2003). *Guidelines for safe recreational water environments. Volume 1: Coastal and fresh waters*. Tech. rep. URL: [http://www.who.int/water\\_sanitation\\_health/bathing/srwe1/en/](http://www.who.int/water_sanitation_health/bathing/srwe1/en/) (see p. 30).

404. — (2004). *Water treatment and pathogen control: Process efficiency in achieving safe drinking water*. Tech. rep. URL: [http://www.who.int/water\\_sanitation\\_health/dwq/en/watreatpath.pdf](http://www.who.int/water_sanitation_health/dwq/en/watreatpath.pdf) (see pp. 15, 21).
405. — (2011). *Guidelines for drinking-water quality*. Tech. rep. URL: [http://www.who.int/water\\_sanitation\\_health/dwq/guidelines/en/](http://www.who.int/water_sanitation_health/dwq/guidelines/en/) (see pp. 2, 3, 14, 15, 22, 30).
406. — (2012). *Animal waste, water quality and human health*. Tech. rep. (see p. 45).
407. — (2013). *Progress on sanitation and drinking-water–2013 update*. Tech. rep. URL: [http://apps.who.int/iris/bitstream/10665/81245/1/9789241505390\\_eng.pdf](http://apps.who.int/iris/bitstream/10665/81245/1/9789241505390_eng.pdf) (see p. 13).
408. — (2014a). *Antimicrobial resistance: Global report on surveillance*. Tech. rep. URL: <http://www.who.int/drugresistance/documents/surveillance-report/en/> (see p. 157).
409. — (2014b). *World health statistics 2014*. Tech. rep. URL: <http://www.who.int/gho/publications/en/> (see p. 75).
410. Wu, E.-Y. and S.-L. Kuo (2012). “Applying a multivariate statistical analysis model to evaluate the water quality of a watershed”. English. In: *Water Environment Research* 84.12, pp. 2075–2085. ISSN: 10614303. DOI: [10.2175/106143012X13415215906979](https://doi.org/10.2175/106143012X13415215906979) (see p. 46).
411. Xie, Y. (2014). *Knitr: A general-purpose package for dynamic report generation in R. R package version 1.7*. URL: <http://yihui.name/knitr/> (see p. 161).
412. Yalow, R. and S. Berson (1960). “Immunoassay of endogenous plasma insulin in man”. English. In: *The Journal of clinical investigation* 39, pp. 1157–1175. ISSN: 00219738 (see p. 24).
413. Zablocki, O. et al. (2014). “High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils”. In: *Applied and environmental microbiology* 80.22, pp. 6888–6897 (see p. 135).
414. Zanutti, A. et al. (2000). “Generalized additive distributed lag models: Quantifying mortality displacement”. In: *Biostatistics* 1.3, pp. 279–292 (see pp. 85, 95).



# Index

- Acidobacteria*, 39
- Actinobacteria*, 35, 39
- Arcobacter*, 147, 149, 156
- Betaproteobacteria*, 39
- Burkholderia*, 142, 149
- C. coli*, 33, 50, 112, 115, 116, 120, 122, 123, 129
- C. jejuni*, 33, 50, 112, 115, 116, 120, 122, 123, 129
- C. sp. nova* I, 112, 120, 122, 123, 129
- Campylobacteraceae*, 142, 147, 149, 156
- Campylobacter*, 3, 16, 33, 35, 41, 47, 49, 50, 52, 55, 56, 58, 63, 64, 66, 67, 69–72, 78, 87, 95–97, 107, 110, 112, 114–116, 120, 122–125, 127–130, 142, 147, 149, 159, 160, 163, 164
- Cryptosporidium*, 16, 20–22, 28, 41, 46, 47, 50–52, 56, 58–60, 63, 64, 66, 69, 71–73, 78, 87, 110, 114, 122, 123, 127, 159, 163
- Cyanobacteria*, 3, 39
- Deltaproteobacteria*, 39
- Dracunculus medinensis*, 3
- E. coli*, 16, 22, 23, 25, 28, 30–33, 41, 42, 46, 47, 50, 52, 56, 58, 59, 63, 64, 66, 69–72, 106, 107, 110, 114–116, 120, 122, 127, 128, 156, 159
- Enterococci*, 34
- Enterococcus*, 42, 107
- Epsilonproteobacteria*, 39
- Firmicutes*, 39
- Gammaproteobacteria*, 39, 41
- Giardia*, 16, 20, 22, 41, 46–48, 50–52, 56, 58–60, 63, 64, 67, 69, 71, 72, 78, 87, 95, 110, 114, 116, 122, 123, 128, 129, 159, 163
- Legionella*, 3, 14, 142, 149, 151
- Mycobacterium*, 25, 40, 142, 149
- Norovirus*, 22, 31, 106, 107
- Planctomycetes*, 39
- Proteobacteria*, 39–41, 43
- Providencia*, 16
- Salmonella*, 3, 78, 87, 95
- Schistosoma*, 3
- Staphylococcus aureus*, 9
- Sulfurospirillum*, 147, 149
- Vibrio cholerae*, 3
- Burkholderia*, 149
- R**, 36, 37, 55, 57, 80, 84, 85, 120, 142
- MySQL**, 78, 166
- Phyloseq**, 142
- RISmed**, 36, 37
- SEED**, 39, 143
- SolexaQA**, 141, 143
- Splitstree**, 143
- dlnm**, 85
- lme4**, 57, 120
- metaAmos**, 166
- mtry**, 54
- ntree**, 54
- party**, 53, 55
- 16S, 25, 26, 30, 34–37, 112, 136–143, 145, 147, 149, 155, 156, 158, 165, 166
- 454, 7, 8, 36, 137
- absence, 46, 52, 63, 72
- abstraction, 10, 47, 52, 58, 60, 64, 86, 87, 93, 94, 107, 108, 110, 122, 128, 138, 162, 163
- abundance score, 142, 143, 147, 149, 156
- account, 76, 79, 83, 84, 87, 97
- adenovirus, 3
- aerobic, 3
- aesthetic, 2
- algae, 2
- algorithm, 53
- align, 141
- allelic profile, 110, 112, 116, 117, 125
- ambient, 75, 78, 86, 95
- Ambion, Inc., 138
- amine, 3
- amino acid, 39, 40, 149
- amoebae, 14, 43
- anaerobic, 3
- analysis, 77, 79, 82, 84, 87, 91, 93–97
- animal, 2–4, 13–15, 21, 29, 30, 32–34, 46, 52, 56, 66, 109, 120, 128, 155, 163
- anneal, 5, 26, 139
- antibiotic, 40, 43, 155
- antimicrobial, 143, 155–158
- Applied Biosystems, 7
- April, 63
- Aqua-Glo<sup>®</sup>, 50
- archaea, 135
- ArcMap, 52
- ArcView 3.2a, 52
- area, 47, 50, 52, 56



- arsenic, 2
- aspirate, 50
- assemblage, 114, 116, 124, 129
- associate, 10, 11, 16, 31–34, 40, 41, 43, 45–47, 53, 54, 69–71, 76–78, 87, 94–97, 106, 107, 125, 127, 130, 137, 142, 149, 155
- association, 34, 46, 53, 70, 76, 77, 84, 94, 159
- assumption, 80, 81, 84, 85, 94, 96, 97, 118, 119
- Astrovirus, 16
- AsureQuality, 52
- attribute, 10, 32, 41, 45, 46, 52, 53, 58, 105, 106, 108, 116, 156, 160
- Auckland, 58
- August, 76
- Australia, 45, 46, 75, 95, 149
- automate, 6, 37, 143
- autumn, 33, 60, 63, 70, 71
  
- bacteria, 3, 14, 16, 21–23, 25, 26, 28, 29, 35, 39–41, 43, 64, 69, 71, 72, 110, 135, 142, 149, 157–159
- barrier, 4, 15, 20
- base, 6–8, 140
  - pair, 6, 140, 141, 143
- Basel, 7
- basis, 81, 85, 86
  - function, 85, 86
- beef cattle, 56, 63, 66, 72
- Belgium, 107
- bias, 55, 96, 136
- Big Huia Creek, 47
- binomial, 56, 119, 120
- biofilm, 14, 34, 40, 43
- biological, 1, 3, 46, 55
- biopore, 9
- biplot, 80, 90
- bird, 123–125, 129, 145, 155, 157, 163
- blood agar, 50, 112, 123, 164
- blue, 6, 51
- Bolton's broth, 49, 112, 114
- bootstrap, 54
- borehole, 16, 17, 45, 47, 77
- Brisbane, 149
- bubble plot, 91
- BX 60 fluorescence microscope, 50
- by-product, 2, 20, 30, 42
  
- calculate, 52, 80, 82, 83
- calculation, 96, 97
- calendar, 60, 63
- California, 7, 9, 50, 138
  
- camera, 8
- campground, 32, 37, 106–110, 120–122, 125, 127–130, 137, 138, 142, 143, 145, 155, 156, 158–164, 166
- Campylobacteriosis, 31
- campylobacteriosis, 16, 31, 32, 41, 87, 90, 91, 93–97, 106, 135, 160
- Canada, 76, 95
- canal, 77
- cancer, 2
- carbohydrate, 39, 149
- carbon, 5
- case, 15, 16, 31–33, 41, 77, 78, 80, 82, 84, 87, 89–91, 93, 94, 96, 97, 106, 107, 118, 119, 136, 155, 160, 163
- catchment, 2–4, 10, 11, 15, 16, 26, 33, 45–47, 52, 55, 56, 58, 64, 67, 69, 70, 72, 73, 108, 110, 116, 120, 121, 127–129, 138, 162
- categorical, 53, 56, 70, 119, 120
- caterpillar plot, 67
- cattle, 30, 32, 120, 123, 124, 145, 157, 159
- Central Environmental Laboratories, 50, 114
- centrifuge, 50
- centroid, 79, 84
- challenge, 3, 11, 35, 40, 45, 136, 162, 164, 166
- chemical, 1–5, 8, 13, 20, 21, 25, 29, 30, 42
  - contamination, 2
  - pollution, 2
  - treatment, 15, 121, 129, 130, 160
- chemiluminescence, 8
- chicken, 30, 32, 156
- chimaera, 140, 141
- chloramine, 21, 40
- chlorination, 4
- chlorine, 2, 21, 40
- cholera, 15
- Christchurch, 33
- Christmas, 47
- classification, 35, 53, 54, 140, 143
- cleave, 7
- climate change, 13
- clinical, 149
- cluster, 6, 8, 117, 140, 141, 145, 147, 155, 158
- cluster analysis, 46
- co-worker, 16, 22, 30, 46, 70, 81, 95, 149
- coefficient, 86, 97, 117–119
- coliform, 22, 23, 26, 28, 42, 106, 107
- Colilert<sup>®</sup>, 50, 114
- colon, 2
- colonies, 23, 112, 164

- 
- Colorado, 106
  - colour, 2, 6
  - community, 10, 78, 80, 84, 90, 96
  - computer, 114, 137, 143, 157, 161, 165–167
  - concentration, 2–4, 20–22, 28, 33, 42, 45, 46, 56, 58–60, 63, 64, 66, 71, 76, 107, 116, 120, 122, 127, 138, 157–159, 164, 167
  - conduct, 76, 84, 94, 97
  - confidence interval, 67, 90, 143
  - confirmatory, 50
  - constant, 63, 66, 67, 80, 82, 87, 97, 117, 118
  - consumption, 13, 15, 17, 20, 25, 32, 33, 41, 46, 106, 107, 128, 159
  - contaminant, 1, 20, 138, 158
  - contaminate, 13, 41, 58, 63, 70, 128, 155, 157
  - contamination, 2, 4, 5, 13, 15, 22, 26, 28–33, 41, 42, 60, 70–73, 106, 107, 129, 138, 156–158, 160, 163
  - continuous variable, 53, 56
  - Copan Transystem, 110
  - coregionalised variable, 82
  - correlate, 22, 46, 76, 77, 79, 80, 82, 90, 91, 106, 118
  - correlation, 42, 76, 80, 94, 95, 164
  - covariance, 79–81, 83
  - cow, 156
  - creek, 16, 47, 52, 77, 108
  - cross-basis, 86, 87
  - cross-sectional, 31, 33, 107, 138
  - cryptosporidiosis, 16, 33, 76, 87, 89, 91, 93, 94, 96, 160
  - culture-dependent, 10, 29, 30
  - current study, 45–47, 52, 55, 69, 77–79, 81, 84, 85, 87, 93, 96, 97, 107, 108, 110, 112, 115, 116, 119, 120, 127–130, 141, 157, 158
  - cut-point, 53, 119
  - cyst, 20, 48, 50–52, 116, 149
  - Düren, 114
  - dairy cattle, 56, 63, 64, 66, 67, 70, 72, 73
  - data, 1, 10, 22, 32, 33, 35, 52–56, 70, 72, 76, 78–84, 87, 94–97, 116–120, 128, 129, 136, 140, 143, 160–162, 165–167
  - database, 29, 76, 78, 87, 89, 93, 96, 112, 124, 141, 143, 149, 160, 166
  - dataset, 54, 78, 79, 95, 142, 157
  - death, 15, 75
  - December, 31, 47, 60, 105, 107, 138, 155
  - deer, 120, 123
  - degenerate, 7
  - density, 52, 55, 56, 63, 64, 66, 67, 72, 73
  - dependent variable, 81, 119
  - dichotomy, 56, 120
  - dideoxy, 5
  - disadvantage, 53
  - discriminant analysis, 46
  - disease, 10, 16, 31–33, 40, 41, 45, 71, 75–79, 84, 87, 91, 94–96, 135, 156–158, 160
    - outbreak, 76
  - disinfectant, 21, 43
  - disinfection, 2, 3, 20, 21, 40
  - disk, 49, 50
  - distance, 79–81, 83, 108
  - distribution
    - network, 4, 39, 42, 47, 77, 94–96, 159
  - distribution zone, 17, 47, 58, 77
  - dog, 107, 120, 156
  - domestic, 15, 105, 120, 123, 124, 127, 145, 155
  - Don River, 76
  - Don Whitley Scientific Limited, 49, 112
  - downstream, 140, 163
  - drinking water, 1–4, 10, 11, 13–17, 20–23, 25, 28, 30–34, 36, 37, 39–43, 45–47, 56, 58, 60, 69–73, 75, 77, 78, 84, 87, 90, 93–97, 106–108, 120, 127–130, 135, 137, 142, 149, 155, 156, 158–160, 162, 167
  - duck, 30, 120, 123, 127, 129, 145
  - dye, 6–8
  - Dynabeads GC-Combo kit, 50
  - Earth, 1, 13
  - eigenvalue, 79
  - eigenvector, 79
  - electrofluorescence, 114
  - electronic, 9, 31, 36, 37, 76, 91
  - electrophoresis, 6
  - electrophysiological, 9
  - eliminate, 15
  - eliminative, 4, 5, 41, 45
  - England, 16, 75, 76
  - enrich, 49, 112, 140
  - ensemble, 53–55
  - enteric, 45
    - disease, 14–16, 45
    - pathogen, 22
    - virus, 22
  - enterococci, 22, 30, 31
  - enterovirus, 3
  - entropy, 53

- enumeration, 47, 50, 110, 114, 115, 122
- environment, 2, 3, 10, 11, 22, 29, 30, 43, 76, 136, 147, 156, 157, 160
- enzyme, 6, 9, 24, 35, 40, 114, 135
- Epicentre®, 139, 164, 165
- epidemiological, 10, 107, 162
- EpiSurv, 78, 87, 89, 93, 96, 160
- equation, 29, 51, 54, 83, 86, 117, 118, 141
- error, 54, 118, 119, 141
- estimate, 10, 11, 15, 16, 45, 58, 64, 70, 72, 80–84, 86, 87, 89, 95, 97, 105, 117–120, 127–130, 141–143, 156, 162
- eutrophication, 46
- evidence, 32–34, 39–41, 72, 75, 94–97, 129, 147, 156–158, 160
- example, 1–3, 6, 13, 15–17, 22, 23, 25, 29, 33–35, 39, 42, 43, 45, 46, 51, 53, 64, 67, 71, 72, 85, 94, 96, 106, 118, 119, 129, 130, 135, 141, 149, 156, 161, 162, 166
- explanatory variable, 53–56, 66, 67, 72, 84, 86, 117, 118, 120
- exponential, 119
- exposure, 2, 22, 23, 32, 84–86, 95, 96
  
- factor, 45, 46, 56, 76, 77, 83, 86, 94–97
  - analysis, 45, 46, 79
- faeces, 3, 16, 22, 26, 28–31, 39, 41, 106, 107, 109, 110, 112, 114–116, 120, 123–125, 127, 129, 130, 138, 139, 142, 145, 147, 149, 156, 157, 160, 162, 163, 165, 167
- fatty acid, 3
- February, 107, 138, 155
- fertiliser, 2
- field, 10
- Filta-Max®, 47, 50, 110, 114
- filter, 17, 20, 23, 47, 49–51, 110, 114, 120, 121, 127, 138, 163, 164
- filtration, 4, 15, 16, 20, 21, 97, 110, 121, 127–129, 138, 164
- first-generation, 5, 10
- fixed effect, 119, 120
- flank, 7, 25, 35, 139
- fluorescent, 5–9, 23, 50
- fluorophore, 7
- foodborne, 71, 97
- forecast, 72
- fragment, 6, 8, 136, 137, 139, 140
- framework, 41, 52, 69, 72, 119, 128
- freshwater, 1, 13, 30, 31, 34, 42, 115
- functional factor, 143, 149, 155, 156, 158
  
- G00183, 47
- G00197, 47
- G01679, 55, 58
- gamma, 119
- Garmin Limited, 110
- gastroenteritis, 15, 16, 75–77, 80, 87, 89, 90, 93, 95, 96, 106, 160
- gastrointestinal, 16, 33, 34
- gastrointestinal illness, 11, 16, 33, 41, 45, 71, 75–77, 80, 84, 87, 91, 94–97, 106, 107, 129, 130, 142, 159
- gel, 6, 110, 138
- gene, 23, 25, 26, 30, 34–36, 39, 40, 43, 110, 112, 115, 136–141, 145, 155, 160
- genome, 5–7, 10, 136
- genotype, 115
- genus, 50, 112, 141, 147, 156
- geosmin, 3
- geospatial, 10, 52, 58, 94, 108, 121, 160, 161
- geostatistics, 80, 81, 83, 84, 91
- Germany, 50, 112, 114
- giardiasis, 16, 32, 33, 87, 91, 93, 94, 160
- Gini Index, 53
- global, 45, 75
- GMI Inc., 50
- goodness of fit, 72, 117
- GPSmap 62, 110
- grass, 3
- grassland, 56, 63
- green, 6, 51
- GridION, 9
- ground, 47
- groundwater, 14, 45, 47, 55, 58–60, 70, 89
- Guatemala, 106
- Gulf of Taranto, 106
- gully, 77
  
- haemolysin, 9
- Hamilton, 47, 59, 64
- hay, 3
- hazard, 11, 142, 155–158
  - assessment, 142, 143, 145, 158
- Helicos BioSciences Corporation, 9
- HeliScope, 9
- helminth, 3
- Hicks Road Spring, 55, 58
- HiSeq, 7, 8
- Hoffmann-La Roche, 7
- homogenise, 50
- Hopkirk Research Institute, 48, 49, 110, 138

- 
- horse, 50, 112
  - hospitalisation, 16, 45, 76
  - human, 1–3, 10, 13–15, 20–22, 25, 26, 29–32, 34, 39, 41, 46, 71, 77, 97, 107, 125, 128, 129, 135, 136, 149, 156, 160
  - hydrocarbon, 3
  - hydrochemical, 46
  - hydroxyl, 5
  - hygiene, 15, 45
  - hypervariable region, 35, 139
  - hypothesis, 11, 77, 137, 160
  
  - identify, 45, 56, 69, 70
  - IDEXX Laboratories Inc., 50, 110, 114
  - igneous volcanic soil, 56, 64
  - illness, 2–4, 15, 41, 76, 84, 93–96, 106, 107, 125, 129, 135, 149
  - Illumina, 7, 36, 137–140, 158
  - importance score, 64
  - impurity, 53
  - in-depth, 4, 37, 107
  - incidence, 33, 76, 77, 79
    - rate, 75, 80, 84, 89–91, 94–97
  - incorporate, 5–9
  - incubate, 23, 25, 112, 114, 164
  - independent variable, 46
  - indicator, 1, 21–23, 26, 28, 30, 31, 33, 34, 42, 69, 71, 72, 107, 115, 155–157, 159
  - infection, 15, 16, 32, 33, 41, 45, 71, 97, 106, 125, 149, 157
  - infectious, 95
  - intake, 10, 47, 77, 108, 120, 122, 123, 125, 127, 129, 145, 147
  - interaction, 52, 55, 69, 72
  - interpolation, 81, 84, 95
  - investigate, 10, 46, 69, 75–77, 79, 86, 94, 95
  - Invitrogen Corporation, 50, 138
  - Ion Torrent, 7, 8
  - irrelevant, 64
  - isolate, 21, 23, 47, 49, 50, 63, 110, 112, 114–117, 122, 124, 125, 128–130, 143, 149, 156, 160, 164
  - Italy, 106, 110
  
  - January, 56, 63, 64, 70
  - Japan, 50
  - John Snow, 15
  
  - kriging, 81–84, 91, 94, 95, 97
  - laboratory, 16, 21, 23, 29, 33, 34, 42, 106, 108, 110, 112, 128, 138, 159, 160, 163
  - lag, 81, 84–87, 93–96, 159
  - lake, 16, 77, 105, 108, 121, 145
  - Lake Karapiro, 47, 64
  - Lake Ontario, 76
  - land cover, 52, 55, 58
  - lane, 6
  - Lauda-Königshofen, 50
  - learning data, 53, 54, 56
  - legionellosis, 16
  - level, 56, 64, 69, 71, 72
  - library, 76
  - Life Technologies, 7
  - ligase, 6
  - ligate, 7, 8
  - linear, 117–120, 128
  - link function, 119
  - literature, 10, 31, 36, 37, 76, 137
  - lithology, 55, 58
  - load, 46, 72
  - logarithm, 56, 67, 141
  - London, 15
  - longitudinal, 47, 77
  - Lower Huia Dam, 47, 58
  - Lower Hutt, 47
  - lung, 2
  
  - Macherey-Nagel GmbH & Co. KG, 114
  - machine learning, 52, 53, 167
  - Macs-VA500 microaerophilic workstation, 49, 112
  - magnification, 50, 51
  - magnitude, 72, 80, 84, 94, 97
  - Maine, 50, 110, 114
  - mammal, 40, 123, 145
  - manure, 2, 46
  - map, 4, 52, 58, 82, 84, 95, 108, 136, 160–162
  - March, 46, 47, 56, 58, 64, 70, 71, 105
  - Marienfeld GmbH & Co. KG, 50
  - marshy, 3
  - Massachusetts, 9, 49, 114, 138
  - Massey Genome Service, 114
  - Massey University, 31, 48, 49, 76, 108–110, 138, 143, 166
  - mathematical model, 46
  - matrix, 79, 80, 85–87, 118
  - May, 63, 70, 71
  - measure, 1, 3–5, 9, 10, 15, 16, 22, 24, 33, 41, 42, 45, 53, 55, 71, 80–82, 96, 97, 107, 110, 117, 118, 129, 130, 142, 143, 165
  - Melbourne, 45

- metabolism, 39, 40, 149
- metagenome, 34, 36, 136, 137, 142, 143, 145, 147, 149, 155–158, 160, 166, 167
- metagenomics, 29, 34, 36, 42, 110, 136, 137, 140, 155–158, 167
- methyl-isoborneol, 3
- microaerophilic, 49, 50, 112, 114
- microbe, 3, 13, 25, 29, 34, 40, 42, 43, 45–47, 52, 55–58, 60, 64, 66, 69, 71, 72, 76, 128, 130, 135–137, 147, 156, 157, 159, 160
- microbial
  - burden, 14, 41
  - community, 34–36, 39, 40, 42, 43, 136, 137, 143, 147, 155–158, 160
  - contamination, 3, 4, 14, 29, 30, 45, 58, 71, 72, 107, 128, 137, 159
  - indicator, 46, 107
  - load, 14, 15, 70–72, 96, 97
  - quality, 32, 107, 115, 129
- microbiological, 15, 20, 29, 40, 41, 45, 115, 128–130, 159
- microbiome, 39, 135
- microorganism, 20, 135
- microscope, 114
- Millipore Corporation, 49, 114
- Minnesota, 50
- MiSeq, 7, 8, 138–140, 164
- model, 46, 52, 53, 55–57, 66, 67, 69, 72, 81, 82, 84–87, 96, 97, 117–120, 128, 129, 159
- molecular, 10, 158
- molecule, 7–9
- monitor, 87
- month, 22, 33, 47, 60, 63, 70, 76, 86, 87, 94, 105–107, 138
- morbidity, 45, 75
- mould, 3
- multiple barrier, 3–5, 45
- multiplex, 138, 140, 156, 167
- multivariable, 56, 57
- multivariate, 46, 79, 90, 95, 142, 145, 155
- musty, 2, 3
- mycobacterium, 149
- 
- Nanodrop<sup>®</sup>, 138, 165
- nanopore, 9
- nanoscale, 9
- negative, 20, 23, 112, 123, 135, 157, 158
- NeighborNet, 143, 149, 155
- network, 4
- New Jersey, 46
- New Orleans, 50
- New Zealand, 1, 10, 11, 13, 15–17, 31–33, 36, 37, 40, 41, 45–47, 52, 58, 69–72, 75, 77, 78, 80, 84, 87, 89, 93–97, 105–107, 109, 121, 127–129, 137, 142, 155, 158–160, 163, 165
- node, 53
- non-
  - linear, 119, 159
  - parametric, 46, 52, 69, 72, 82
  - significant, 32, 66, 67, 72, 127
- norovirus, 3
- North Island, 47, 91, 93, 95, 120, 121, 127
- Norway, 106
- November, 56, 63, 76
- nucleotide, 5–9, 140, 145, 165
- 
- objective, 10, 46, 53, 77, 107, 137
- octamer, 7, 8
- October, 60
- odds, 66, 67, 120, 125, 127
- odour, 2, 3
- off-flavour, 2, 3
- Ohio River, 76
- oligonucleotide, 7
- Olympus, 50
- oocyst, 48, 50–52
- organism, 15, 21–23, 28, 30, 31, 33–35, 40–42, 52, 58, 69, 72, 95, 96, 115, 122, 127–129, 135, 136, 141, 149, 151, 155–157, 164
- Orongorongo River, 47
- Oroua River, 58, 59, 70
- outbreak, 15, 16, 31, 32, 45, 76, 106, 107, 129, 130, 160
- outcome variable, 55, 56, 67, 72, 84, 87, 117–120
- outdoor, 105–107, 127, 129, 130
- overdispersion, 86
- overfit, 55
- Oxford, 9
- Oxford Nanopore Technologies, 9
- oxidation, 3
- ozonation, 4
- ozone, 106
- 
- PacBio RS, 9
- Pacific Biosciences, 9
- paired, 140, 141, 143
- Pall Corporation, 138
- Palmerston North, 50
- parameter, 1, 46, 52, 54, 83, 85, 86, 117–119
- parametric, 46, 52, 55, 72

- Passaic River, 46  
 passerine, 120, 123–125, 127, 145  
 pathogen, 1, 3–5, 10, 11, 14–16, 20–22, 26, 30, 31, 33, 34, 36, 41–43, 45, 46, 70, 75, 78, 87, 96, 97, 106, 107, 125, 129, 142, 149, 155–159, 167  
 pathogenic, 3  
 patient, 39, 149  
 pattern, 79, 81, 91, 93–95  
 percentile, 86, 94, 96  
 pesticide, 2  
 phenolic, 3  
 phosphodiester, 5  
 Phred, 141  
 phylogeny, 34, 141, 142, 147, 156  
 physical, 1, 2  
 pig, 30, 156  
 Poisson, 86, 119  
 pollution, 45, 46, 95  
 colonies, 6  
 polyacrylimide, 6  
 polygon, 52  
 polymerase, 5, 6, 8, 9  
 polynomial, 85, 86  
 population, 47, 52, 58, 71, 72, 80, 87, 89, 90, 93, 94  
 Porirua, 47  
 positive, 20, 23, 33, 40, 43, 50, 51, 55, 58–60, 63, 66, 67, 69–72, 112, 114, 120, 122–124, 127, 135  
 possum, 110, 120, 123, 145  
 precipitation, 13, 76, 95  
 predict, 46, 52–56, 63, 64, 72  
 predictor, 54, 55, 64, 71, 72, 117–119, 159  
 presence, 45, 46, 52, 63, 69, 71, 72, 76, 77  
 present study, 77, 82, 84, 86, 94, 129, 137, 155, 157, 158  
 presumptive, 50, 51  
 prevalence, 45  
 preventive, 4, 5, 15, 41, 45  
 previous, 13, 21, 25, 30, 45, 69–72, 75–77, 84, 85, 94, 95, 97, 106, 125, 128, 129, 143, 149, 156, 157  
 primer, 5–9  
 principal component, 79, 80  
 proactive, 45, 46  
 probability, 54, 64  
 proportion, 52, 56  
 prospective, 47, 84  
 protein, 25, 34, 35, 39, 40, 138, 149  
 protocol, 50, 109, 112, 114, 115, 139, 167  
 protozoa, 3, 20–22, 28, 29, 41, 58, 59, 64, 69, 71, 72, 110, 159  
 public health, 1–3, 10, 11, 14, 41, 46, 115, 127–130, 142, 143, 145, 155–158, 167  
 PubMed, 36, 76  
 PubMLST, 112, 124  
 pukeko, 120, 123, 124, 127, 129, 145  
 pump, 106, 110, 155, 163  
  
 Q-score, 141, 145, 155  
 quality, 1, 9, 45, 46, 69  
 Quanti-tray, 114  
 quantify, 77, 84  
 quasi-Poisson, 86, 87  
 Qubit<sup>®</sup>, 138  
  
 R, 53  
 rabbit, 120, 123, 145  
 radiation, 4  
 radioactive, 5  
 rail, 120, 127, 129  
 rainfall, 75–78, 80, 86  
 rancid, 3  
 random effect, 56, 67, 119, 120, 128  
 rank, 75  
 rate, 45, 46, 54  
 raw  
     reads, 140, 141, 145  
     water, 17, 20, 39, 41, 42, 46, 47, 69–73, 95–97, 107, 120, 159, 166  
 reactive, 5  
 reads, 136, 137, 140, 141, 145  
 real-time, 7, 9  
 recreation, 1, 4, 10, 23, 25, 30, 32, 105–107, 115, 120, 121, 129, 130  
 rectal, 2  
 recursive, 53  
 red, 6  
 reference, 1, 13, 29, 31, 46, 108, 110  
 regionalised variable, 80–84  
 regression, 52–54, 56, 72, 73, 79, 117–119, 125, 128, 129  
 regulation, 45, 128, 129  
 relationship, 46, 52, 55, 69, 72, 75–79, 84–86, 90, 94–97, 117, 118, 129, 135, 137, 159  
 replicate, 7, 25  
 research, 16, 31, 32, 36, 37, 39, 42, 105, 128, 137, 157, 162, 166, 167  
 reservoir, 46  
 residual, 83

- residue, 2, 4
- resistance, 40, 43, 143, 155, 157, 158
- resource, 13, 32, 40, 105, 143, 157, 166, 167
- response
  - category, 53
  - value, 53
  - variable, 53
- result, 46, 50, 51, 53, 54, 58, 66, 67, 70, 72
- retrospective, 77, 84
- retrospective longitudinal study, 15
- reversible, 7, 8
- reversible termination, 7
- risk, 1, 2, 4, 5, 10, 11, 22, 32, 33, 41, 42, 46, 107,
  - 109, 127–130, 155, 160
  - assessment, 4, 10, 45, 46, 69, 70, 115, 128
  - factor, 4, 32, 33, 69, 76, 84, 95, 106, 107
- river, 45–47, 52
  - flow, 33, 76–78, 84, 86, 87, 94, 96, 97,
    - 159–161, 165, 167
- riverbed, 52
- RNALater<sup>®</sup>, 138
- rock, 2, 4
- rodent, 120, 123
- roof, 16, 17, 32–34, 77, 89, 90, 95, 97, 108, 110,
  - 122, 138
- rotavirus, 3
- ruminant, 55, 56, 120, 123, 124, 127, 129, 145,
  - 155
- runoff, 2, 13, 14, 46, 76
- rural, 80, 89–91, 95, 97
- rurality, 79, 80, 84, 90
  
- S00009, 47, 64
- S00041, 47, 58, 59, 64, 67, 70
- S00088, 58, 70
- S00092, 47, 58
- S00120, 47
- S00121, 47
- S00124, 67
- S00298, 58, 59, 64, 67, 70
- S00383, 47
- S00434, 47
- S00865, 47, 58–60, 64, 70
- safe, 3, 11
- safety, 4, 10, 40, 41, 45, 70, 96, 107, 159
  - standard, 1, 28
- salmonellosis, 16, 31–33, 75, 87, 89, 91, 93, 94,
  - 160
- saltwater, 1
- Sample, 109
- sample, 16, 23, 24, 26, 29, 33, 36, 47–49, 51, 54,
  - 56, 58–60, 63, 64, 66, 69–71, 106,
    - 108–110, 112, 114, 116, 117, 120,
      - 122–125, 127, 129, 130, 136–143, 145,
        - 149, 156–158, 160, 162–165, 167
    - collection, 15, 47, 60, 109, 110, 127, 138,
      - 162–164, 167
  - Sanger, 5, 6
  - sanitation, 45
  - schistosoma, 14
  - science, 105
  - Scopus, 31, 36, 137
  - Seadon Well, 47
  - season, 34, 46, 55, 56, 60, 63, 64, 70–72, 87, 109,
    - 127, 138, 155, 159
  - second-generation, 5–8, 10
  - selective, 50
  - semivariance, 81, 83
  - September, 46, 47, 56, 58, 63, 71
  - septic, 3
  - sequence, 5–8, 10, 25, 26, 29, 35, 36, 40, 110,
    - 112, 115, 124, 125, 136, 137, 139–143,
      - 145, 149, 155, 160, 161, 165, 166
  - sequencing, 5–10, 23, 34–36, 114, 136–139, 156,
    - 158, 160, 164, 165
      - chemistry, 6
      - generation, 9
      - platform, 7–9
      - technology, 5–7, 9, 10, 36, 137
  - sewage, 2
  - Seward Ltd, 50
  - Shannon Entropy, 53
  - shapefile, 52
  - sheep, 120, 123, 145
  - shigellosis, 16
  - significant, 32, 40, 53, 67, 72, 76, 97, 105
  - site, 47, 52, 55, 58–60, 63, 64, 67, 70, 71, 78, 86,
    - 87, 93, 94, 96, 97
  - skew, 56
  - skin, 2
  - slide, 50
  - snowmelt, 77
  - soil, 2, 4
    - temperature, 55, 56, 64
  - SOLiD, 7, 8
  - Sorvall RT7 Benchtop centrifuge, 50
  - source, 2–5, 10, 11, 14–17, 22, 28–30, 32, 33, 36,
    - 39–42, 45–47, 52, 56, 58, 60, 63, 64, 66,
      - 69–72, 77, 87, 89, 90, 94, 96, 106–110,

- 115, 120–122, 124, 125, 127–130, 135,  
137, 142, 155–158, 160, 163, 165, 167
- South Africa, 149
- South Island, 47, 91, 93, 96, 120, 121, 125, 127
- sparrow, 120, 123
- spatial, 46, 80–84, 91, 93–96, 108, 116, 162, 163
- species, 25, 33–35, 110, 112, 136, 141, 142, 145,  
149, 156
- spectrophotometer, 138, 165
- spline, 86
- spring, 26, 33, 77, 106, 108, 110, 122, 129
- starling, 120, 123
- statistical, 72
- analysis, 69, 128
  - approach, 45, 46
  - method, 69
  - modelling, 10
  - significance, 56, 57, 120
  - technique, 52, 107, 117
- Stomacher<sup>®</sup>, 50
- strand, 5, 7–9
- straw, 3
- stream, 16, 33, 47, 52, 77, 108, 121, 125, 145
- strength, 53, 70, 72, 117, 130, 157, 159
- study, 76, 77, 86, 95, 99, 121
- period, 78, 80, 84, 86, 87, 89, 91, 93, 96, 121
- sub-nanogram, 9
- sulphide, 3
- summary, 5, 10, 40, 114, 115, 130, 138, 142, 145,  
158, 165
- summer, 33, 60, 63, 70, 71, 106, 107, 109, 138,  
155
- supernatant, 50
- supplies, 77, 78, 95, 97
- supply, 3, 10, 11, 15–17, 28, 32, 33, 36, 41, 42,  
106–108, 128, 137, 138, 159, 160, 163
- network, 17, 45
  - system, 2–4, 10, 36, 39, 106, 137, 159, 163
- supply system, 16
- surface, 45, 47, 52, 55, 58, 59, 70, 71, 77, 84, 87,  
90, 91, 94
- surveillance, 10, 78, 94, 96, 97
- swampy, 3
- swan, 120, 123
- Switzerland, 7
- Sydney, 46
- synthesize, 5, 6, 34, 40
- tag, 6, 7
- taint, 2
- Taipei, 46
- Taiwan, 46
- taste, 2
- taxa, 35, 140, 142, 143, 147, 149, 156
- technique, 5, 6, 9–11, 24–26, 32, 36, 39, 42, 43,  
46, 52, 55, 79, 81, 82, 94, 95, 98, 110,  
119, 128, 137, 155, 159, 160, 167
- temperature, 75, 76, 78, 86, 95
- template, 5–10
- temporal, 46, 71, 72
- tephra lapilli soil, 56, 63
- Thermo Fisher Scientific Inc., 138
- thesis, 1, 10, 11, 162, 165–167
- third-generation, 5, 9
- threshold, 46, 66
- Tokyo, 50
- tool, 10, 11, 16, 45, 46, 52, 69, 70, 72, 80, 95,  
141, 155, 167
- Toronto, 76
- Totaranui, 108, 121, 122, 128, 163
- tourist, 105, 106
- training data, 53
- transform, 79–82, 85, 86
- transmission, 15, 75, 157, 158
- transmit, 14, 16, 75
- transport, 48, 50
- treatment, 3, 4, 13–15, 20–22, 28, 29, 97, 108,  
121
- plant, 4, 14, 15, 17, 28, 39, 77, 87, 96, 97
- tree, 53–55
- Tuakau, 47, 58–60, 64
- turbidity, 2, 20, 21, 34, 76, 77, 164
- unbiased, 54, 136, 155
- unhygienic, 106
- univariable, 56, 57
- unpaired, 140, 145
- Upper Hutt, 47
- upstream, 47, 52, 162
- urban, 80, 89–91, 95
- urinary
- bladder, 2
- variable, 52–56, 60, 63, 64, 69–72, 79–82, 84–86,  
89, 95, 120
- importance, 54, 56, 63, 70, 71
- variance, 79, 80, 83, 97
- variation, 46, 67, 72, 79, 80, 84, 90, 91, 93, 94,  
96, 97
- virulence, 40, 143, 155, 156, 158
- virus, 3, 22, 41, 106



- vortex, 50
- Waikato, 16, 33
- Waikato River, 47, 58–60, 64, 70
- Waingawa River, 47
- Wainuiomata River, 47
- Waiorohi Stream, 58, 59, 64, 70
- Waitakere, 16
- Wakaito river, 16
- Wales, 16, 75, 76
- waste, 2
- wastewater, 14, 22, 23, 35, 46
- water, 45
  - quality, 1–3, 10, 32, 43, 46, 47, 50, 69, 114, 115, 120, 127, 128, 155–159, 167
  - source, 45, 47, 55, 58–60, 70–72, 77, 79, 80, 84, 87, 90, 91, 96, 97
  - system, 45
  - treatment, 2, 20, 39–41, 43, 45, 70–72, 120, 121, 127–130, 159, 160, 167
- waterborne, 15, 16, 31, 33, 36, 40–42, 45, 46, 71, 75–77, 94, 97, 106, 107, 160
- Waterborne Inc., 50
- weak, 72
- weather, 13, 76, 80
- Web of Knowledge, 31, 36, 137
- Web of Science, 76
- weka, 120, 123, 127, 129
- well, 45, 50, 70, 77, 90, 95
- Wellington, 47
- West Sussex, 50
- wetland, 56, 64
- Whakarewarewa Forest Spring, 47
- Whangaiterenga, 108, 121, 123
- Whatamango Bay, 108, 122, 129
- winter, 33, 60, 63, 71
- Wisconsin, 139, 164
- worldwide, 13, 15, 16, 40, 105
- yersiniosis, 16
- yield, 79
- zero, 67, 80, 85, 118, 120, 143
- zone, 15, 17, 47, 58, 108, 162