Source Attribution Models using Random Forest for Whole Genome Sequencing Data

A thesis presented in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy in
Statistics

Helen L. Smith

School of Mathematical and Computational Sciences
Massey University, Palmerston North
New Zealand

2025

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

at Massey University, Palmerston North,

New Zealand.

2025

Table of Contents

List of Figures

List of Tables

List of Acronyms

ADHB	Auckland District Health Board
AIR	Actual Impurity Reduction
ARPHS	Auckland Regional Public Health Services
BIGSDB	Bacterial Isolate Genome Sequence Database
BLAST	Basic Local Alignment Search Tool
BURST	Based Upon Related Sequence Types
CA	Correspondence Analysis
CAP	Canonical Analysis of Principal Coordinates
CART	Classification And Regression Tree
CC	Clonal Complex
CCorA	Classical Correspondence Analysis
cgMLST	Core-Genome Multilocus Sequence Typing
cgSNP	Core-Genome Single Nucleotide Polymorphism
COGs	Clusters of Orthologous Genes
DBI	Distribution-Based Imputation
DNA	Deoxyribonucleic Acid
ESR	Institute of Environmental Science and Research
GBS	Guillain-Barré Syndrome
GWAS	Genome-Wide Association Study
HSL	Hue, Saturation, Lightness
HSV	Hue, Saturation, Value
IBS	Irritable Bowel Syndrome
MCMC	Markov Chain Monte Carlo
MCPHS	MidCentral Public Health Services
MDA	Mean Decrease in Accuracy
MDI	Mean Decrease in Impurity

mEpiLab	Molecular Epidemiology and Public Health Laboratory
ML	Machine Learning
MLST	Multilocus Sequence Typing
MMD	Minimal Multilocus Distance
MoH	Ministry of Health
MPI	Ministry for Primary Industries
NGS	Next-Generation Sequencing
nMDS	Non-Metric Multidimensional Scaling
NZFSSRC	New Zealand Food Safety Science and Research Centre
OOB	Out-Of-Bag
PCA	Principal Component Analysis
PCO	Principal Coordinates Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
RF	Random Forest
rMLST	Ribosomal Multilocus Sequence Typing
RNA	Ribonucleic Acid
SACNZ	Source Assigned Campylobacteriosis in New Zealand
SNP	Single Nucleotide Polymorphism
sRGB	Standard Red, Green, Blue
ST	Sequence Type
tRNA	Transfer RNA
wgMLST	Whole-Genome Multilocus Sequence Typing
WGS	Whole Genome Sequencing

Abstract

Foodborne diseases, such as campylobacteriosis, represent a significant risk to public health. Preventing the spread of *Campylobacter* species requires knowledge of sources of human infection. Current methods of source attribution are designed to be used with a small number of genes, such as the seven housekeeping genes of the original multilocus sequence typing (MLST) scheme, and encounter issues when presented with whole genome data. Higher resolution data, however, offers the potential to differentiate within source groups (i.e., between different ruminant species in addition to differentiating between ruminants and poultry), which is poorly achieved with current methods.

Random forest is a tree-based machine learning algorithm which is suitable for analysing data sets with large numbers of predictor variables, such as whole genome sequencing data. A known issue with tree-based predictive models occurs when new levels of a variable are present in an observation for prediction which were not present in the set of observations with which the model was trained. This is almost certain to occur with genomic data, which has a potentially ever-growing set of alleles for any single gene.

This thesis investigates the use of ordinal encoding categorical variables to address the 'absent levels' problem in random forest models. Firstly, a method of encoding is adapted, based on correspondence analysis (CA) of a class by level contingency table, to be unbiased in the presence of absent levels. Secondly, a new method of encoding is introduced which utilises a set of supplementary information on the category levels themselves (i.e., the sequence information of alleles) and encodes them, as well as any new levels, according to their similarity or dissimilarity to each other via the method of principal coordinates analysis (PCO). Thirdly, based on the method of canonical analysis of principal coordinates (CAP), the encoding information of the levels from the CA on the contingency table is combined with the encoding information of the levels from the PCO on the dissimilarity matrix of the supplementary levels information, with a classical correspondence analysis (CCorA). Potential issues when using out-of-bag (OOB) data following variable encoding are then explored and an adaptation to the holdout variable importance method is introduced which is suitable for use with all methods of encoding.

This thesis finishes by applying the CAP method of encoding to a random forest predictive model for source attribution of whole genome sequencing data from the Source Assigned Campylobacteriosis in New Zealand (SACNZ) study. The advantage of adding core genes and accessory genes as predictor variables is investigated, and the attribution results are compared to the results from a previously published study which used the asymmetric island model on the same set of isolates and the seven MLST genes.

Acknowledgements

This thesis is the joint achievement of my supervisory team, my family, and somewhat myself. My four supervisors are a team of exceptional people, both academically and personally, and they have kept me smiling and mostly sane over the last five years.

My foremost thanks goes to my primary supervisor Associate Professor Jonathan Marshall. I have some experience in what makes a good supervisor, and I can say without a doubt that you are one of the best. Thank you for painstakingly mentoring me in the language of R and keeping me grounded throughout the whole project. Thank you for knowing the answer to every question I ever asked, you are my hero. My favourite quote of yours is "I think you should do as little as possible" which, although you don't yourself follow, did help to nudge me through the procrastinating!

Distinguished Professor Nigel French, thank you for saying yes when I turned up at your door all those years ago asking if I could do a PhD despite having no topic and no funding. You believed me that I wanted to do it and that I could do it. On top of that, you are one of the funniest people on this planet and your sense of humour has had me laughing well after our meetings many a week. And of course, thank you for sharing your wealth of knowledge on *Campylobacter* and helping me fit the pieces of the puzzle together.

Professor Patrick Biggs your unwavering support has been truly appreciated. Thank you for wrangling giant sets of genes for me and then explaining what they mean, and then explaining it again (and again). You have always been able to see each piece of work from a unique perspective and the result is much more widely understandable as a result. Thank you also for bringing your family to visit me. Working remotely means my two worlds rarely interact, and I loved it when they did.

Doctor Adam Smith, thank you for being my friend ☺. And thank you for massively improving my writing, my grammar, and, my, overuse of, commas. Thank you also for generously helping me expand my broader statistical knowledge to help prepare me for the scary outside world. I am genuinely lucky to have had such a great supervisory team.

I also gratefully acknowledge the financial support I received from the School of Fundamental Sciences (now the School of Mathematical and Computational Sciences, Massey University), and the New Zealand Food Safety Science and Research Centre (NZFSSRC), without which I could not have completed this work. I also received a travel scholarship from the New Zealand Statistical Association and a Massey University COVID-19 Doctoral Student Bursary.

The bulk of the data used in this study is from the Source Assigned Campylobacteriosis in

New Zealand (SACNZ) study. This was a large study involving many individuals, researchers, and organisations. I am extremely grateful to have had the use of this data and I am a little embarrassed that I didn't even need to step foot inside a lab to receive it. I would like to extend my gratitude to the individuals who contributed the data in this study, and to acknowledge the contributions of the organisations that were involved in the collection of the data, the sequencing of the genomes, the cleaning and analysis of the data, and the advisory groups behind the scenes; in particular the following organisations:

Public Health Services

- Auckland Regional Public Health Services (ARPHS)
- MidCentral Public Health Services (MCPHS)

Diagnostic laboratories

- Labtests
- North Shore Hospital
- Middlemore Hospital
- Auckland Hospital – LabPlus
- MedLab Central

Source Sampling

- Meat Industry Association
- Poultry Industry Association of New Zealand
- Cattle plants
  - Alliance Levin
  - Silver Fern Farms Pacific
  - Landmeats Whanganui
  - ANZCO Foods Eltham
- Sheep plants
  - Alliance Dannevirke
  - Silver Fern Farms Takapau
  - ANZCO Foods Rangitikei
- Poultry companies
  - Tegel
  - Inghams
  - Turks
  - Brinks

AsureQuality Limited

Eurofins Food Analytics NZ Limited

Ministry of Health (MoH)

CBG Health Research Ltd

UMR Market Research

Institute of Environmental Science and Research (ESR): John Gray, Kristin Thom, Sally Ladbrook, Ashley Orton, Paula Scholes, Andrew Crooke, Mehnaz Adnan, Ben Waite, Tina von Pein

Ministry for Primary Industries (MPI): Elaine Taylor

Massey University: Dr. Ahmed Fayaz

Molecular Epidemiology and Public Health Laboratory (mEpiLab), Massey University: Lynn Rogers, Rukhshana Akhter

Advisory Groups

SACNZ Advisory Group (May 2016 – September 2017)

Prof Steve Hathaway (MPI), Dr Donald Campbell (MPI), Dr Rob Lake (ESR), Prof Michael Baker (Otago University), Dr Phil Shoemack (Toi Te Ora Public Health), Dr Tammy Hambling (ESR) Dr Lisa Oakley (MoH), Prof Nigel French (Massey University), Elaine Taylor (MPI)

Working Groups (June 2016 – September 2017)

Working Group 1 (Study design) participants: Peter Cressey (Chair, ESR), Dr Donald Campbell (MPI), Dr Jill Sherwood (ESR) Prof Michael Baker (Otago University), Prof Nigel French (Massey University), Dr Penelope Neave (Auckland District Health Board (ADHB)), Dr Phil Shoemack (Toi Te Ora Public Health) Dr Tomasz Kiedrzynski (MoH), Dr Annette Bolton (ESR), Dr Graham Mackereth (ESR), Dr Lisa Lopez (ESR)

Working Group 2 (Sampling) participants: Dr Rob Lake (Chair, ESR), Dr Anne Midwinter (Massey University), Dr Jill Sherwood (ESR), Dr Paula Scholes (ESR), Dr Chris Hewison (ESR), Dr Andrea McNeill (MoH), Dr Anne Marie Perchec Merien (MPI)

Working Group 3 (Genomics and modelling) participants: Prof Nigel French (Chair, Massey University), Dr Brent Gilpin (ESR) Dr Claire McDonald (ESR) Dr David Wilkinson (Massey University) Dr Claire Newbern (ESR) Dr Jonathan Marshall (Massey University)

Finally, I would like to thank my wonderful family. Isla, Felix and Elise, you have never asked me why I was doing a PhD or what I will do with it, you just knew it was important. Thank you for hanging out with me in my office on the days you didn't go to school. It was fun having you help me choose colours for plots – some of which made it into this thesis. Thank you for decorating my walls with your beautiful artwork, and thank you for understanding when I couldn't come and watch your cross-country or Kapa Haka or Christmas sing-a-long. Hopefully we can spend some more time together now ♥. Aidan, you have been by my side for the better part of my life, and as for most things, I couldn't have started, or completed, this thesis without your support. Thank you for bearing the load with me. Nana Jane and the wider circle of family and friends, thank you for helping with the kids so I could get some hours behind the computer, thank you for waiting in the playground when I was late for school pick ups, and thank you for

encouraging me to keep going to the end.

And it has taken a while! This thesis has seen the life of one old dog, one puppy, several lambs, oceans of zucchinis, two pre-schools, two *au pair*, a few photography awards, and many overdue library notices ... and now it is time to stop and smell the roses.

Best decorated office ever!

Publications Arising From Thesis

Presented in Chapter 4

Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N.H., and Marshall, J.C., Lost in the Forest: Encoding Categorical Variables and the Absent Levels Problem, *Data Mining and Knowledge Discovery* 38, 1889–1908 (2024), DOI 10.1007/s10618-024-01019-w.

In addition, the following publication by Singh et al. (2025) uses the PCO method from Chapter 4: Singh, N., Thystrup, C., Hassen, B.M., Bhandari, M., Rajashekara, H., Hald, T., Manary, M.J., McKune, S.L., Hassen, J.Y., Smith, H.L., Marshall, J.M., French, N., and Havelaar, A.H., Transmission pathways of *Campylobacter jejuni* between humans and livestock in rural Ethiopia are highly complex and interdependent, *Gut Pathogens* 17:26, 1-16 (2005), DOI 10.1186/s13099-025-00691-7.

Presented in Chapter 5

Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N.H., and Marshall, J.C., CAP-encoding: Encoding Categorical Variables using Canonical Analysis of Principal Coordinates. *Machine Learning* (2025). In review.

Presented in Chapter 6

Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N.H., and Marshall, J.C., Out of (the) bag - encoding categorical predictors impacts out-of-bag samples. *PeerJ Computer Science* (2024), 10:e2445, DOI 10.7717/peerj-cs.2445.

Chapter 1

Introduction

1.1 Rationale and Importance of this Research

Campylobacteriosis is gastroenteritis caused by infection of humans with species of the *Campylobacter* genus. Campylobacteriosis is a major public-health problem world-wide. In New Zealand, campylobacteriosis is the highest occurring notifiable disease. In 2023, 6089 cases were notified (116.6 cases per 100,000 people), accounting for 43.8% of disease notifications (ESR, 2023) (figure 1.1), and the true incidence is likely to be much higher (Cressey and Lake, 2012; Pascoe et al., 2024). As a result, campylobacteriosis is the largest contributor to the economic costs of foodborne diseases in New Zealand (Scott et al., 2000).

Most *Campylobacter* infections in developed countries are due to *C. jejuni* (approximately 90% of cases) followed by *C. coli* (approximately 10% of cases) (Cody et al., 2012; Nohra et al., 2016; Wong et al., 2007). Infection typically results in acute gastroenteritis with associated fever, abdominal cramping, diarrhoea, vomiting and headaches (Allos, 2001; Blaser, 1997; Epps et al., 2013). Most infections are self-limiting with symptoms clearing within a week; however, infection with *C. jejuni* is recognised as an important contributor to more severe conditions, including Irritable Bowel Syndrome (IBS) (Zilbauer et al., 2008) and Guillain Barré Syndrome (GBS) (Kuwabara, 2007; Scallan Walter et al., 2020).

Campylobacter species are ubiquitous in the environment and form part of the commensal microbiota of numerous vertebrate host species, including farmed animals. The principal source of human infection is chickens (Nohra et al., 2020; Pascoe et al., 2024), although, due to the ubiquity of *C. jejuni* and *C. coli* in animal intestines, and its shedding into the environment, transmission can occur via numerous routes including contaminated water (Gilpin et al., 2020), consumption of contaminated milk (Fernandes et al., 2015) and meat (Cody et al., 2019; Rivas et al., 2021), and less frequently by contact with infected animals (Newell et al., 2017; On et al., 2019; Thépault et al., 2020) and people (Same and Tamma, 2018).

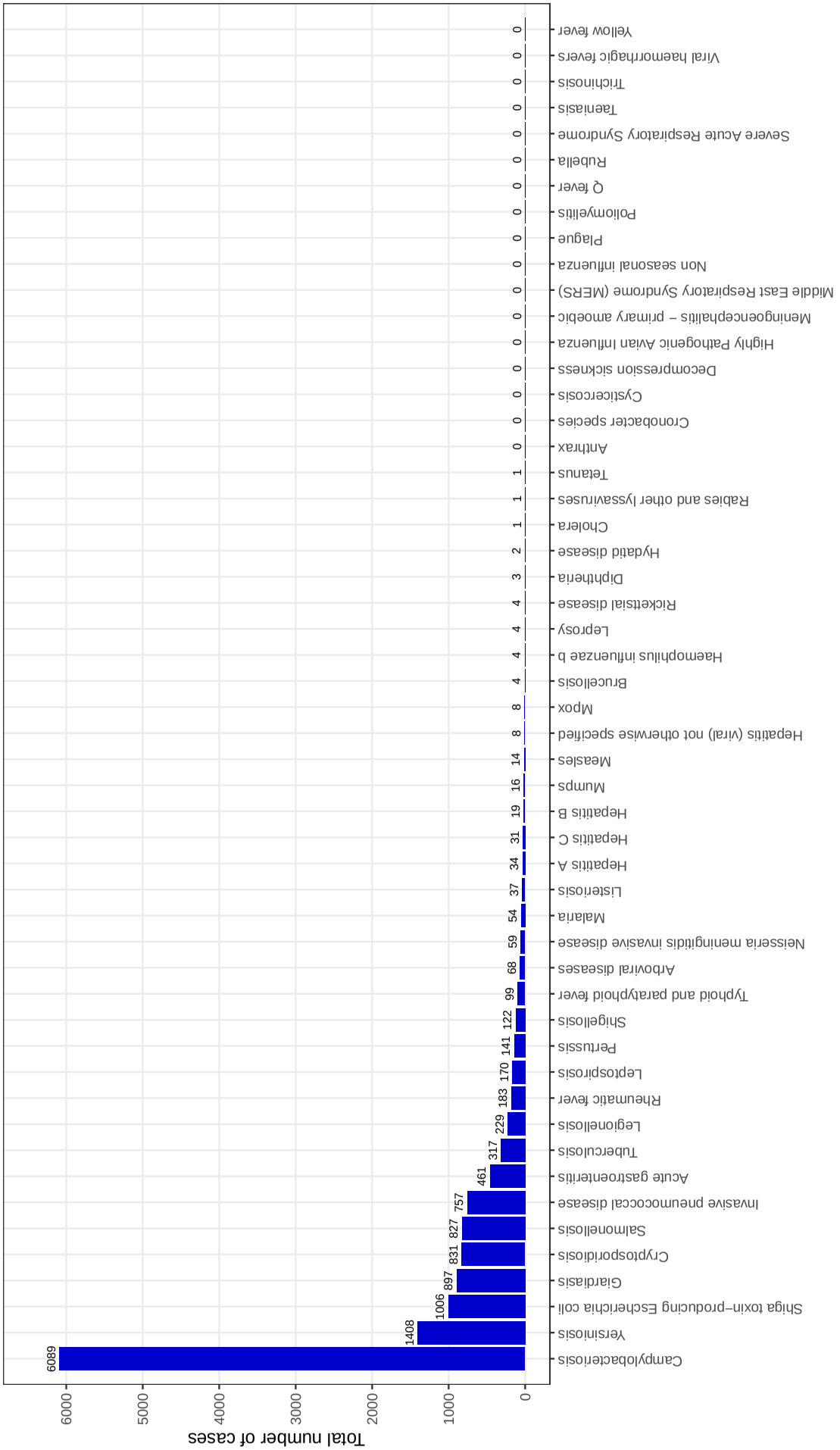

Figure 1.1: The incidence of notifiable diseases in New Zealand in 2023 (ESR, 2023).

Quantifying the relative contribution of different sources of *Campylobacter* infection is necessary to prioritise food-safety interventions and their efficacy in terms of reducing infections in humans (Mullner et al., 2010; Sears et al., 2011). The process of assigning cases of human disease to their most likely origin is known as source attribution. Multiple approaches to source attribution have been developed, including epidemiological methods (Domingues et al., 2012; Pires et al., 2010), comparative risk and exposure assessment (Pintar et al., 2017), expert knowledge elicitation (Hald et al., 2016; Havelaar et al., 2008), and microbiological methods (Hald et al., 2004; Liao et al., 2019; Miller et al., 2017; Mullner et al., 2009a,b; Sheppard et al., 2009; Strachan et al., 2009).

Campylobacter populations are highly structured with groups of clonally related lineages often sharing phenotypic properties, such as association with host species (Sheppard et al., 2010a,b). Current source attribution methods show greater discriminatory power between major groups of sources (e.g., poultry *versus* ruminants) than within these major groups (e.g., cattle *versus* sheep within ruminants). These methods generally use relatively low-resolution genomic data, such as the 7-loci multilocus sequence typing (7-gene MLST) scheme (Dingle et al., 2001). Utilising the higher resolution offered by whole genome sequencing (WGS) may increase the genomic differentiation within source group which is expected to better discriminate between sources. Existing statistical models, however, are poorly scalable to the whole genome level (Sheppard et al., 2012).

Random forest is a method of supervised machine learning that creates an ensemble of decision trees and uses bagging (bootstrap aggregating) and random subspacing to prevent overfitting. Random forest is well suited to sets of data with a high number of variables and/or high numbers of variable levels, such as genomic data (Auti et al., 2023; Chen and Ishwaran, 2012; Montesinos López et al., 2022). However, due to the variable nature of alleles and the continual evolution of new genetic variants, genomic data is subject to the absent-levels problem (Au, 2018). Absent levels are levels which are present in new observations for prediction but which were not present in the observations used to train the model. There is currently no established solution for dealing with absent levels in random forest models (Au, 2018).[1]

1.2 Research Objectives

This thesis approaches the problem of source attribution of *Campylobacter jejuni* and *Campylobacter coli* at the whole genome level using random forest. It initially evaluates the effect of absent levels on random forest predictive models using data from the Source Assigned Campylobacteriosis in New Zealand (SACNZ) study (Lake et al., 2021). It then develops new methods for encoding categorical predictor variables which are unbiased in the presence of absent levels. The effect of encoding on calculations made using out-of-bag (OOB) samples, such as OOB error and measures of variable importance is investigated. The new methods of encoding categorical variables are then applied to genome sequence data, incorporating different numbers

[1] https://github.com/imbs-hl/ranger/issues/94

of genes utilising the multilocus sequence typing (MLST) schemes (e.g., 7-gene MLST *versus* core-genome MLST (cgMLST) *versus* whole-genome MLST (wgMLST)) in order to attribute the human infections of *C. jejuni* and *C. coli* to potential animal sources using random forest. Finally, the new methods are compared with results from the asymmetric island model, on the same data, published by Lake et al. (2021).

1.3 Thesis Structure

The remainder of this thesis is organised as follows:

In **Chapter 2**, the microbiology of the *Campylobacter* genus is described and the international *Campylobacter* typing scheme is introduced. Current methods for source attribution are reviewed and their limitations are outlined. The random forest predictive model is presented for classification and the issue of absent levels is raised. Finally, the concept of dissimilarity measures for distance-based methods of analysis is discussed, and using an illustrative example, the importance of selecting an appropriate measure is shown.

In **Chapter 3**, based on data from the SACNZ study (Lake et al., 2021) the diversity of *Campylobacter* species in New Zealand is explored, including the extent of new alleles in observations for prediction (i.e., absent levels, as described in the chapter 2).

In **Chapter 4**, using the data presented in chapter 3, the effect that absent levels have on random forest predictions for source attribution of *Campylobacter* species using WGS data as predictors is evaluated. Two new methods for encoding categorical predictors are introduced: (i) a target-based encoding approach which uses class probability information and encodes absent levels according to the *a priori* hypothesis of equal class probability (the CA-unbiased-encoding method), and (ii) a target-agnostic encoding approach which encodes absent levels according to their similarity to each of the other levels in the training data (the PCO-encoding method).

Chapter 4 has been published with *Data Mining and Knowledge Discovery* – "Lost in the Forest: Encoding Categorical Variables and the Absent Levels Problem" by H.L. Smith, P.J. Biggs, N.P. French, A.N.H. Smith, and J.C. Marshall (2024), *Data Mining and Knowledge Discovery* 38, 1889–1908, DOI 10.1007/s10618-024-01019-w.

In **Chapter 5**, a new method for encoding categorical predictor variables is introduced which utilises the method of canonical analysis of principal coordinates (CAP) (the CAP-encoding method). This method combines the advantages of the target-based CA-unbiased-encoding method and the target-agnostic PCO-encoding method which were introduced in chapter 4. The concept of the CAP-encoding method is illustrated using a simulation study and then the method is applied to three real-world datasets, including a subset[2] of the SACNZ data and two non-genomic datasets.

Chapter 5 is under review with *Machine Learning* – "CAP-encoding: Encoding Categorical Variables using Canonical Analysis of Principal Coordinates" by H.L. Smith, P.J. Biggs, N.P. French, A.N.H. Smith, and J.C. Marshall (2025).

[2]the *C. jejuni* isolates

Chapter 6, investigates how target-based *versus* target-agnostic encoding of categorical predictor variables for random forest can bias performance measures based on out-of-bag (OOB) samples, such as variable importance and misclassification rates. The new 'independent holdout method' for calculating variable importance is introduced and evaluated with different methods of encoding.

Chapter 6 has been published with *PeerJ Computer Science* – "Out of (the) bag - encoding categorical predictors impacts out-of-bag samples" by H.L. Smith, P.J. Biggs, N.P. French, A.N.H. Smith, and J.C. Marshall (2024), *PeerJ Computer Science* 10:e2445, DOI 10.7717/peerj-cs.2445.

In **Chapter 7**, the CAP-encoding method which was presented chapter 5 is used for the source attribution of *Campylobacter* species in New Zealand. The effect of using the full genome (wgMLST data) compared to the 7-gene MLST data and the cgMLST data is evaluated, and the results are compared against those of the asymmetric island model on 7-gene MLST data (Lake et al., 2021). The difference in performance when the dissimilarity matrix is adjusted to account for clonal structure and genetic recombination is also investigated. Lastly, the variable importance measure introduced in chapter 6 is applied to identify the most important genes with regards to source attribution.

Finally, in **Chapter 8**, the General Discussion, the findings of this thesis are summarised and possible directions for future research are discussed.

Chapter 2

Literature Review

2.1 Introduction

Quantitative estimates of the relative contributions of different sources to human campylobacteriosis are essential for the identification of transmission pathways, as well as the development and monitoring of food-safety interventions. Source attribution modelling relies on comparing the genomic profiles of human cases of *Campylobacter* with those of putative sources. To develop suitable models, it is important to understand how genomic variation occurs, starting with microbiological processes. In this review of the literature, first, an overview of the microbiology of *Campylobacter* is provided. Secondly, the concept of source attribution is explained and the array of known microbiological source attribution methods are detailed along with their strengths and limitations. The concepts of the random forest machine learning algorithm and the issue with absent levels are further described. Finally, the use of distance-based methods for multivariate analysis is discussed.

2.2 Microbiology

2.2.1 What is a Gene?

A gene is a sequence of nucleic acids (deoxyribonucleic acid (DNA) or ribonucleic Acid (RNA)) that code for a functional product, like a protein, that contribute to the expression of a particular trait or a particular function in a cell. DNA is a long polymer of nucleotides. Each nucleotide consists of a monosaccharide sugar (deoxyribose), a phosphate group, and one of four nitrogen bases (adenine (A), guanine (G), cytosine (C), or thymine (T)).

The nitrogen bases of one DNA strand pair with the bases of a second strand (of reverse orientation) and the two strands twist helically to form a double helix. The base pairs are fixed - the pyrimidine cytosine always pairs with the purine guanine (with three hydrogen bonds) and the pyrimidine thymine always pairs with the purine adenine (with two hydrogen bonds). To fit the large polymer into a small cell, DNA is packaged up with proteins and supercoiled into chromosomes (figure 2.1).

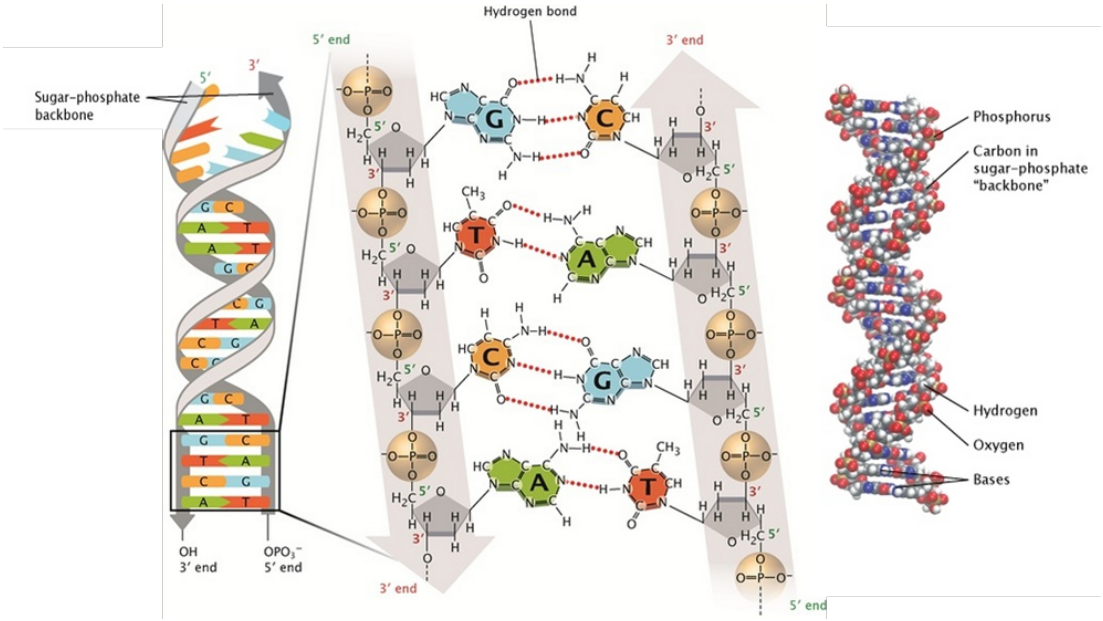

Figure 2.1: The Structure of DNA.[1]

2.2.2 The Genetic Code

Three adjacent nucleotides are known as a codon. Each codon codes for a specific amino acid. In this way, the sequence of nucleotides specifies a sequence of amino acids and therefore the structure of a protein, and is referred to as the genetic code. There are four nucleotide bases and 64 (4^3) possible combinations (codons). As there are only 20 amino acids in the standard genetic code, some amino acids may be represented by more than one codon (i.e., there is degeneracy in the genetic code).

2.2.3 Bacteria

Bacteria are single-celled microscopic organisms that lack a nucleus and are classified as prokaryotes. Bacterial DNA is generally a single, circular DNA strand (chromosome), supercoiled with proteins, and floating freely in the cell in a region termed the nucleoid (figure 2.2). Bacteria may also carry extrachromosomal genetic material in the form of plasmids, which are smaller than chromosomes and replicate independently of the rest of the genome. Plasmids encode few genes which, although are nonessential, often benefit the survival of the organism and confer selective advantage to the bacteria through mechanisms such as antibiotic resistance. Plasmids also provide a mechanism for horizontal gene transfer (bacteria can pick up new plasmids from other bacterial cells during conjugation, or from the environment), and are a source of bacteriophages (viruses that infect bacteria) (Harrison et al., 2015).

[1] Image adapted from Pray (2008)

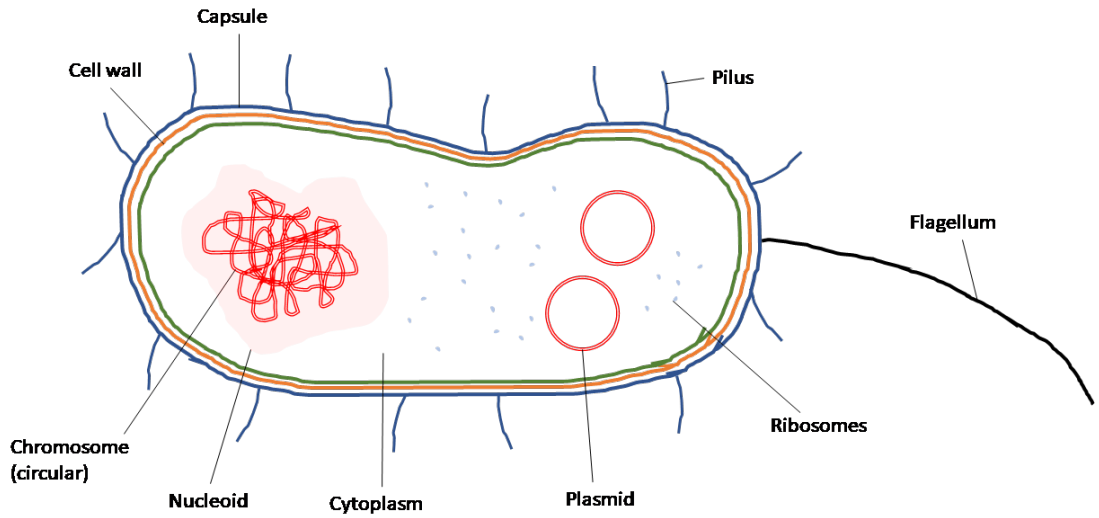

Figure 2.2: Simple bacterial cell structure.

2.3 *Campylobacteraceae*

The *Campylobacter* genus belongs to the family *Campylobacteraceae* and contains 66 species (LPSN, 2024), with new species likely to be discovered with time (Bian et al., 2020; Silva et al., 2020). Most *Campylobacter* species are gram-negative, non-spore forming, spiral-shaped, motile bacteria with a polar flagellum at one or both ends of their cell (Nachamkin et al., 2008) (figure 2.3). Two species, *C. jejuni* and *C. coli*, are responsible for most cases of campylobacterosis in humans and are the subject of this review.

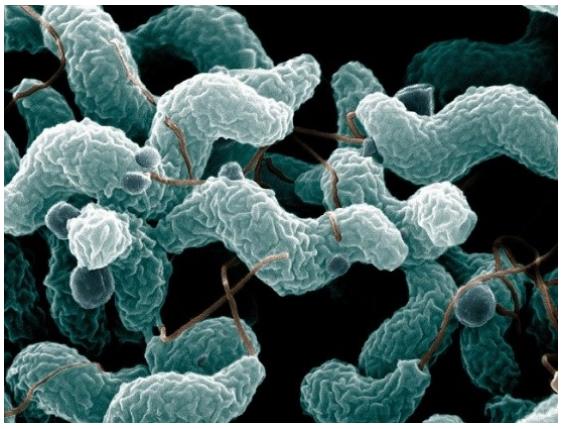

Figure 2.3: *Campylobacter jejuni*.[2]

C. jejuni and *C. coli* each have a circular chromosome roughly 1.7 Mb long (Chen et al., 2013; Parkhill et al., 2000; Pearson et al., 2013; Taylor et al., 1992) which encodes for approximately 1700 genes (Parkhill et al., 2000). The genomes of the two species are similar, matching by around 85% at the nucleotide level (Sheppard et al., 2013a). These *Campylobacter* species

[2]Image sourced from `https://www2.le.ac.uk/projects/vgec/schoolsandcolleges/Microbial%20Sciences/bacteria-passport/campylobacter-jejuni-1`

have high allelic diversity (i.e., variability within a gene); however, this variability is not consistent across the genome – some loci show very little variation while others are highly variable (Parkhill et al., 2000; Sheppard and Maiden, 2015). Variation in the nucleotide sequence of a gene is either the result of a mutation event (i.e., a substitution, insertion, or deletion of one or few nucleotides) or of a recombination event (i.e., a restructuring of a larger part of the genome using external DNA) (Sheppard and Maiden, 2015). In addition, new genetic material can be introduced via horizontal transfer from other bacteria ('conjugation'), intracellular DNA remnants ('transformation'), or bacteriophages ('transduction') (Fearnhead et al., 2014) and can generate mosaic genes, different parts of which have different evolutionary histories (Maiden et al., 2013). The genetic diversity of *Campylobacter* is driven by high levels of this horizontal genetic exchange (Wilson et al., 2009), yet the populations retain a distinct structure corresponding to ancestral lineages (Sheppard and Maiden, 2015). In other words, *Campylobacter* species show evidence of both clonal relatedness through substitutions and panmixis through frequent recombination events.

Genetic Sequencing

DNA sequencing is the extraction of the exact order of nucleotide bases in a strand of DNA. Sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e., clusters of co-regulated genes (operons)), full chromosomes, and entire genomes. Historically, there have been three milestones of genetic sequencing methods: Sanger sequencing, next-generation sequencing (NGS, also known as high-throughput sequencing), and third-generation sequencing (also known as long-read sequencing). NGS is frequently used due to its affordability, speed, and accuracy. NGS requires breaking long strands of DNA into small segments randomly, then using Polymerase Chain Reaction (PCR) amplification and cloning to generate a pool of small DNA fragments. The fragments are then sequenced from either end to create a library of sequenced regions, called 'reads'. The individual reads are then computationally merged into longer stretches of overlapping sequences called 'contigs', which in turn are joined to form even longer stretches of sequences called 'scaffolds'. Finally, the aligned scaffolds are matched against a reference genome (e.g., by BLAST (Basic Local Alignment Search Tool)) to identify the individual genes and their locations on the chromosome (figure 2.4).

2.3.1 *Campylobacter* Typing Schemes

The ability to accurately identify the strains of infectious agents that cause disease is central to the investigation of epidemiology and infections, surveillance, and public health decisions (Maiden et al., 1998). An isolate of *Campylobacter* is a single colony isolated from agar plates where microorganisms are cultured. Typing schemes for *C. jejuni* and *C. coli* help decipher the relationships between disease-associated isolates and isolates sourced from animal or environmental populations (Dingle et al., 2001, 2005). The ultimate goal of a typing scheme is to assign isolates to molecular types that each share a recent common ancestor (Maiden et al., 1998). Although historically there have been many schemes for discriminating *Campylobacter*

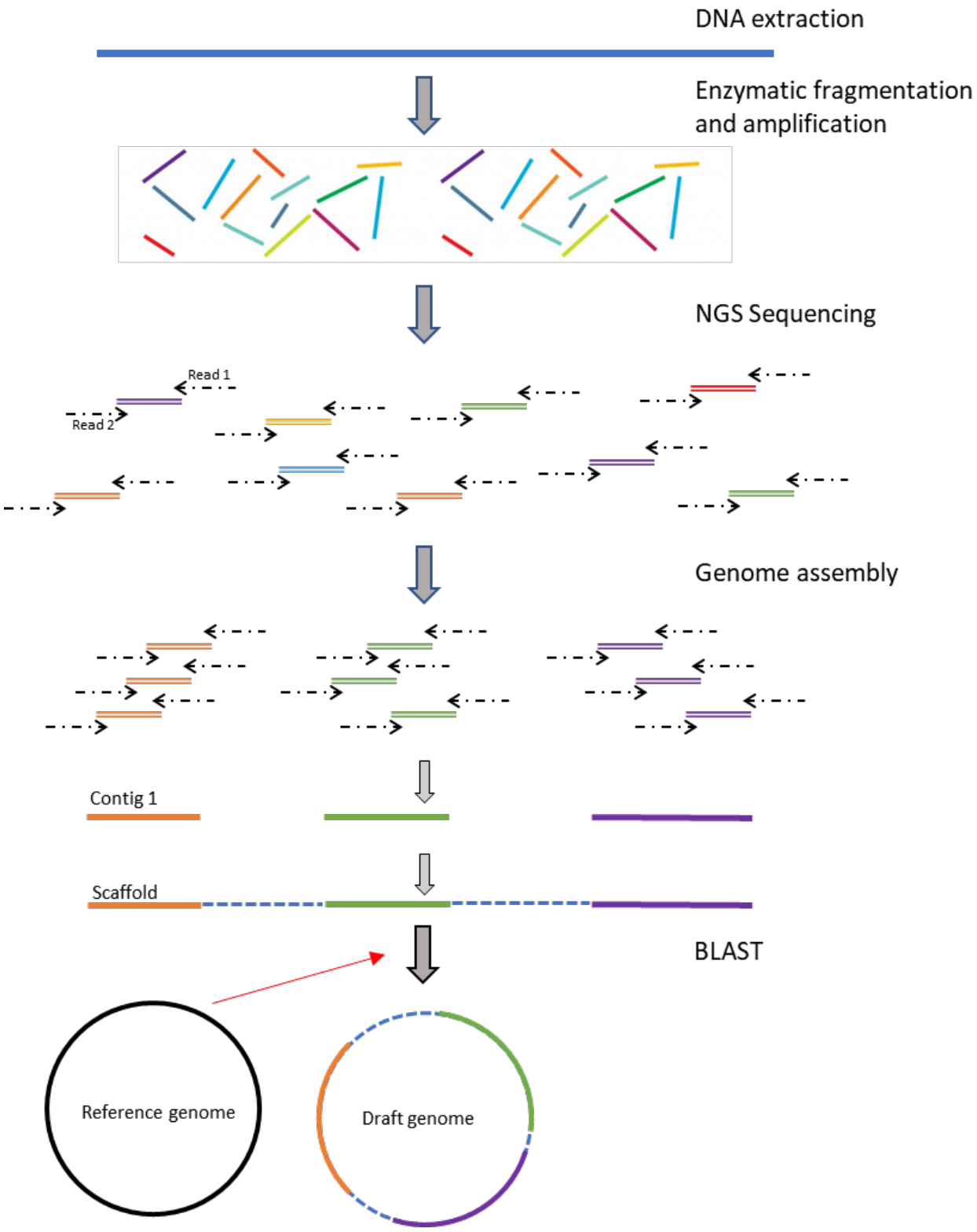

Figure 2.4: Genome sequencing process.

isolates (described in Guzmán-Martín et al. (2019)), the broadly adopted approach is the nucleotide sequence-based typing method known as multilocus sequence typing (MLST). MLST has the advantage of being able to overcome some of the difficulties of interpreting bacterial population structures that display elements of both clonality and panmixis.

MLST uses nucleotide sequencing to directly assign variants of genes, known as alleles, which reside at a particular locus to a molecular type. A change in nucleotide sequence may or may not influence the functioning of the gene. A mutation that changes the sequence of a triplet codon can be: (i) synonymous (the new codon specifies the same amino acid as the original codon), (ii) non-synonymous (the new codon specifies a different amino acid), (iii) nonsense (the new codon specifies a termination codon), or (iv) readthrough (the new codon converts a termination codon into one specifying an amino acid). In *Campylobacter*, recombination events occur with a much higher frequency than point mutations (Yu et al., 2012) and have a greater impact on the total similarity of sequences between strains. In MLST, the number of different nucleotides between alleles is ignored. Each unique sequence for each locus is assigned a unique allele number (incrementing with the time of discovery) whether the sequences differ at a single nucleotide position or at many positions. Thus, the allele is the unit of comparison rather than the nucleotide sequence. Each allelic change is, therefore, counted as a single genetic event, regardless of the nature of the change. MLST is a portable system with a central set of databases hosting MLST allele sequences and sequence type (ST) profile tables for over 100 different microbial species and genera (Jolley et al., 2018).

C. jejuni and *C. coli* share a common MLST scheme (Dingle et al., 2005). The original MLST scheme only looked at the nucleotide sequences of internal fragments of seven housekeeping genes accounting for approximately 0.1% of the genomic sequence (Dingle et al., 2001). The corresponding allelic profile from the seven loci define the ST of each isolate and subsequently the clonal complex (CC) or lineage. MLST can now be assigned from whole-genome sequence information and applied gene-by-gene (Jolley and Maiden, 2010; Sheppard et al., 2012). A more recent and comprehensive MLST scheme comprises the core genome (cgMLST), defined as a set of 1343 loci which are present in most members of *C. jejuni* and *C. coli* (Cody et al., 2017). The defined core genome has been validated against isolates from the United Kingdom, Europe, and North America, but may not be wholly transferrable to other continents or time periods. The highest level of resolution is whole-genome MLST (wgMLST) (Cody et al., 2013; Maiden et al., 2013), in which all the loci of a given isolate are compared to equivalent loci in other isolates. The whole-genome resolution is particularly suited to studies of closely related strains that have subtle genetic differences. Lower resolution is suited to studies of strains of less closely related strains that have greater genetic differences (Maiden et al., 2013).

2.4 Source Attribution

The process of assigning cases of human disease to their most likely origin is known as ‘source attribution’. Many approaches to source attribution have been developed, including epidemiological methods (Domingues et al., 2012; Pires et al., 2010), expert knowledge elicitation (Hald

et al., 2016; Havelaar et al., 2008; Mughini-Gras et al., 2022), and microbiological methods (Falcao et al., 2024; Sheppard et al., 2009; Strachan et al., 2009), among others. Microbiological methods of source attribution encompass comparative risk and exposure assessment (Pintar et al., 2017) and microbial subtyping methods (Hald et al., 2004; Liao et al., 2019; Miller et al., 2017; Mullner et al., 2009a; Pritchard et al., 2000; van Pelt et al., 1999; Wilson et al., 2008).

2.4.1 Frequency-Matching Methods

The microbial subtyping approach involves differentiating bacterial isolates of the same species by phenotypic and genotypic subtyping methods. Frequency-matching methods of source attribution then infer probabilistically the most likely sources of human cases by comparing the relative frequency of subtypes from each potential source with subtypes from human cases. Bayesian inference is often applied to link the models derived from human occurrences with the models derived from animal occurrences (tables 2.1, 2.2; figure 2.5).

The models fit to counts of occurrences in humans can be broadly categorised as either Poisson ($\boldsymbol{Y} \sim \mathsf{Poisson}(\boldsymbol{\lambda})$; e.g., Hald model and its modifications) or Multinomial ($\mathbf{Y} \sim \mathsf{Multinomial}(n, \mathbf{p})$; e.g., Dutch model, asymmetric island model, and Dirichlet model). The models fit to counts of occurrences in animals typically take the form of a multinomial model ($\boldsymbol{X}_k \sim \mathsf{Multinomial}(n_k, \boldsymbol{\pi}_k)$ for animal k). The corresponding models for $\boldsymbol{Y}$ (the human model) and $\boldsymbol{X}_k$ (the animal model) are then linked by one of several methods (table 2.1, figure 2.5). A particular challenge is that $\boldsymbol{\pi}_k$ (the proportions of cases from each source) is a parameter in both the human model and the animal model (either directly or indirectly) and; for reasons including differential exposure, virulence, survivability, etc.; the values of $\boldsymbol{\pi}_k$ might differ between the two models. The different methods deal with this issue in method-specific ways (figure 2.5).

Table 2.1: Summary of frequency-matching source attribution methods. Each method has a human model, an animal model, and a method of linking the two models together. The data variables and model parameters are detailed in table 2.2.

Method	Human model	Details (human model)	Link	Details (animal model)	Animal model
Hald model	$Y_i \sim \mathsf{Pois}(\lambda_i)$	m_k is estimated from consumption data Priors: $s_k \sim \mathsf{U}(0, 0.01)$ $q_i \sim \mathsf{U}(0, 10)$	$\lambda_i = \sum_{k=1}^{K}(s_k q_i \pi_{ik} prev_k m_k)$	$\pi_{ik} = \frac{X_{ik}}{n_k}$ $prev_k = \frac{n_k}{T_k}$	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$
Modified Hald model	$Y_i \sim \mathsf{Pois}(\lambda_i)$	$m_k = 1$ Priors: $s_k \sim \mathsf{Exp}(\gamma_k)$ $log(q_i) \sim \mathsf{N}(0, \tau)$ $\tau \sim \mathsf{Gamma}(0.001, 0.001)$ $r_{ik} \sim \mathsf{Beta}(\alpha_{ik}, \beta_{ik})$ Here α and β are estimated by fitting a Beta distribution to the posterior product $r_{ik} = \pi_{ik} x prev_k$ to estimate α_{ik}, β_{ik}.	$\lambda_i = \sum_{k=1}^{K}(s_k q_i \pi_{ik} prev_k m_k)$ $\lambda_i = \sum_{k=1}^{K}(s_k q_i \pi_{ik} prev_k)$ $\lambda_i = \sum_{k=1}^{K}(s_k q_i r_{ik})$	$n_k \sim \mathsf{Binomial}(T_k, prev_k)$ Priors: $\pi_{ik} \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$ $prev_k \sim \mathsf{Beta}(1, 1)$	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$
sourceR	$Y_i \sim \mathsf{Pois}(\lambda_i)$	$m_k = 1$ $prev_k = 1$ $\sum s_k = 1$ Priors: $s_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$ $q_i \sim \mathsf{DP}(1, Gamma)$	$\lambda_i = \sum_{k=1}^{K}(s_k q_i \pi_{ik} prev_k m_k)$ $\lambda_i = \sum_{k=1}^{K}(s_k q_i \pi_{ik})$	Priors: $\pi_{ik} \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$ Note that this is the only jointly fit model	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$
Dutch model	$\mathbf{Y} \sim \mathsf{MN}(n, \boldsymbol{p})$	$s_k \sim \mathsf{U}(1)$	$p_i = \sum_{k=1}^{K} \pi_{ik} s_k$	$\pi_{ik} = \frac{X_{ik}}{n_k}$	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$

Continued on next page.

Table 2.1: Summary of frequency-matching source attribution methods – continued.

Method	Human model	Details (human model)	Link	Details (animal model)	Animal model
Asymmetric island model	$\mathbf{Y} \sim \mathsf{MN}(n, \boldsymbol{p})$	Priors: $s_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$	$p_i = \sum_{k=1}^{K} \pi_{ik} s_k$	μ, M, and R are estimated from the known source cases with a pseudo-likelihood approximated by a leave-one-out approach. $\pi_{ik} = \sum_j \frac{M_{jk}}{I_j} \prod_{l=1}^{L} \begin{cases} \mu_k & \text{if } a^l \text{ is novel} \\ (1-\mu_k) R_k B^l_{a^l k} & \text{if } a^l \neq g^l \\ (1-\mu_k)[1 - R_k(1 - B^l_{a^l k})] & \text{if } a^l = g^l \end{cases}$ $\sum \pi < 1$ because not all possible mutations and/or recombinations have been observed Priors: $\mu_k \sim \mathsf{Beta}(1, 1)$ $R_k \sim \mathsf{Beta}(1, 1)$ $M_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$
Dirichlet model	$\mathbf{Y} \sim \mathsf{MN}(n, \boldsymbol{p})$	Priors: $s_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$	$p_i = \sum_{k=1}^{K} \pi_{ik} s_k$	$\sum \pi = 1$ Priors: $\pi_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$	$\mathbf{X}_k \sim \mathsf{MN}(n_k, \boldsymbol{\pi}_k)$

Table 2.2: Data variables and model parameters for frequency-matching source attribution methods.

Input Variables	
$\mathbf{Y} = \{Y_i\}$	vector of counts of genotype i in the human isolates where $i \in (1 : I)$
$\mathbf{X}_k = \{X_{1k}, X_{2k}, \ldots, X_{Ik}\}$	vector of counts of genotype i in the isolates from source k, where $k \in (1 : K)$ is the index for animal sources (e.g., poultry, sheep, cattle)
a^l	allele (observed sequence) at loci (chromosomal position) l, where $a^l \in (1^l : A^l)$ and $l \in (1 : L)$
Model Parameters	
λ_{ik}	expected number of human cases of genotype i from source k
$\boldsymbol{\lambda} = \{\lambda_i\}$	vector of expected number of human cases of genotype i over all K sources ($\lambda_i = \sum_{k=1}^{K} \lambda_{ik}$)
n	total number of human isolates ($n = \sum_{i=1}^{I} Y_i$)
n_k	number of positive isolates from source k ($n_k = \sum_{i=1}^{I} X_{ik}$)
$\boldsymbol{p} = \{p_1, p_2, \ldots, p_I\}$	vector of proportions of human cases with each genotype $i \in (1 : I)$
m_k	estimated quantity of food source k consumed by the human population
$\boldsymbol{\pi}_k = \{\pi_{1k}, \pi_{2k}, \ldots, \pi_{Ik}\}$	vector of proportions of cases from source k with each genotype $i \in (1 : I)$
s_k	source effect - the ability of source k to transmit infection (the attribution probability).
q_i	type effect - the ability of genotype i to cause disease in humans
$prev_k$	prevalence of infection, over all genotypes, on source k
T_k	total number of source samples (both positive and negative)
r_{ik}	probability of a randomly chosen individual from source k having genotype i ($r_{ik} = \pi_{ik} prev_k$)
μ_k	probability, per locus, that a genotype sampled from source k contains a novel mutant allele
M_{jk}	probability of sampling an allele from source k that has already been observed in source j
R_k	probability, per locus, that a genotype sampled from source k has undergone recombination
I_j	total number of genotypes observed on source j
$B^l_{a^l k}$	probability of sampling allele a at locus l in source k

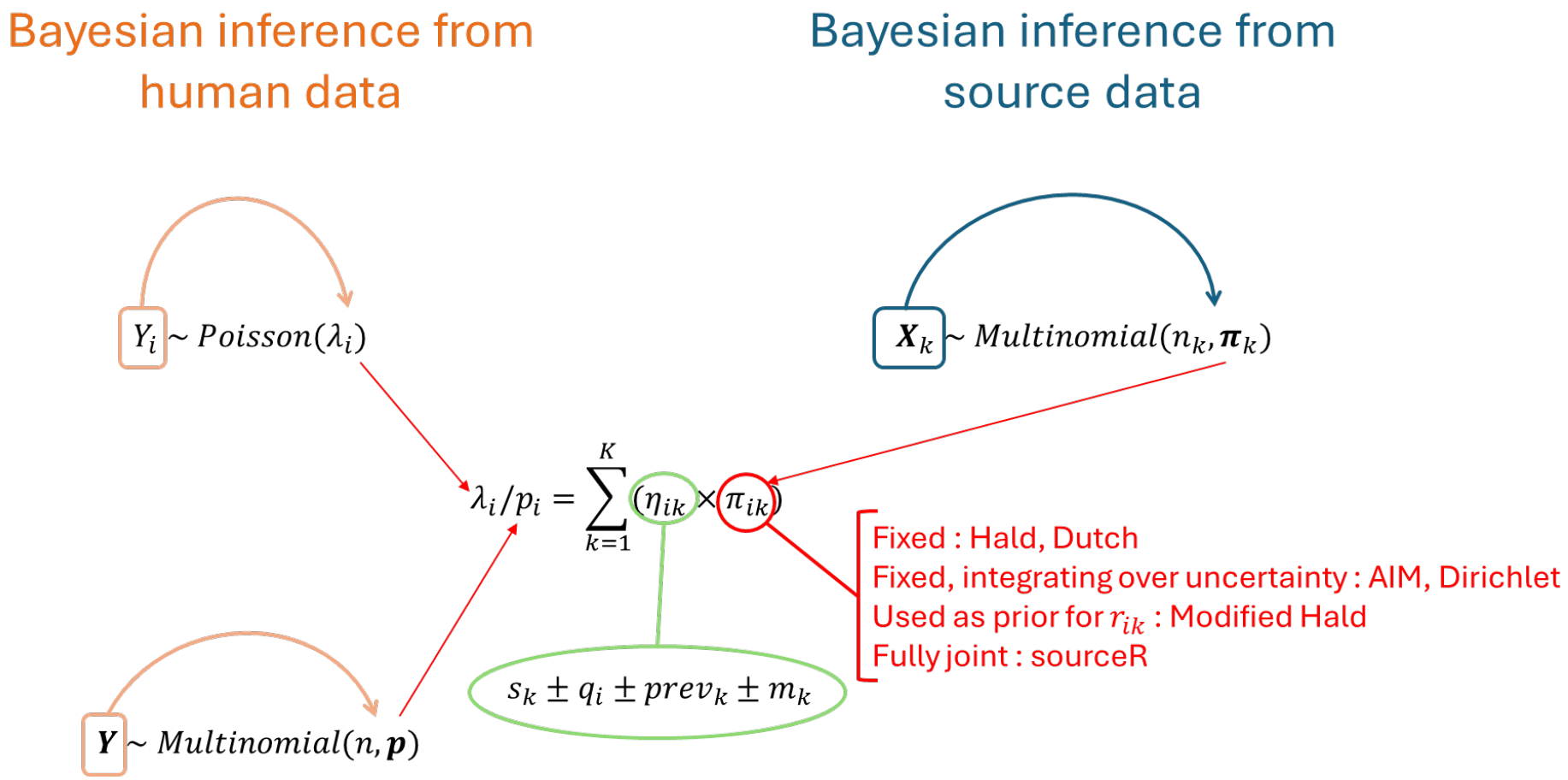

Figure 2.5: Illustrative representation of the frequency-matching source attribution methods. The treatment of π_k (the proportions of cases from each source) is method dependent. The data variables and model parameters are detailed in table 2.2.

2.4.1.1 $Y \sim$ Poisson$(\lambda) : X_k \sim$ Multinomial(n_k, π_k) Models

In this group of models, the human Poisson model and the animal Multinomial model link via a variation of:

$$\lambda_i = \sum_{k=1}^{K} \lambda_{ik} = \sum_{k=1}^{K} (s_k q_i \pi_{ik} prev_k m_k)$$

where π_{ik} (the proportion of cases from source k which are of genotype i) is estimated using the source data (X_k). The inclusion of a type specific effect (q_i) allows the distribution of genotypes in humans ($q_i \pi_{ik} prev_k m_k$) to differ from the distribution of genotypes in animals (π_{ik}). In this way, the source-specific effect (s_k) acts as an attribution weight.

The **Hald Model** (Hald et al., 2004) includes parameters for the ability of individual sources and genotypes to transmit disease to humans (s_k and q_i respectively). s_k and q_i are estimated from Y_i and are assigned priors of $q_i \sim \mathsf{Uniform}(0, 10)$ and $s_k \sim \mathsf{Uniform}(0, 0.01)$. Values of parameters π_{ik}, $prev_k$, and m_k are fixed according to the observed data, and so this model allows for no uncertainty in these quantities. To be effective, this model requires the distribution of genotypes to vary among sources. This model is over-parameterised as the number of independent data points (I) is less than the number of parameters to be estimated ($I + K$).

The **Modified Hald Model** (Mullner et al., 2009a) simplifies and extends the Hald model to allow for environmental sources. As the concept of consumption is redundant for environmental sources, the model sets the consumption parameter as $m_k = 1$ and, for food sources, the consumption information is incorporated into the model via the estimated source-specific parameter (s_k). Starting with the source data, $\mathbf{X}_k$, where $\mathbf{X}_k \sim \mathsf{Multinomial}(n_k, \boldsymbol{\pi}_k)$ and $n_k \sim \mathsf{Binomial}(T_k, prev_k)$, the relative frequencies of within-source genotypes (π_{ik}) and the overall prevalence of the organism in each source ($prev_k$) are estimated using $\mathsf{Dirichlet}(1, 1, \ldots, 1)$ and $\mathsf{Beta}(1, 1)$ priors leading to Dirichlet and Beta posteriors for π_{ik} and $prev_k$ respectively. A new parameter, r_{ik}, is then calculated from the product of the posterior π_{ik} and $prev_k$ samples, and independent $\mathsf{Beta}(\alpha_{ik}, \beta_{ik})$ distributions are fit to these by equating the first two moments. These independent Beta distributions are then used as priors for r_{ik} in the attribution stage. The final human attribution is then $\lambda_i = \sum_{k=1}^{K} (s_k q_i r_{ik})$ and the remaining priors are $s_k \sim \mathsf{Exp}(\gamma_k)$ and $\log(q_i) \sim \mathsf{Normal}(0, \tau)$; $\tau \sim \mathsf{Gamma}(0.001, 0.001)$. Modelling q_i as a random effect from a lognormal distribution with an adaptive-shrinkage Gamma prior helps with the issue of poor identifiability arising from over-parameterisation.

sourceR (Miller et al., 2017) is a further adaptation of the Modified Hald model. The consumption and prevalence parameters are fixed, $m_k = 1$ and $prev_k = 1$, and the information on consumption and prevalence is incorporated into the estimated source-specific effects (s_k). A further identifiability issue (that only the product $s_k q_i$ is identified) is resolved by constraining s_k to sum to 1. This is achieved by changing the prior from an Exponential distribution to $s_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$. In addition, a Dirichlet process is used to cluster strains (and their q_i's) by behaviours such as virulence, pathogenicity, and survivability with new priors $q_i \sim \mathsf{DP}(1, \mathsf{Gamma})$. As with the above models, $\boldsymbol{X}_k \sim \mathsf{Multinomial}(n_k, \boldsymbol{\pi}_k)$ with prior $\boldsymbol{\pi}_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$, but, in this case, the model is jointly fit. This model is the only fully

joint source attribution model currently developed. This means the uncertainty arising from both human and source data will be incorporated in both the proportion of human cases attributed to each source and the proportion of each genotype on each source. The model can be extended to include a separate estimation of attribution across independent locations (e.g., urban and rural dwellers, or different centres) and independent time periods, as described in Miller et al. (2017).

2.4.1.2 $\boldsymbol{Y} \sim$ Multinomial$(n, \boldsymbol{p})$: $\boldsymbol{X}_k \sim$ Multinomial$(n_k, \boldsymbol{\pi}_k)$ Models

In this group of models, the human Multinomial model and the animal Multinomial model link via a variation of:

$$\pi_i = \sum_{k=1}^{K} \pi_{ik} s_k$$

Here, s_k, the source effect, directly represents the probability, over all genotypes, of human cases arising from source k, and is known as the attribution probability.

For the Poisson models, the type-specific parameter (q_i) allows the probabilities of each type to differ between the human and source models. In the Multinomial models, however, the q parameters are dropped. By doing so, there is an inherent assumption that each genotype has a similar probability of infecting humans, which is unlikely to hold. As the human and animal type probabilities are expected to differ, the $\boldsymbol{\pi}_{ik}$'s are modelled using the animal model only, $\boldsymbol{X}_k \sim \mathsf{Multinomial}(n_k, \boldsymbol{\pi}_k)$, and then each posterior realisation of $\boldsymbol{\pi}_k$ is used to model the attribution probabilities (s_k's) in the human model, $\boldsymbol{Y} \sim \mathsf{Multinomial}(n, \sum_{k=1}^{K} \boldsymbol{\pi}_k s_k)$.

The simplest of these models is the **Dutch Model** (van Pelt et al., 1999). This model has a uniform prior on the attribution probabilities, $s_k \sim \mathsf{Uniform}(1)$, such that all sources are equally likely to be the source of human infection. In addition, π_{ik} is modelled directly using the proportions of source cases of each genotype i ($\pi_{ik} = X_{ik}/n_k$). Thus, the proportion of human cases of each genotype i is simply estimated to be the proportion of source cases of each genotype i. There is no uncertainty in this model and confidence intervals need to be estimated via bootstrapping methods.

The **Asymmetric Island Model** (Wilson et al., 2008) models the sampling distribution of the allelic profiles, rather than the genotypes, on the animal sources. It may also be considered a population genetic method of source attribution – each source is represented by an island and the collection of human isolates are the population. The model assumes that the observed genotypes, $i = (a^1; a^2; \ldots; a^L)$, are a result of the evolutionary processes of mutation (where a locus has a novel allele), migration (where the genotype has been observed in a different source), and recombination (where the allele at a locus has been observed before but in a different genotype). Probabilities are assigned to each of these processes, on each island source, from the allelic frequencies at given loci.

If the loci are not independent, and there is migration of strains between sources, then i may be a copy of a genotype g from a source j other than k. Mutation and recombination mean that i may contain novel alleles, or comprise a novel combination of existing alleles. Using the multinomial animal model, $\boldsymbol{X}_k \sim \mathsf{Multinomial}(n_k, \boldsymbol{\pi}_k)$, π_{ik} is then estimated via the following

function with Beta priors on μ_k and R_k and a $\mathsf{Dirichlet}(1, 1, \ldots, 1)$ prior on M_k:

$$\pi_{ik} = \sum_j \frac{M_{jk}}{I_j} \prod_{l=1}^{L} \begin{cases} \mu_k & \text{if } a^l \text{ is novel} \\ (1-\mu_k) R_k B^l_{a^l k} & \text{if } a^l \neq g^l \\ (1-\mu_k)[1 - R_k(1 - B^l_{a^l k})] & \text{if } a^l = g^l \end{cases}$$

The probabilities μ, M, and R are estimated from the known source cases with a pseudo-likelihood approximated by a leave-one-out approach. For each posterior sample of these, a Markov Chain Monte Carlo (MCMC) side-chain with various Metropolis-Hastings updates is run to estimate the probabilities s_k for the human model using a $\mathsf{Dirichlet}(1, 1, \ldots, 1)$ prior. This model may be extended to allow attribution to change through time, or to include covariates for the human cases by modelling the s_k probabilities by case through time. This model considers the collection of observed genotypes to be a sample of all possible genotypes, so that $\sum \pi_{ik} < 1$. This means the model can attribute human cases of a genotype that has not previously been observed in one of the K sources by assessing genetic similarity of alleles. The model requires a high degree of genetic diversity within the sources. Low divergence among sources can limit its usefulness, such as when data are sourced from different datasets (e.g., non-local or non-recent) (Smid et al, 2013).

The **Dirichlet Model** (Liao et al., 2019) simplifies the asymmetric island model by using a Dirichlet distribution in place of the evolutionary model for π_k. Modelling the prior for π_k as $\pi_k \sim \mathsf{Dirichlet}(1, 1, \ldots, 1)$, which is the conjugate prior for the multinomial distribution, allows the posterior distribution for π_k to be modelled as $\pi_k \sim \mathsf{Dirichlet}(\boldsymbol{X}_{k+1})$ which can be easily calculated. Because this model now considers the collection of genotypes observed to be the entire population of genotypes, $\sum_i \pi_{ik} = 1$, and as a result the model cannot attribute human cases of a genotype which has not previously been observed in one of the K sources.

Each of the frequency-matching methods have limitations, particularly with increasing genetic resolution. Any chosen level of resolution for a model restricts attribution to that level of resolution. For example, isolates belonging to a single genotype, as defined by the 7-gene MLST scheme, are all attributed in the same way. Increasing the number of genes included in the model will increase the number of unique genotypes so that attribution can continue within a single genotype; however, as the number of unique genotypes increases, the counts of each unique genotype decrease, ultimately to single counts. At this point the Poisson and Multinomial models are ineffective and the genotypes found in humans will ultimately become distinct from those found in the sources. With the exception of the asymmetric island model, these models cannot attribute an isolate unless it has been observed in one of the sources. The asymmetric island model, rightly or not, will attribute it the source with the highest diversity. Another potential issue of the asymmetric island model, is that increasing the number of alleles in the model not only reduces the functionality of the count-based model, but also results in the attribution probabilities changing (Lake et al., 2021) and favouring the source with the lowest variability in its complement of genotypes (figure 2.6).

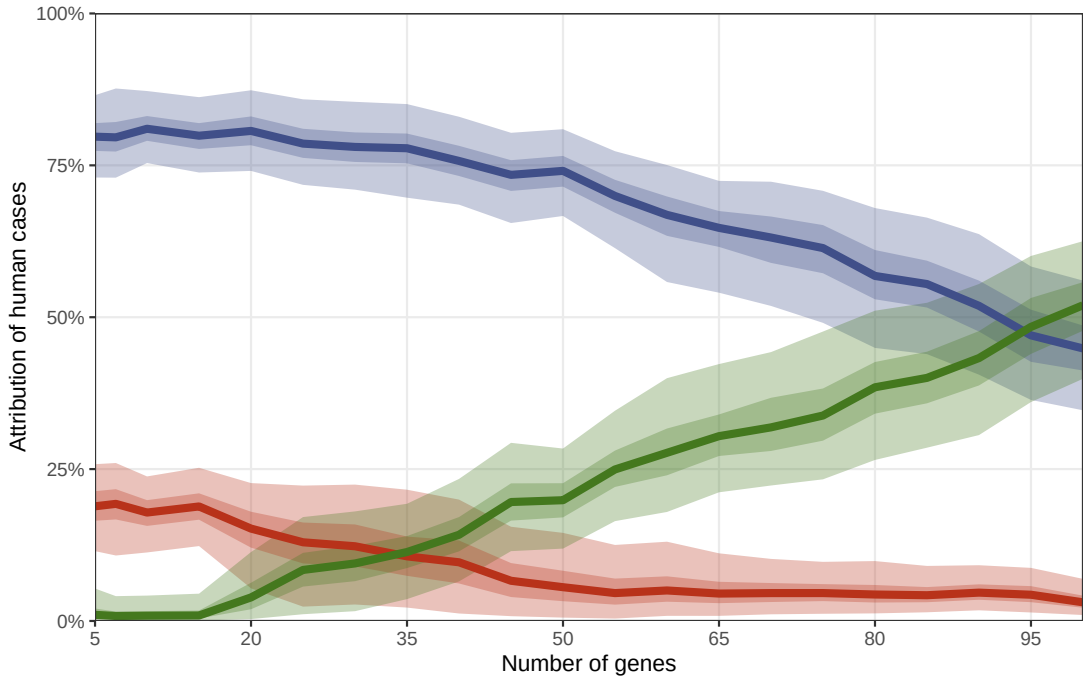

Figure 2.6: Asymmetric island model attribution results. As the number of genes in the model increases, the attribution probabilities change. Each colour represents a different host source.

2.4.2 STRUCTURE

STRUCTURE, developed by Pritchard et al. (2000) and updated several times since (Falush et al., 2003, 2007; Hubisz et al., 2009) is another source attribution model based on population genetics. Similar to the asymmetric island model, STRUCTURE is a Bayesian clustering method using multilocus genotype data to estimate allele frequencies at each genetic locus for each source. By assuming that each locus is independent, the probability distribution of allele frequencies is a multinomial distribution on counts of alleles (rather than on counts of genotypes as in the Dirichlet model above). Each human case is assigned probabilistically to a single source (the no admixture model) based on their genotypes.

Let:

- $\boldsymbol{G}$ be the set of observed genotypes, where g_l^i is the sequence (genotype) of isolate $i \in (1:I)$ at locus $l \in (1:L)$ loci, from both the source and the human isolates
- $\boldsymbol{S}$ be the set of K predefined sources, where $s^{(i)}$, is the source from which isolate i originated
- $\boldsymbol{F}$ be the set of frequencies, where f_{sla}, is the frequency of allele a, at locus l, in source s
- N_l be the number of distinct alleles observed at locus l

Assume that each source $s \in (1:S)$ is modelled by a characteristic set of allele frequencies. Under the assumptions of (i) Hardy-Weinberg equilibrium within populations and (ii) complete linkage equilibrium between loci within populations, each allele (a) at each locus (l) in each

genotype (g) is an independent draw from the appropriate frequency distribution. The probability distribution is therefore $P(G \mid S, F)$. With suitable priors on S and on F, the posterior distribution is $P(S, F \mid G) \propto P(S)P(F)P(G \mid S, F)$. Given the source of each isolate, $P(G \mid S, F) = P(g_l^i = a \mid S, F) = f_{s(i)la}$ – the frequency of allele a at locus l in the source of individual i, independently for each g_l^i. Assuming that the probability that an isolate originated in source s is proportional to the size of each cluster s, the prior on S is $P(S) = \frac{\sum g_{s(i)}}{\sum G}$, independently for all individuals. The prior on F is $f_{sl} \sim \mathsf{Dirichlet}(\lambda_1, \lambda_2, \ldots, \lambda_{A_l}) = \mathsf{Dirichlet}(1, 1, \ldots, 1)$ independently for each source s, and locus l. Then an MCMC algorithm is used to estimate $P(S, F \mid G)$. For source attribution, the model uses the source data as a training set (i.e., the source is known), also known as the burn-in stage, and the human isolates as the test set with source unknown.

Although the STRUCTURE model is potentially applicable to whole genome sequencing (WGS) data (Berthenet et al., 2019; Denis et al., 2023; Mughini-Gras et al., 2021; Saif et al., 2022; Thépault et al., 2017) the computation time increases linearly with each additional locus and using a large number of loci (e.g., > 1000) is impractical (Pérez-Reche et al., 2020). Additionally, increasing the number of loci compounds the attribution errors resulting from even small errors in estimates of allele probabilities. Solutions have been proposed for this in terms of pre-selecting genes based on their discriminatory ability (Collins and Didelot, 2018; Rosenberg et al., 2003; Thépault et al., 2017) or using a method of attribution bias correction (Berthenet et al., 2019). It is, however, circular logic to assess results of self-attribution following pre-selection of variables on the same criteria of self-attribution, so it is difficult to assess these methods. Pérez-Reche et al. (2020) describe a new method using information theory to select loci based on decreasing allelic diversity within sources and/or increasing allelic diversity between sources but there is no record that this has been applied in combination with STRUCTURE and the effectiveness is not known. A further issue is the underlying assumption of complete linkage equilibrium between loci within populations. Both *C. jejuni* and *C. coli* have linked loci and the number of linked loci increases with the numbers of loci (Meinersmann et al., 2002; Suerbaum et al., 2001). Although Falush et al. (2003) adapted STRUCTURE to allow for linked loci, all known published studies utilising STRUCTURE for source attribution of *Campylobacter* have used the unlinked (uncorrelated allele frequency) model (Benshak et al., 2023; Berthenet et al., 2019; Denis et al., 2023; Kovac et al., 2018; Mughini-Gras et al., 2021; Saif et al., 2022; Sheppard et al., 2009; Thépault et al., 2017) and the effect on the results of not meeting this independence assumption has not been quantified.

2.4.3 Minimal Multilocus Distance Method

The minimal multilocus distance (MMD) method by Pérez-Reche et al. (2020) is based on the Hamming distance (d_H). d_H is a variation of the simple matching coefficient and is the number of positions (e.g., alleles, or single nucleotide polymorphisms (SNPs)) at which the corresponding variables are different. The similarity between isolates and sources is interpreted in terms of the cumulative distribution function of d_H, $F_{h,s}(\lambda)$. Using the principles of machine learning, parameters (q_s) which optimise self-attribution probabilities are determined, and for each

isolate to be attributed, $F_{h,s}(\lambda)$ is calculated (for each potential source) and the parameters determined from the machine learning process (q_s) are used to calculate the attribution probability.

Let:

- $\mathbf{g}_h = \{g_{h,l}\}_{l=1}^{L}$ be the genotype of isolate h, from a human where the origin of infection is unknown
- $\mathbf{g}_{i,s} = \{g_{i,s,l}\}_{l=1}^{L}$ be the genotype of isolate $i \in I$ from source $s \in S$; and $g_{i,s} \in G_s$, the set of observed genotypes from all isolates of source s
- $d_H(\mathbf{g}_h, \mathbf{g}_{i,s}) \in \lambda$ be the Hamming distance (the number of mismatches) between $\mathbf{g}_h$ and $\mathbf{g}_{i,s}$, and λ be the range of possible numbers of mismatches (i.e., 1:the total number of SNPs)
- $\lambda_{h,s}(q)$ be the q-quantile, which is the value of λ (for each source) at which the cumulative distribution function, $F_{h,s}(\lambda) = q$, for a given probability q
- $\lambda_{min} = min_s\{\lambda_{h,s}(q)\}$ be the smallest source-specific value of λ
- $\sigma_{h,s} = F_{h,s}(\lambda_{min})$ be the probability, for each source s, that $d_H(\mathbf{g}_h, \mathbf{g}_{i,s}) \leq \lambda_{min}$

Then the probability that a human isolate, h, originated from source s is:

$$p_{h,s} = \frac{\sigma_{h,s}}{\sum_{s' \in S} \sigma_{h,s'}}$$

In figure 2.7, the curves show the source-specific cumulative distribution functions $F_{h,s}(\lambda)$ which gives the probabilities that d_H between an isolate of unknown origin and any genotype from the corresponding source is smaller than λ. The value of q determines the attribution probabilities and therefore has the potential to bias results, however the attribution probability, p_s, is not very sensitive to the specific value of q, provided it is within the range in which self-attribution probability is high. For each source, the optimal self-attribution value, q_s, is determined as the value which minimises $\frac{q_s(1-p_s)}{p_s}$. q is then chosen to be the mean of the optimal self-attribution values, q_s, weighted by the number of isolates in each source. Estimates of uncertainty for the attribution probabilities, p_s, are obtained via bootstrap samples from the set of source probabilities $\{p_{h,s}\}_{h=1}^{I_h}$. Self-attribution values are determined by a Monte-Carlo cross-validation strategy using 50% of the samples (the training set) from each source.

The attribution results of the MMD method, based on core-genome SNP (cgSNP) genotypes, compare favourably against STRUCTURE (Pérez-Reche et al., 2020). d_H can be calculated for a unique genotype not observed in any of the sources. The computational time of MMD is related to d_H and not to the number of loci, and MMD can deal with genotypes comprising thousands of loci with minimal computational effort. Despite this, Pérez-Reche et al. (2020) compared attribution results using both an increasing number of loci and using loci selected based on information theory and found that selecting loci which decrease allele diversity within sources decreased the number of SNPs required to obtain optimal self-attribution.

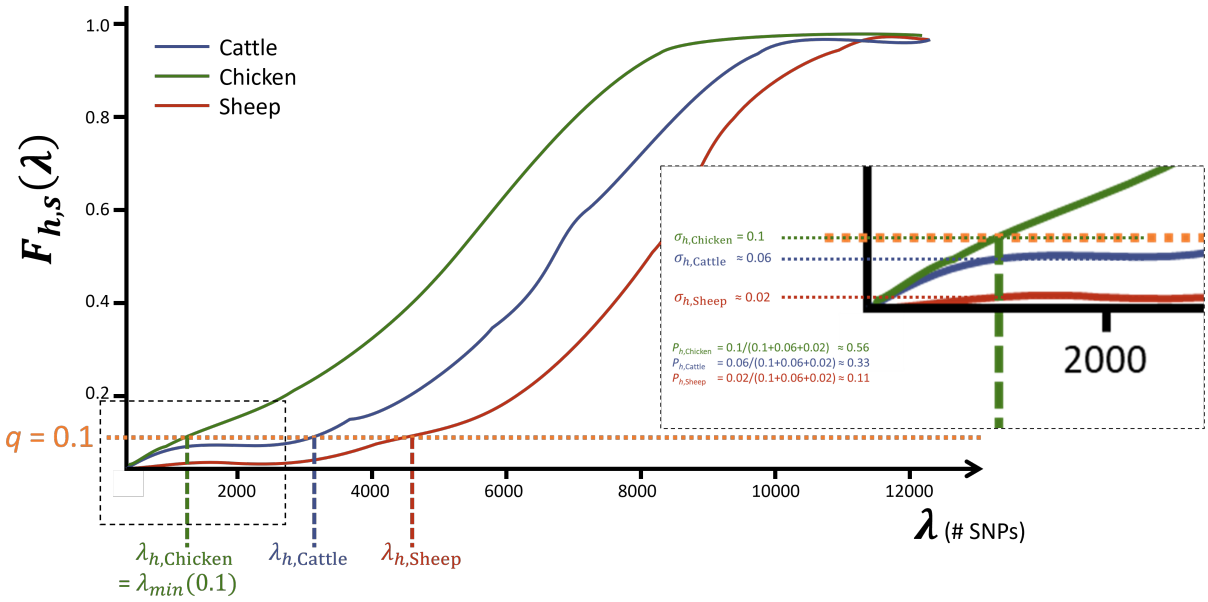

Figure 2.7: Example of the determination of attribution probabilities, $p_{h,s}$, in the MMD method.[3]

2.4.4 Genome-Wide Association Study

A genome-wide association study (GWAS) assesses whether a particular genetic variant is found more often than expected in individuals with a particular phenotype. The method is commonly used to identify genetic variants associated with a particular disease, and it has been adapted as a method for source attribution by looking for genetic variants underlying preferential host colonisation (Sheppard et al., 2013b). The method by Sheppard et al. (2013b) identifies unique 'words' (k-mers) of 30 bp sequences which are more strongly associated with a particular host than would be expected based on neutral patterns of evolution, given the clonal relationships of the bacteria in the sample and their distribution among hosts. Monte Carlo simulations are used to measure the statistical association of a word with the host from which each genome was isolated. Words are simulated based on a reconstructed phylogeny and the correlation of the real word is then compared with the distribution of the correlations of the simulated words to produce a phylogenetically correct p-value. Significant words are then mapped to genes and further investigated.

Collins and Didelot (2018) developed **treeWAS**, an adaptation of the GWAS method that uses a phylogenetic approach. This method reconstructs the phylogeny and computes the homoplasy distribution containing site-specific numbers of substitutions drawn from the empirical dataset. The simulated dataset then resembles the empirical dataset in population structure and genetic composition and includes the effects of mutation and recombination. Association testing is then performed as described in Collins and Didelot (2018). treeWAS therefore circumvents the confounding effects of population structure and recombination inherent in GWAS methods.

[3]Image adapted from "Pérez-Reche et al (2020) Mining Whole Genome Sequence data to efficiently attribute individuals to source populations. Additional file 2: Supplementary figures" available at https ://doi.org/10.1038/s41598-020-68740 -6.

In essence these are methods of selecting loci which may discriminate sources rather than being true source attribution models.

2.4.5 Machine Learning

Recently, there has been a surge in machine learning methods for source attribution of *Salmonella* (Guillier et al., 2020; Guzinski et al., 2024; Lupolova et al., 2019; Munck et al., 2020; Thystrup et al., 2024; Zhang et al., 2019) and more recently *Campylobacter* (Arning et al., 2021; Brinch et al., 2023; Pascoe et al., 2024), with varying levels of success.

2.4.5.1 Unsupervised Machine Learning

Unsupervised machine learning (e.g., principal component analysis (PCA), clustering methods) are useful for revealing underlying structure in the data, but where the structure is not dominated by host source, then these methods struggle to differentiate inherent ancestral structure from host source. Some recent studies have applied **weighted network analyses**, a clustering method, to the source attribution of *Campylobacter* (Brinch et al., 2023; Wainaina et al., 2022). Network analysis is based on weighted networks theory, which provides a direct comparison of genomic distances among isolates of different sources via weighted network links. Each node in the network is an isolate, and links between isolates correspond to their genetic distance. Weaker links imply a greater genetic distance between isolates. The probability of a human isolate being associated with an animal source is computed as the function of the number of links that the human isolate has with other animal isolates (Merlotti et al., 2020). This method can be applied to data of any resolution (e.g., wgMLST, cgMLST, or SNP) (Merlotti et al., 2020; Wainaina et al., 2022).

2.4.5.2 Supervised Machine Learning

Supervised machine learning algorithms are trained on variations in the genomes (e.g., of alleles) isolated from animal sources to predict the source from which a human isolate originated. **Random forest** (Breiman, 1996, 2001) is a tree-based method of supervised machine learning that has been used in a handful of source attribution studies for *Campylobacter* (Arning et al., 2021; Brinch et al., 2023).

For a training set of N independent observations (the animal isolates) on P variables, where $x_n = (x_{n1}, x_{n2}, \ldots, x_{nP})$ is the vector of predictor variables for observation $n = 1, 2, \ldots, N$, and y_n is the corresponding response variable, classification and regression tree (CART) is a greedy recursive binary partitioning algorithm that successively partitions data (the parent node) into two smaller subsets (the left and right child nodes). Each binary partition is based on a decision rule for a single predictor variable chosen to achieve maximal reduction in the impurity of the response variable in the resulting child nodes (Breiman et al., 1984). The Gini index is a common measure of impurity and is simply a measure of the likelihood of an isolate chosen at random being incorrectly classified if it was randomly classified according to the distribution of

sources from the data set. The tree continues to grow until a stopping rule is reached or until each observation has been assigned to a terminal node. A classification can then be predicted for a new observation (i.e., a human isolate) by sending it down the tree according to the decision rules until it arrives at a terminal node.

Individual classification trees tend to overfit to the training data, that is, they yield decision rules that are more specific to the training data than they are to new independent data. Random forest addresses this issue by creating an ensemble of classification trees. The individual trees that make up the ensemble differ from one another because they are each trained on a different random sample of the training observations ('bagging') and predictor variables ('random subspacing'; Amit and Geman, 1997; Breiman, 1996; Ho, 1998). The predictions from the individual trees are aggregated and classifications are made based on the majority vote across the trees. Various control parameters can be set for random forest models, including the number of trees, the number of variables randomly selected as splitting candidates, and tree size (Wright and Ziegler, 2017).

Tree-based source attribution models suffer a range of issues. They are not robust to unequal sample sizes among the sources, and isolates may be over-attributed to sources with the largest number of observations. In addition, the models do not account for sources that were not included in the training data – every human isolate will be attributed to one of the known sources. However, the greatest issue with tree-based methods occurs when a level of a categorical predictor variable (i.e., an allele) is absent when a tree is grown, but is present in a new observation for prediction (i.e., unique strains in the human isolates) (the absent-levels problem *sensu* Au, 2018). In a random forest algorithm, this situation can arise due to sampling variability (i.e., the strain was absent from the observations that were used to train the model), bagging (i.e., the strain was in the training data but absent from the bootstrapped sample used by a particular tree), or tree design (i.e., the strain was present at the top of the tree but absent from a lower subset created by binary splits). When the algorithm encounters a unique strain, there is no immutable *a priori* rule for determining which side of the binary split the isolate should go.

For the algorithm to proceed with a unique strain, a heuristic rule is required. Available heuristics include stopping an affected isolate from proceeding down the tree (Therneau et al., 2022), using a surrogate decision rule that mimics the original split's partitioning (Hothorn and Zeileis, 2015; Therneau et al., 2022), directing all affected observations down the branch with more training observations (Hothorn and Zeileis, 2015), directing all affected observations down the same branch (i.e., "left" or "right heuristic") (Liaw and Wiener, 2002; Wright and König, 2019), directing all affected observations down both branches simultaneously (Saar-Tsechansky and Provost, 2007), randomly directing affected observations down a left or right branch (Hothorn and Zeileis, 2015), and binary encoding predictors. The choice of treatment of absent levels can dramatically alter a model's performance and potentially lead to systematic bias (Au, 2018).

Another approach is to have the model interpret allele numbers quantitatively (the equivalent of treating a categorical variable as ordinal) as this allows for the classification of unique strains.

Although this has been the approach taken by several studies (Arning et al., 2021; Guzinski et al., 2024; Tanui et al., 2022), this treatment may be detrimental to random forest predictions (Wright and König, 2019). It is also problematic if the ordering of the alleles has some degree of association with the source. For example, the open-access PubMLST database (Jolley et al., 2018) defines alleles numerically and in a sequential manner based on sequence deposition. In this instance, treating alleles as numeric would not be appropriate because allele "1" is not necessarily more related to allele "2" than it is to allele "500". However, it is likely that isolates have been added to the database in groups according to host source, so that their numeric order may partition into contiguous chunks by host. The numeric order thus provides information on likely host sources which is external to the data in a particular study, potentially biasing class assignment (chapter 7, appendix D.1).

Many models use variable selection as a first step in the analysis (e.g., Arning et al. (2021); Brinch et al. (2023); Guzinski et al. (2024); Munck et al. (2020); Pascoe et al. (2024); Thystrup et al. (2024)) with the aim of reducing the overall quantity of data while retaining the variables which have discriminatory information (Collins and Didelot, 2018; Pérez-Reche et al., 2020; Rosenberg et al., 2003; Thépault et al., 2017). This both reduces the risk of unique genotypes and also helps decrease computational time, but it can counter the benefits of using high resolution data (e.g., Munck et al. (2020) reduced 3002 *Salmonella* core genes down to 17; and Pascoe et al. (2024) reduced 1343 *Campylobacter* core genes down to 15). Increasing in popularity is the transformation of nucleotide sequence data into *k*-mers (Kokot et al., 2017) and unitigs (formed from *k*-mers) (Jaillard et al., 2018) which avoids the problem of unique alleles (Brinch et al., 2023; Thystrup et al., 2024), and this approach has also been used in combination with a boosting algorithm (e.g., **xgboost** and **logit-boost**) (Arning et al., 2021; Brinch et al., 2023; Pascoe et al., 2024).

2.4.6 Whole Genome Sequencing Data for Source Attribution

Current working models for source attribution have been developed using a small number of genes and using data based on alleles. Many studies using these methods with 7-gene MLST data have successfully identified the relative contribution of different host sources to human infection in several countries (e.g., Mullner et al. (2009b); Sheppard et al. (2009); Strachan et al. (2009); Thépault et al. (2018)). A limitation of attribution methods using low resolution data is the ability to discriminate between some sources, such as between different ruminant sources (Cody et al., 2019; McCarthy et al., 2007; Mughini-Gras et al., 2012; Ogden et al., 2009). In addition, some strains, known as generalist strains, are found in multiple hosts (e.g., ST474 in New Zealand) and are identical across the seven MLST loci and cannot be discriminated further based on only these seven genes. Increasing the number of genes included in a model will increase the number of genotypes so that attribution can continue within a single ST. Lake et al. (2021) show that the results of using 7-gene MLST, 50-gene ribosomal MLST, or 13 genes from the core genome for attribution (as described in Thépault et al. (2017)) alters the attribution of each human case.

WGS data provides information on allelic variation across the entire genome, improving the discriminatory power of data for source attribution (Cody et al., 2017). For example, WGS data identifies discrete clades within ST474 which are more source specific (figure 2.8). WGS isolates can be compared across various degrees of similarity, providing different resolutions for genotyping. For example, SNP analysis compares SNPs across the parts of the individual genomes that align to a reference genome, while cgMLST and wgMLST use gene-to-gene comparisons. The increasing availability of whole genome data allows for the ultimate level of discrimination among close strains.

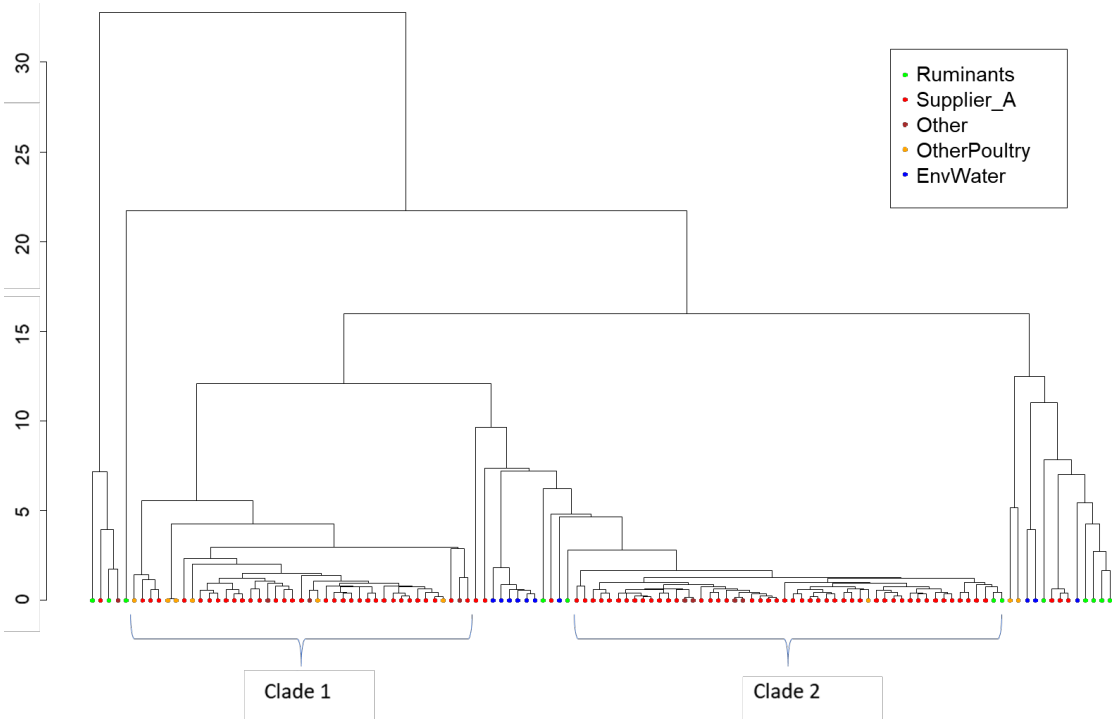

Figure 2.8: Complete linkage dendrogram of simple matching distances between isolates of a single sequence type (ST474) based on core genome allelic profiles. The higher resolution allows isolates to cluster into clades which are more source specific.

2.4.6.1 Limitations of using WGS data for Source Attribution

Although it is widely accepted that using a greater number of loci will have more successful attribution (Berthenet et al., 2019; Collins and Didelot, 2018; Merlotti et al., 2020; Munck et al., 2020; Pérez-Reche et al., 2020; Thépault et al., 2017), there is a balance between increasing the number of loci and model failure (figure 2.9). A major limitation of using cgMLST as input data is the substantial number of missing alleles, which occurs when no matches can be found in the PubMLST database. This may be because of incomplete assembly or because of the variable nature of the *Campylobacter* genome which results from horizontal gene transfer and mutations. In addition, when using a large number of genes, human isolates which have a unique genotype (a genotype not present in the combined set of source isolates) are likely to contribute a large proportion of isolates. The number of non-unique genes will also decrease with increasing

numbers of human isolates, until conceivably every gene contains an allele unique to the human population.

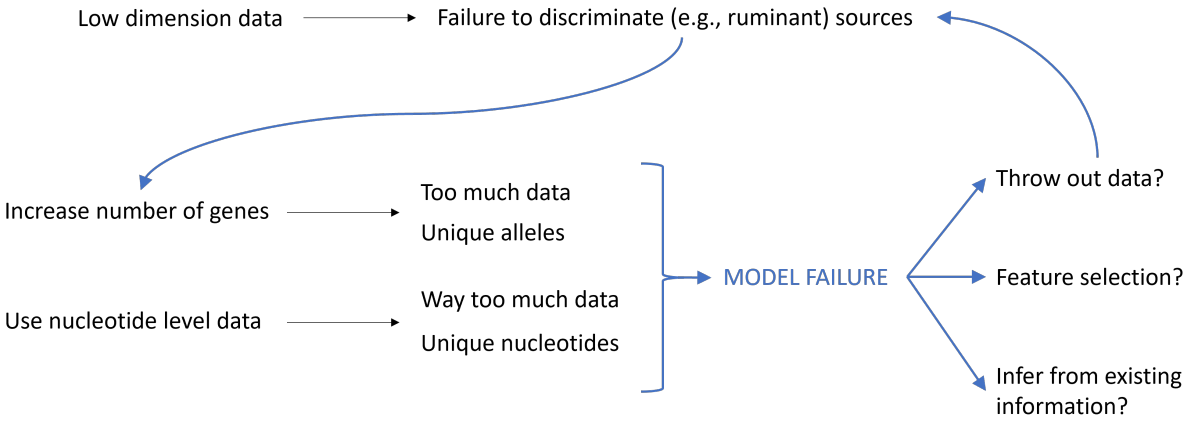

Figure 2.9: The cycle between increasing data resolution and subsequently reducing the number of predictor variables.

As the number of genes increases, the number of genotypes increases, but the counts of each particular genotype decrease, ultimately to single counts, at which point the genotypes found in humans will be distinct from those found in the sources. The early frequency-matching models (Hald, modified Hald, sourceR, and Dutch models) and the Dirichlet model then fail because they cannot attribute an isolate to a source unless its genotype has been observed in one of the sources (Liao et al., 2019). The asymmetric island model can attribute human cases of a genotype that have not been observed in a source. This model, in its original form, suffers from inherent model limitations in that increasing the number of genes results in the attribution probabilities changing. This may be solved, however, by estimating common recombination and mutation rates across all sources.[4] STRUCTURE fails because of lengthy computation time as well as the increased likelihood of loci linkage (thereby breaking the assumption underlying the model). Machine learning models may fail because of the inability to attribute genotypes which, as well as being unique, contain alleles which have not been encountered in the source populations. To illustrate the difference, table 2.3 shows four isolates, two from host sources and two from humans. The first human isolate (Human 1) is a unique genotype but does not contain any unique alleles; the second human isolate is a unique genotype and it contains a unique allele, d3, for Gene D. One potential solution is to discard isolates which have unique genotypes (e.g., Human 1 and Human 2), and/or variables which have unique alleles in the set of human isolates (e.g., Gene D in the above example), however this may negate the benefits of increasing the resolution of the data. The MMD method navigates these issues, but preliminary results indicate a low attribution success for ruminant sources (Pérez-Reche et al., 2020) which is a historical issue with attribution methods (French and Marshall, 2014). The reasons for 'failure' of common source attribution models when presented with large numbers of variables is illustrated in table 2.4.

[4]https://github.com/jmarshallnz/islandR

Table 2.3: The difference between a unique genotype (Human 1) and a unique genotype with a unique allele (Human 2).

Source	Gene A	Gene B	Gene C	Gene D
Animal 1	a1	b1	c1	d1
Animal 2	a1	b2	c2	d1
Human 1	a1	b2	c1	d1
Human 2	a1	b2	c2	d3

Table 2.4: Reasons for 'failure' of common source attribution models (annotated with a cross).

Model	**Handles unique genotypes**	**Handles large data**	**Handles unique alleles**	**Potential model issues**
Hald	×	×	×	-
Modified Hald	×	×	×	-
sourceR	×	×	×	-
Dutch	×	×	×	-
Asymmetric island	✓	×	✓	inconsistent attribution
Dirichlet	×	×	×	-
STRUCTURE	✓	×	✓	slow computation
MMD	✓	✓	✓	poor ruminant attribution
Machine learning	✓	✓	×	-

2.5 Distance-Based Analyses

A unique allele for a given gene may vary from another allele for the same gene by as little as one nucleotide, or may differ at every nucleotide. A measure of this dissimilarity can be achieved with a distance measure which quantifies how similar any two alleles are, on the basis of their nucleotide sequences. For example, the Hamming distance gives the number of nucleotides at which two alleles differ; and the simple matching coefficient gives the proportion of nucleotides at which two alleles match.

The concept of using distances between isolates for source attribution is not new. Merlotti et al. (2020) used pairwise distances as input for a weighted network analysis where nodes correspond to genomes, and links to genetic distances. Here, both phenotypic data such as serotyping, or molecular data such as 7-gene MLST, cgMLST, wgMLST, or SNP, can be used to establish distances between nodes. The MMD method (Pérez-Reche et al., 2020) also uses pairwise Hamming distances in the calculation of attribution probabilities. An option for tree-based methods that has not been explored in depth, is to replace a unique allele at any branch with its 'closest' match, determined by its pairwise distance with all other alleles in the source data. With this treatment, there is no need to discard the gene, nor the isolate (e.g., see figure 2.10).

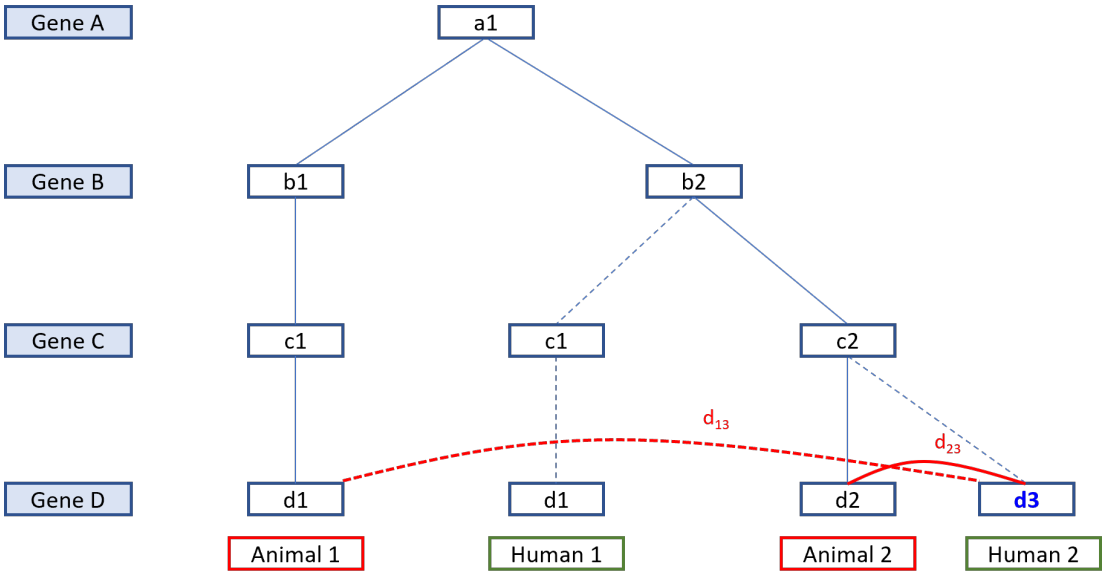

Figure 2.10: Replacement of a unique allele with its 'closest' match. An isolate with a unique allele at Gene D (d3) may be assigned to the branch with the smallest genetic distance (d_{13} *versus* d_{23}).

Clearly the choice of dissimilarity measure will define the definition of 'similar'. For example, describing two colours as being similar or dissimilar depends on the measure with which they are defined and then compared. A poorly informative definition of colour is its name (e.g., blue, green, purple). Far more informative is defining a colour using the standard RGB (sRGB) colour space that defines the chromaticities of the three primary colours red, green, and blue (e.g., (0, 0, 128), (34, 139, 34), (160, 32, 240)). Yet another option is one of the non-standardised, cylindrical, colourspaces: HSV (hue, saturation, value) (e.g., (240, 100, 25.1), (120, 60.7, 33.9), (277, 87.4, 53.3)) or HSL (hue, saturation, lightness) (e.g., (0.667, 1, 0.502), (0.333, 0.755, 0.545), (0.769, 0.867, 0.941)).

The R statistical computing environment (R Core Team, 2024) contains a collection of 502[5] built-in colour names. For each colour, the red, blue, and green values in hexadecimal are available,[6] as are the HSV[7] and HSL[8] values. Calculating pairwise distances of the colours using each of the four descriptions illustrates the impact that choice of both distance measure, and variable characteristic, may have in defining 'similarity' (figures 2.11, 2.12). The Levenshtein distance (i.e., the minimum number of single-character edits required to change one word into the other) on the colour names defines two colours as similar if they are spelt similarly and is not capturing the essential characteristics of 'colour' (figure 2.11(a), 2.12(a)). The Euclidean distance of colours in the sRGB space results in a colour spectrum which is separating colours more intuitively while also capturing differences in lightness of colours (figure 2.11(b), 2.12(b)).

[5]the full collection has 657 named colours as it includes both American and British spellings for the same colour (e.g., 'gray' *versus* 'grey')

[6]sRGB values were obtained using the `seqinr` package (Charif and Lobry, 2007)

[7]HSV values were obtained using the `DescTools` package (Signorell et al., 2024)

[8]HSL values were obtained using the `farver` package (Pedersen et al., 2024)

The Euclidean distance of colours in the HSV colour spaces is emphasising differences in saturation (figure 2.11(c), 2.12(c)), and in the HSL space it is giving the familiar rainbow pattern (figure 2.11(d), 2.12(d)).

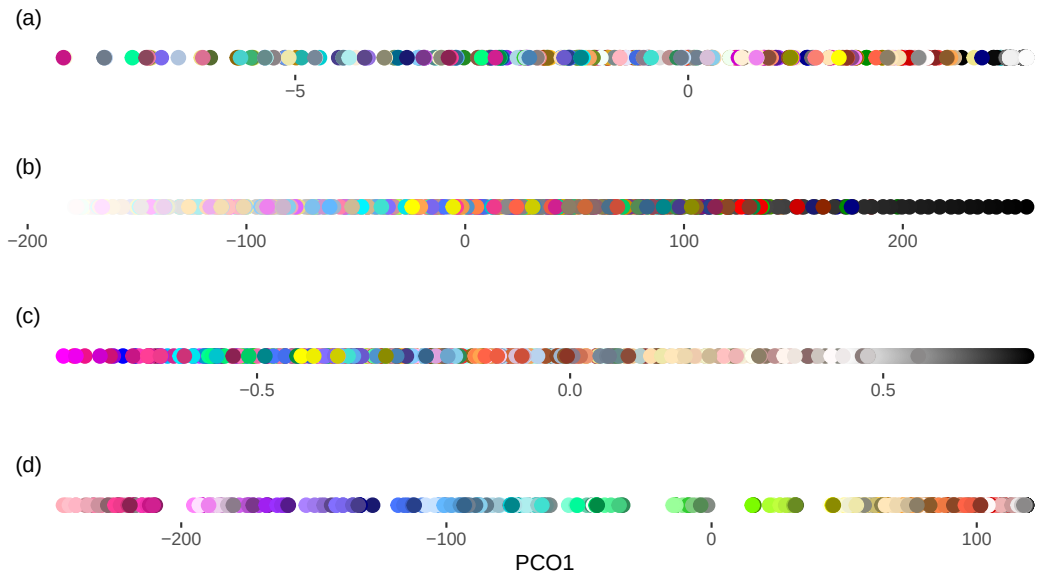

Figure 2.11: One dimensional principal coordinate ordination of 502 named R colours based on four measures of comparison; (a) Levenshtein distance of colour names, (b) Euclidean distance of colours in the sRGB space, (c) Euclidean distance of colours in the HSV space, (d) Euclidean distance of colours in the HSL space.

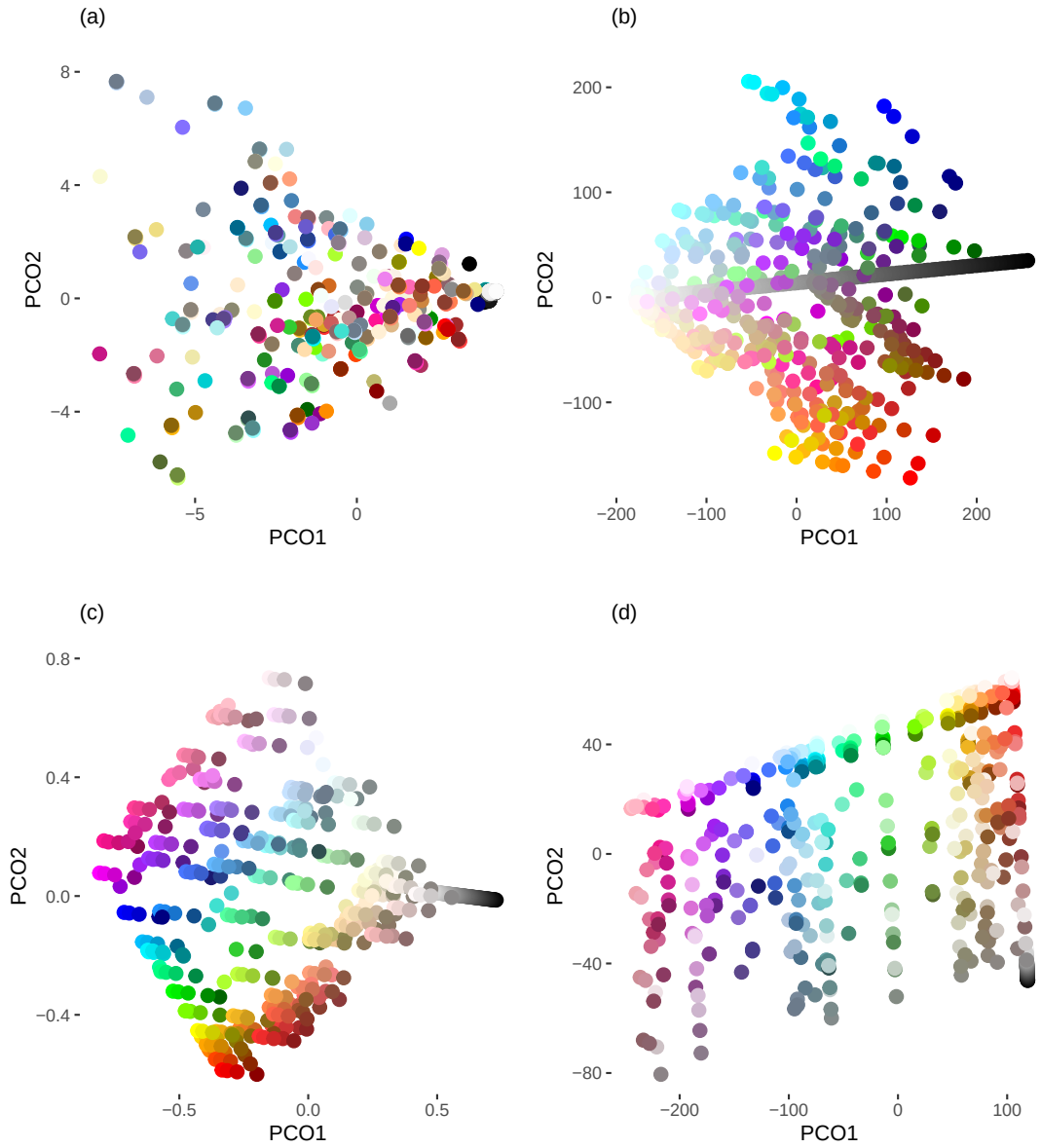

Figure 2.12: Two dimensional principal coordinate ordination of 502 named R colours based on four measures of comparison; (a) Levenshtein distance of colour names, (b) Euclidean distance of colours in the sRGB space, (c) Euclidean distance of colours in the HSV space, (d) Euclidean distance of colours in the HSL space.

2.6 Conclusion

The problem of attributing species of *Campylobacter* to host source has been intensively investigated in the past two decades. A wide range of approaches have been applied which attempt to solve this problem. Epidemiological approaches have focused on the contribution of sources at the point of exposure, rather than the original reservoir. Frequency-matching and population genetics models using microbial subtyping data have focused on the original source, however, they have experienced issues with the unique genotypes that are inevitable with high resolution WGS data. The focus has now shifted to machine learning approaches such as random forest and boosted algorithms which are showing promise using high resolution genomic data. The tree-based approach, however, is complicated by the presence of new alleles (absent levels) in human isolates. The problem of absent levels in random forest will be explored in this thesis and applied to source attribution of *Campylobacter* species, using WGS data.

Chapter 3

SACNZ core-genome MLST Data Set: Implications for Source Attribution

3.1 Introduction

Campylobacter species are highly diverse organisms (Cody et al., 2017; Dingle et al., 2001) which are constantly evolving and the collection of recognised alleles is steadily growing (Jolley et al., 2018). The distribution of alleles among genes is not consistent – some genes are represented by only a few alleles, while other genes are highly diverse with a large number of alleles across all host sources. Most source attribution models have been developed using a small number of genes and using data based on alleles. These models have limited ability to discriminate between ruminant sources (Cody et al., 2019; Mughini-Gras et al., 2012; Mullner et al., 2009a) and between strains with low diversity across the modelled loci (e.g., ST474 in New Zealand) (Lake et al., 2021). Increasing the number of genes in a model will increase the diversity so that attribution can continue within a single sequence type (ST) and there will be greater discriminatory power for source attribution (Cody et al., 2017). The corresponding increase in number of genotypes is, however, generally problematic for source attribution models.

In this chapter the diversity and population structure of *Campylobacter* species in New Zealand is described, based on data from the Source Assigned Campylobacteriosis in New Zealand (SACNZ) study (Lake et al., 2021). The association of host sources with specific clonal complexes is then summarised and the limitations of source attribution models illustrated based on 7-gene multilocus sequence typing (7-gene MLST) data. Finally, the distribution of alleles among host sources is investigated and the frequency of unique alleles in the human isolates for prediction is examined. This is crucial in understanding the extent of the problem of unique genotypes caused by high resolution data and the importance of this thesis.

3.2 SACNZ Data Description

The data used in this thesis is from a source-assigned case-control study of notified human cases of campylobacteriosis in the Auckland and MidCentral Public Health Services regions of New Zealand between 2018-2019 (the SACNZ study) (Lake et al., 2021). *C. jejuni* and *C. coli* isolates were cultured from these human cases, as well as from poultry, sheep and cattle processors located within the Auckland and MidCentral District Health Boards. Whole genome sequencing (WGS) was carried out on the study isolates, with the microbiology and WGS procedures being described elsewhere (Lake et al., 2021). Following sequencing, draft genomes were assembled using the nullarbor2 pipeline[1] with default settings. Core-genome multilocus sequence typing (cgMLST) allele sequences were found by BLAST (Basic Local Alignment Search Tool) analyses (Altschul et al., 1990) against known alleles from the PubMLST *Campylobacter* database (Cody et al., 2017) and an allele number was assigned.[2] Previously found and novel alleles were then aligned using mafft (Katoh and Standley, 2013; Katoh et al., 2002).

The SACNZ dataset consists of 1211 isolates from four sources: cattle (n=168), chicken (n=205), sheep (n=187), and human (n=651). Each isolate has an allelic profile consisting of the pattern of alleles across 1343 genes. The allelic designation for each gene identifies the unique aligned sequence for a previously described allele or a novel allele sequence. In addition, each isolate has sequencing information (i.e., the exact sequence of nucleotides, recorded as A, T, C, G, or missing) of each allele, for each gene. In addition to the full cgMLST data described above, each isolate was also assigned a ST using the 7-gene MLST scheme for *C. jejuni* and *C. coli*. The 7-gene MLST targets loci within seven housekeeping genes, each approximately 500 bp in length with each unique nucleotide sequence being assigned an allele number from the PubMLST database (Jolley et al., 2018). This seven number allelic-based profile is assigned a ST and STs are then clustered into related groups called clonal complexes (CC), named after the central genotype ST (Dingle et al., 2002).

3.2.1 Source Isolate Species

Most of the isolates, from both humans and sources, were *Campylobacter jejuni* (human: 616/651 (94.6%), cattle: 139/168 (82.7%), sheep: 118/187 (63.1%), poultry: 172/205 (83.9%)), and the remainder were *Campylobacter coli*.

3.2.2 Distribution of 7-gene MLST Sequence Types by Host Source

There were 79 STs across the source isolates (figure 3.1) with a further 33 STs seen only in the human isolates. Some STs were associated with certain sources; for example, ST53, ST45, and ST50 were widespread and across multiple sources; whereas ST42, ST61, and ST3072 were strongly associated with ruminants; and ST6964, and ST48 with poultry (figure 3.1). A full description of the phylogeny of isolates from human cases and sources is found at Lake et al.

[1] https://github.com/tseemann/nullarbor
[2] https://github.com/jmarshallnz/cgmlst

(2020). A large proportion of STs were found in multiple host sources and attribution of these STs is challenging without additional information.

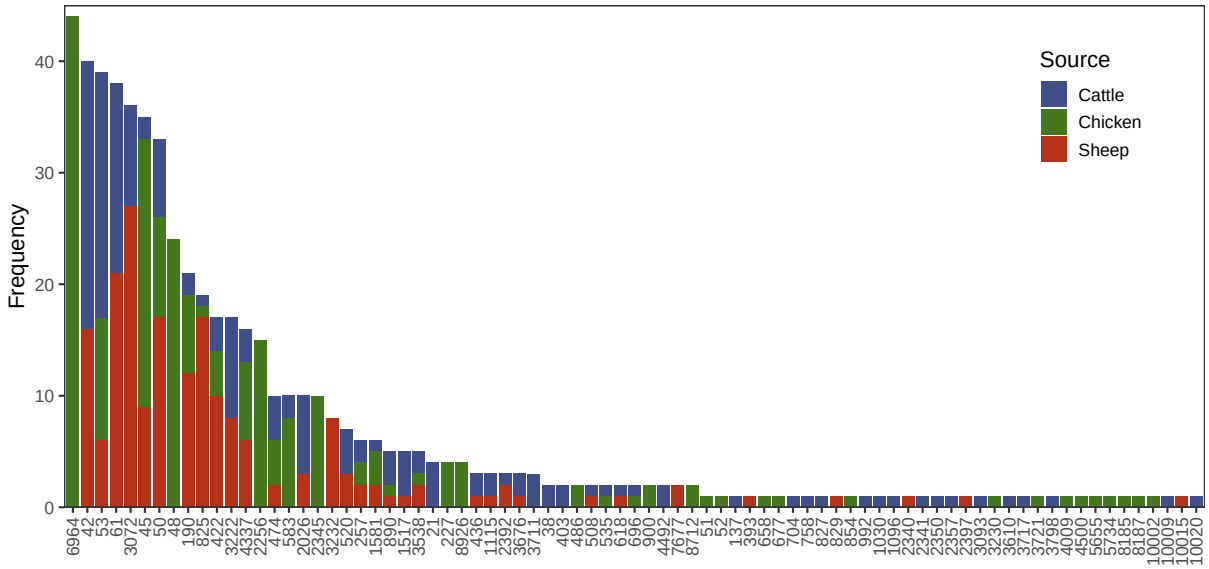

Figure 3.1: Distribution of 7-gene MLST sequence types by host source.

3.2.3 Distribution of cgMLST Sequence Types by Host Source

Although the set of genes in the dataset are defined by the cgMLST scheme as core genes (Cody et al., 2017), no isolate had the full complement of all 1343 genes. In theory, the cgMLST scheme defines a set of loci present in at least 95% of isolates. The current scheme is representative of *Campylobacter* isolates from the United Kingdom, Europe, and North America, and may not be representative of New Zealand isolates. In this study, 20 cgMLST genes were found in fewer than 95% of the isolates, and three cgMLST genes were found in fewer than 60% of the isolates. However, 1303 (97.0%) of the cgMLST genes were found in every isolate, and 1323 (98.5%) of the cgMLST genes were found in 95% of the isolates. The number of alleles per gene ranged from three to 467 (median 50) and the total number of alleles was 70,689. Of these, 49,424 were found in the animal isolates, and 54,718 were found in the human isolates. The aligned sequence length of the genes ranged from 95 to 4554 nucleotides (median 825). The number of nucleotides that differed between any pair of alleles (the Hamming distance) in aligned sequences ranged from one to 2595 (median 42).

With traditional source attribution models, for an isolate to be attributed to a source it needs to have alleles which are present in at least one source and which are not equally represented by all sources. Almost all of the isolates in the study differed by at least one allele across the genome - there were 1202 unique genotypes among the 1211 isolates, and no allele was shared by every isolate. 15,971 alleles (32.3%) were seen in the set of animal isolates but not in the set of human isolates. Among the animal isolates, 12,613 (25.5%) alleles were seen in all three animal sources, and 25,311 alleles (51.2%) were seen in only one single animal source; however,

of these, only 11,396 (45.0%) were also found in the set of human isolates, 17,570 (69.4%) were found in single isolates only, and 22,746 (89.9%) were found in five or fewer isolates (figure 3.2).

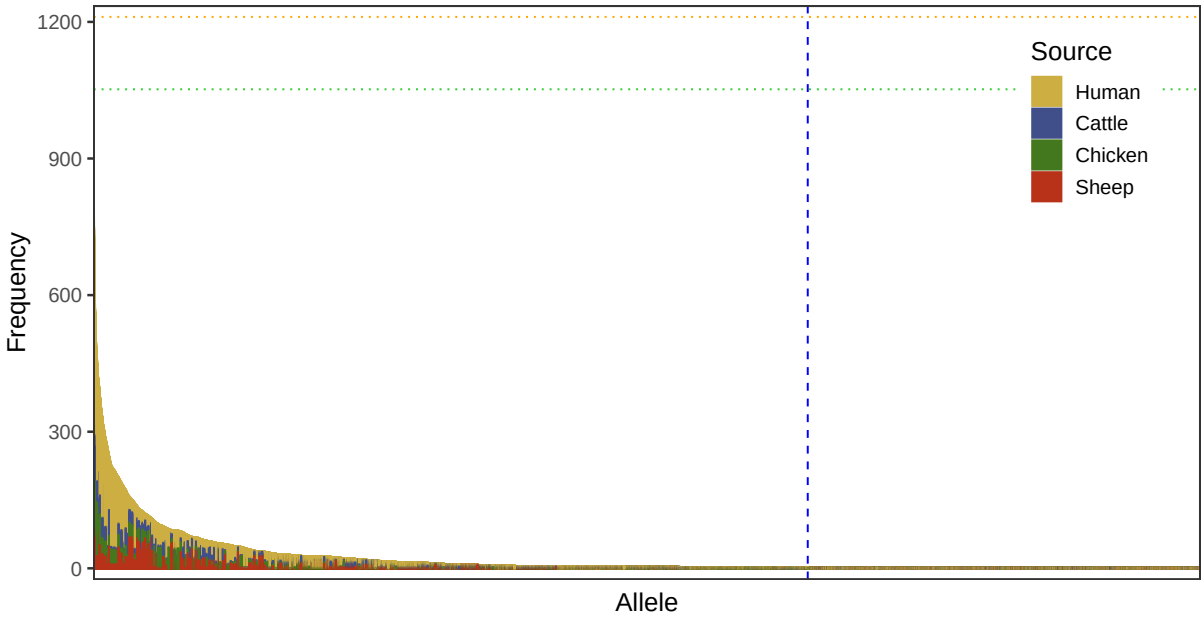

Figure 3.2: The number of instances of each SACNZ *Campylobacter* allele in each source. Most alleles were found in only one or a few isolates. The blue vertical line indicates the threshold for single isolates, the orange horizontal line indicates the total number of isolates, and the green horizontal line indicates the maximum number of isolates with any one allele.

Almost all isolates (167/168 cattle, 204/205 chicken, and 187/187 sheep) contained alleles unique to their respective source; however, in each case only a small number of isolates contained these source-specific alleles (4%, 23%, and 3% respectively). Most (96%) human isolates contained at least one allele that was not found in the animal isolates. Overall, the human isolates had 21,265 alleles (38.9%) that were specific to them (i.e., they were not seen in the set of animal isolates). The high number of unique alleles may be, at least in part, due to insufficient sampling. Rarefaction curves indicate that, for each source, sampling was not likely to have been sufficient to gather data on all the STs in the population (figure 3.3).

Although most of the isolates had only a small proportion of their genome that was unique, these unique alleles were distributed across almost all genes. Almost all (99.6%) genes contained at least one allele in the human isolates that was not found in the animal isolates (i.e., an 'absent level' – an allele that was not in the set of training data but is in the set of observations for prediction), and on average, for each gene, 28.6% of the alleles were unique and there were 30 human isolates with a unique allele. As 95.4% of isolates, and 99.6% of genes contain absent levels, removing affected isolates and/or genes is not a viable solution. At this resolution of data, the percentage of the genome without an absent level diminishes with increasing numbers of human isolates (figure 3.4), until conceivably every gene contains an allele unique to the human population. For source attribution methods that are unable to attribute unique alleles, this

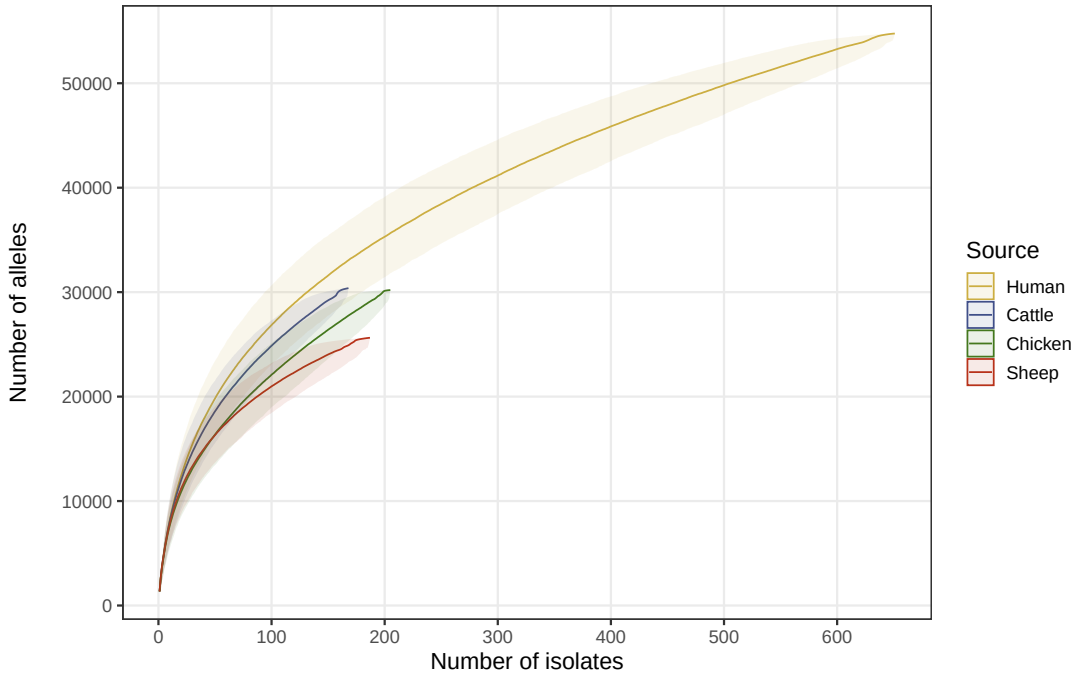

Figure 3.3: Rarefaction curves for alleles of *Campylobacter* isolates in New Zealand.

would mean there would be no suitable isolates remaining.

Higher resolution data, such as the nucleotide sequencing information, would allow a greater proportion of the genome to be used for source attribution. This is because there are five options for nucleotides (A, T, G, C, or missing) at each locus, compared with many hundreds, or even thousands, for the allelic data. There is, therefore, lower likelihood of a nucleotide being unique at each locus, or rather the collection of nucleotides at each locus is more likely to have been seen in the training data. Another option is to use a distance-based approach as the basis of a classification criteria for instances of unique genotypes (figure 2.10). A large number of the unique alleles differed from a known allele (i.e., one that was in the set of animal isolates) by only a single nucleotide (figure 3.4). Classifying alleles according to the number of different nucleotides (SNPs) from known alleles significantly increased the percentage of the core genome that is suitable for analysis (figure 3.4).

The genomic diversity of *Campylobacter* plays a role in the ability to differentiate among host sources. It is assumed that the allelic diversity of *Campylobacter* species among sources is greater than the diversity within sources. Measures of diversity aim to capture aspects of genomic structure within a host source. Five measures of diversity were calculated across each host source - (i) Richness (the total number of alleles per gene), (ii) Simpson's measure of evenness (the probability that two alleles drawn at random are from different genes), (iii) Pielou's evenness (the consistency of abundance of alleles over each gene), (iv) Chao's diversity (an estimate of richness extrapolated to account for different sample sizes), and (v) Shannon diversity index (a combined estimate of richness and evenness).

The diversity of *Campylobacter* genes (i.e., the variation in alleles) was affected more by

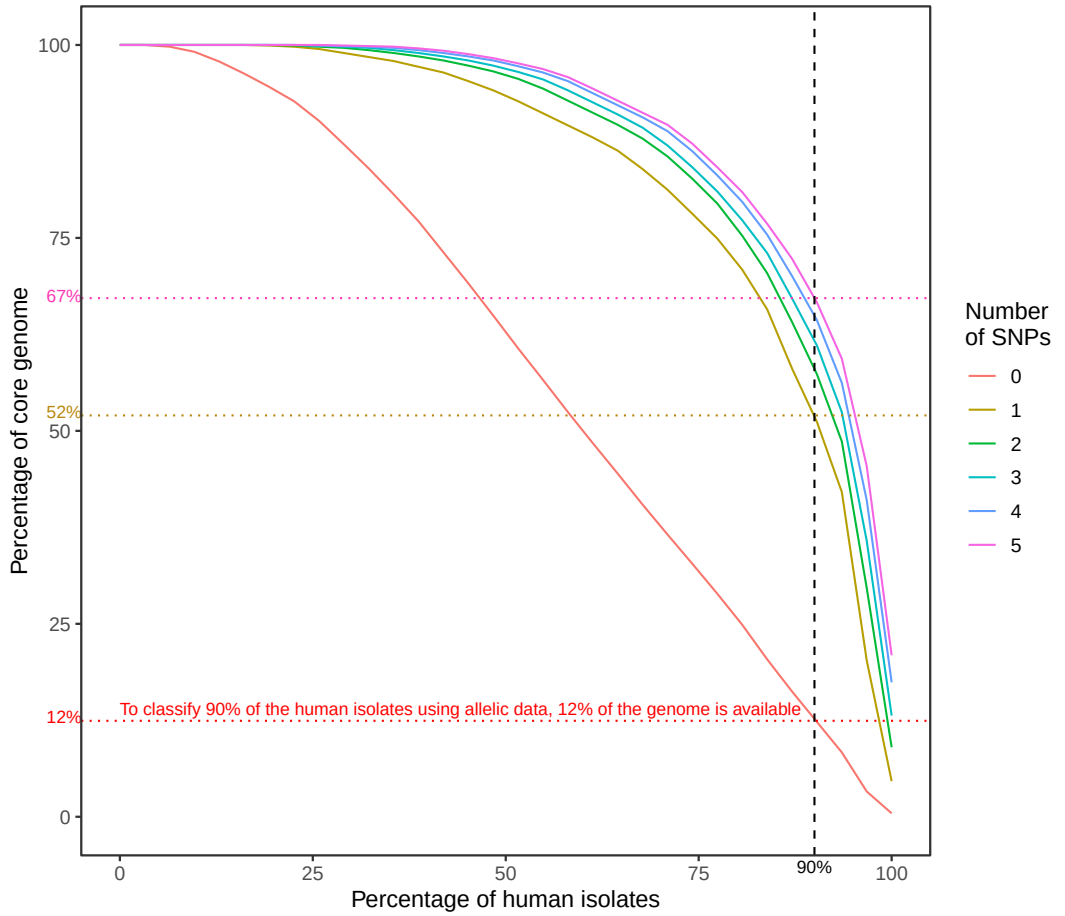

Figure 3.4: The percentage of the core genome available to classify human isolates, based on allelic data, and according to the number of nucleotide differences (SNPs) from a known allele. As the number of SNPs which differ from a known allele increases, the greater the percentage of genome that is available for source attribution models.

gene than by host source (figure 3.5). Despite having the smallest sample size, cattle had the highest number of STs (49) and alleles (30,356), followed by poultry (38 and 30,168 respectively), then sheep (32 and 25,609 respectively). Sheep had fewer alleles per gene, on average, than both cattle and poultry, however Chao's measure, which accounts for differences in sample size, suggests poultry may have the greatest number of alleles per gene, on average. The allelic profiles of the set of genes were also examined for correlation. As many expected cell counts were zero, the Chi-Squared test statistics cannot be reliably tested for significance, however, the actual test statistics (χ^2) for every pair of genes was large (range 141 to 198,604; with a median of 23,910) (figure 3.6), indicating a high level of linkage between loci.

To visualise patterns of variation in the isolates, the simple matching coefficient was calculated between pairs of isolates. The simple matching coefficient calculates the proportion of alleles at which two isolates match (including missing alleles) to the total number of alleles. This generated a similarity matrix which reflects the similarity and dissimilarity between the genomes of the individual isolates. A non-metric multidimensional scaling (nMDS) ordination,

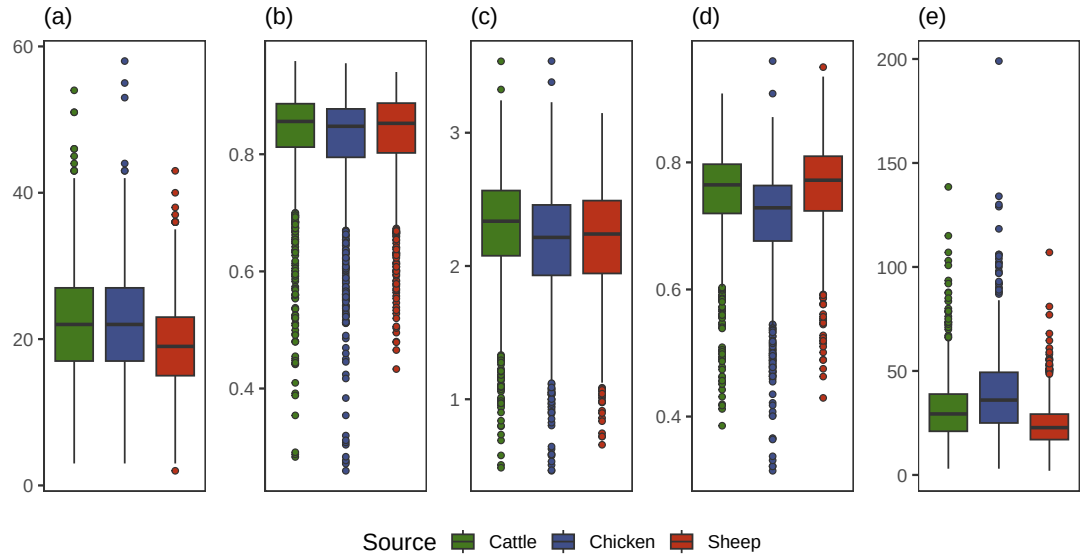

Figure 3.5: Diversity of *C. jejuni* and *C. coli* across each host source as measured by (a) Richness, (b) Simpson's measure of evenness, (c) Shannon diversity index, (d) Pielou's evenness, and (e) Chao's diversity measure for each gene.

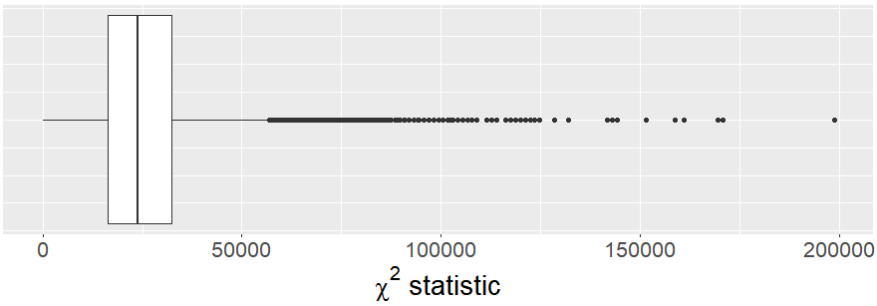

Figure 3.6: χ^2 test statistics for pairwise correlations of alleleic profiles of genes.

shows that the variation amongst the isolates is dominated by CC, which is nested in species, and there is no clear clustering of isolates according to source (figure 3.7).

Differences between allelic profiles of isolates from each of the three source groups were tested using (i) permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001) (including post-hoc pairwise tests), and (ii) a dissimilarity-based multivariate extension of Levene's test (PERMDISP) (Anderson, 2006), both using the simple matching coefficient. All three source groups were highly significantly different from one another ($p < 0.001$), although the amount of variation in the isolates which can be attributed to source is very small (6% of the total variation). There was no difference in the levels of dispersion for each of the source groups ($p > 0.8$). The small but significant difference between genomic profiles of isolates from different sources is not captured in the nMDS ordination (figure 3.7) – it is being obscured by the dominant effect of CC.

A dissimilarity-based multivariate analogue to a residual plot, from which the variation due

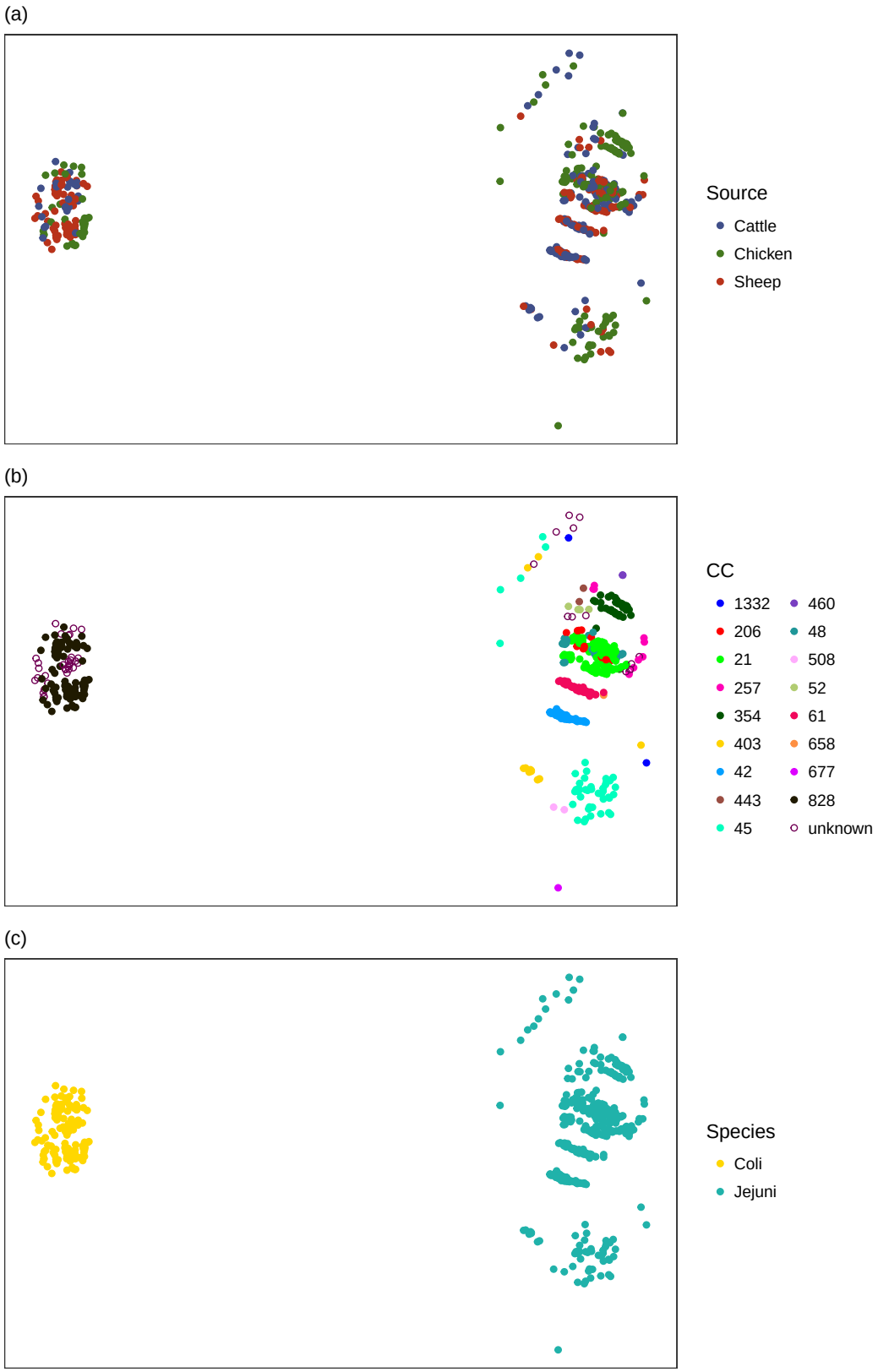

Figure 3.7: nMDS (2D stress = 0.06) of isolates based on the simple matching coefficient, coloured according to (a) source, (b) clonal complex, and (c) species.

to CC is removed, was calculated following the method described in Anderson (2017). An (N × N) "residualised" Gower matrix $\mathbf{G}^{[R]} = \{g_{ij}^{[R]}\}$ was calculated from $\mathbf{G}^{[R]} = (\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})$, where $\mathbf{H}$ is the "hat" matrix calculated on the model matrix $\mathbf{X}$ for CC (the effects of which are to be removed). A residualised distance matrix was then obtained as $\mathbf{D}^{[R]} = \{d_{ij}^{[R]}\}$ where $d_{ij}^{[R]} = \sqrt{g_{ii}^{[R]} - 2g_{ij}^{[R]} + g_{jj}^{[R]}}$. After removing the effect of CC, the difference in dispersion showed some difference between sheep and both beef ($p = 0.06$) and poultry ($p = 0.03$) but poultry and beef were not significantly different from each other ($p = 0.96$). The nMDS ordination was repeated using the residualised distance matrix (figure 3.8). This still failed to show any clear pattern relating to source, although the stress is high (0.22), even in three dimensions (stress = 0.17).[3]

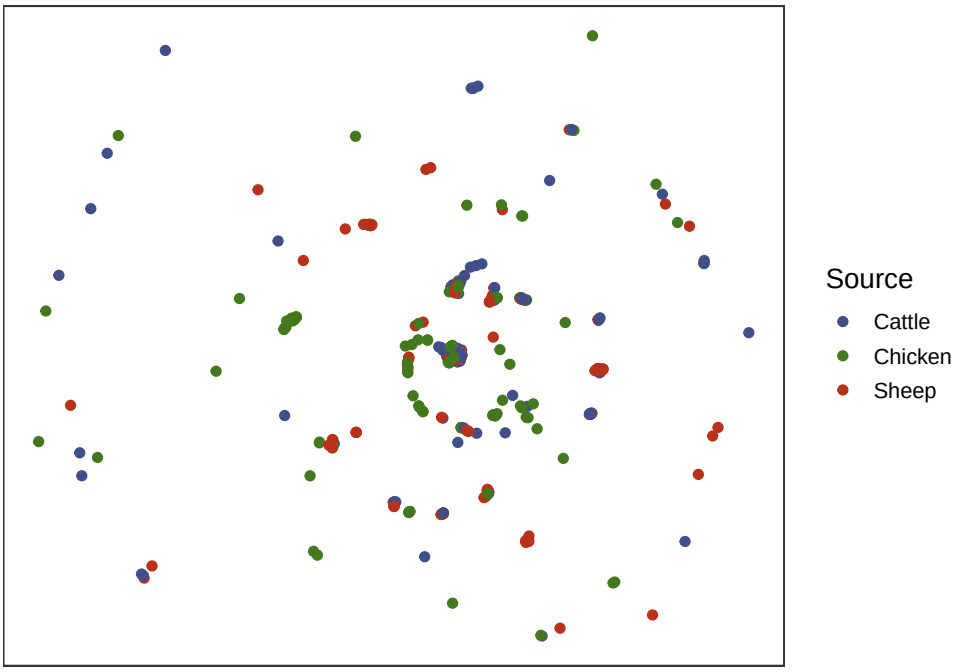

Figure 3.8: nMDS (2D stress = 0.22) of isolates showing clustering of isolates according to source, with the effect of clonal complex removed.

3.3 Implications for Source Attribution Models

C. jejuni and *C. coli* are highly diverse organisms with new genotypes regularly being identified (Jolley et al., 2018). The genomic profiles of isolates from different sources do differ, however, the variation is dominated by species and CC, and the variability attributed to host source is

[3]Stress gives a measure of how accurately the high-d relationships are represented in the low-d ordination. A rule of thumb is that stress < 0.1 is good, and stress > 0.3 indicates that the points are arranged in a random order.

small. Because *Campylobacter* shows evidence of both clonal relatedness through substitutions and panmixis through frequent recombination events, variation among isolates has been traditionally classified at the allelic level (i.e., rather than at the nucleotide level), as a large number of different nucleotides may be due to a single homologous recombination event. The allelic diversity, however, appears to be greater across genes within the genome than across host sources. Source attribution models that use only a small number of genes are limited by the diversity of the loci in the model.

A logical solution is to increase the number of genes in the model, however, this results in a large proportion of the human isolates having a unique genotype (i.e., being distinct from the genotypes found in the animal sources). The counts of each genotype decrease with each additional gene, and with whole genome data many genotypes are found in single isolates only. Currently, almost every host has source-specific alleles, but these are either not found in human isolates, or are found in very few isolates, so they are not useful for frequency-matched source attribution models. In addition, the large number of genes means a high level of linkage is observed between loci, rendering the population genetics model STRUCTURE (Pritchard et al., 2000) unsuitable.

Using higher resolution data, such as nucleotide sequencing information, reduces the potential for unique presentations. At such high resolution, however, the quantity of data increases by several orders of magnitude. For example, an organism with 1000 genes, and 1000 alleles per gene, and 1000 nucleotides per allele yields 1 million nucleotides for analysis. This means any errors are magnified and the resulting variation can obscure any patterns being sought by the analysis.

Another approach is to use the higher resolution sequencing information as the basis of a distance-based measure with which to compare unique alleles to known alleles. In combination with random forest, this novel approach is explored in chapters four and five of this thesis.

Chapter 4

Lost in the Forest - New Methods of Encoding Categorical Predictors

4.1 Introduction

A classification tree is a method of supervised machine learning that predicts a categorical response variable by way of a series of binary decisions. Each decision, or split, is made based on a single predictor variable to maximise predictive accuracy with respect to the response variable. Individual classification trees tend to overfit to the training data, that is, they yield decision rules that are more specific to the training data than they are to new independent data. Random forest is a tree-based algorithm that addresses this issue by creating an ensemble of classification trees. The individual trees that make up the ensemble differ from one another because they are each trained on a different random sample of the cases ('bagging') and predictor variables ('random subspacing'; Amit and Geman, 1997; Breiman, 1996; Ho, 1998). The predictions from the individual trees are aggregated and classifications are made based on the majority vote across the trees.

4.1.1 The 'Absent Levels' Problem

An inherent issue with tree-based predictive models occurs when a level of a categorical predictor variable is absent when a tree is grown, but is present in a new observation for prediction (the 'absent levels' problem *sensu* Au, 2018). In a random forest, absent levels can arise due to sampling variability (i.e., the level was absent from the observations that were used to train the model), bagging (i.e., the level was in the training data but absent from the bootstrapped sample used by a particular tree), or partitioning of the data by the tree (i.e., the level was present at the top of the tree but absent from a lower subset created by binary splits). When the algorithm encounters an absent level, there is no immutable *a priori* rule for determining which side of the binary split the observation should go. When this happens, the observation is effectively 'lost in the forest'.

Missing data heuristics allow the random forest algorithm to proceed with an absent level.

Methods include stopping an affected observation from proceeding down the tree (Therneau et al., 2022), using a surrogate decision rule that mimics the original split's partitioning (i.e., surrogate splits) (Hothorn and Zeileis, 2015; Therneau et al., 2022), directing all affected observations down the branch with the most training observations (Hothorn and Zeileis, 2015), directing all affected observations down both branches simultaneously but weighted according to the number of observations from each child node (e.g., distribution-based imputation (DBI)) (Quinlan, 1993; Saar-Tsechansky and Provost, 2007), and randomly directing affected observations down a left or right branch (Hothorn and Zeileis, 2015). The `scikit-learn` Python module's implementation of random forest (Pedregosa et al., 2011) treats absent levels as missing values. If missing levels are present in the training data then absent levels get assigned to an explicit missing category, otherwise they get mapped to the child node that has the most samples.[1]

It is important to distinguish between absent levels and missing data, however. Unlike missing data, absent levels are fully observed and known. Treating an observation with an absent level as though it were missing data necessitates a loss of information and is not recommended (Ishwaran et al., 2008). Au (2018) thoroughly investigated the properties of missing data heuristics with random forest models and compared them to the naïve heuristic of directing all observations with absent levels down the same branch (i.e., 'left' or 'right' heuristic) as is implemented in many random forest applications (Liaw and Wiener, 2002; Wright and König, 2019). Au showed that the choice of heuristic can dramatically alter a model's performance and potentially lead to systematic bias in prediction. Decision tree-based methods are widely used; it is almost certain that a number of these models have been inadvertently affected by the absent-levels problem in practice. To date, there remains no compelling solution for dealing with absent levels in random forest models.[2]

4.1.2 Variable Encoding

The key to dealing with absent levels lies in how categorical variables are encoded. The random forest algorithm can, in theory, process categorical variables in their raw state, comparing all $2^{k-1} - 1$ possible binary splits for a nominal predictor variable with k distinct levels. There are, however, significant potential gains in efficiency from imposing an order on a nominal predictor variable. An ordered categorical predictor with k levels can be treated the same way as a numerical predictor with k unique ordered values. This reduces the number of potential partitions from $2^{k-1} - 1$ to $k - 1$ and the allocation of each level to one side of the binary split is constrained only by whether it is above or below the split point. There are several approaches used to encode, or convert, categorical variables into numerical format for analysis.

Integer encoding (also called label encoding) is the simplest method of encoding. For each categorical variable X, with k distinct values, the observed levels are mapped to the integers one to k and new levels, which were not observed during training, are encoded as missing values.

[1] https://scikit-learn.org/stable/modules/ensemble.html#random-forests
[2] https://github.com/imbs-hl/ranger/issues/94

A major issue with this method is that despite there being no intrinsic relationship between the levels and the numbers being used to replace them, an ordering ($1 < k$) is imposed.

Indicator encoding (also called one-hot encoding) avoids imposing an order on a nominal categorical variable. Each categorical variable X, with k distinct values, is transformed to k binary indicator variables and observations are encoded to indicate the presence (1) or absence (0) of the dichotomous variable. Indicator encoding removes any uncertainty over where to send an observation with an absent level as they can be encoded with a zero vector. However, it can result in the dataset becoming very wide and sparse, which in turn can present computational challenges and inconsistent results (Au, 2018; Cerda et al., 2018; Hastie et al., 2009; Reilly et al., 2022). With indicator encoding, the feature importance of the original variable is distributed among separate binary variables which may cause bias for tree based algorithms as the impurity reduction induced by a single indicator is rarely enough to be selected for splitting. Dummy encoding, i.e., indicator encoding with $k - 1$ categories, has similar properties.

Target-based encoding methods differ from integer encoding and indicator encoding in that they incorporate information about the target values associated with a given level. For the case of two-class (binary) classification, ordering a nominal predictor by the proportion of observations with the second response class in each level leads to identical splits in the random forest optimisation as considering all possible 2-partitions of the predictor levels if the encoding is repeated at every split (Breiman et al., 1984; Fisher, 1958). Two popular software implementations for random forest, the `randomForest` and `ranger` R packages, adopt this optimisation. For multiclass classification, there is no available sorting algorithm that leads to splits which are equivalent to considering all $2^{k-1} - 1$ possible partitions (Wright and König, 2019). The R package `ranger` (Wright and Ziegler, 2017) offers a target-based encoding method that encodes each predictor variable according to the first principal component of the weighted covariance matrix of class probabilities, following Coppersmith et al. (1999).[3] For computational efficiency the encoding of the predictor variables occurs once on the entire dataset prior to bagging. In each of these methods absent levels are encoded with the highest rank, effecting the 'right' heuristic.

4.1.3 Encoding of Absent Levels

In addition to reducing computational complexity, ordinal encoding of predictor variables allows absent levels to be encoded, integrated with existing levels, and subsequently used for prediction, thereby circumventing the absent-levels problem. The `randomForest`[4] and `ranger` R packages encode absent levels with the highest rank (equivalent to integer encoding as $k + 1$), which ensures observations with an absent level will always 'go right', as per the 'right' heuristic for missing data. Assigning all observations with absent levels to the same branch will keep the observations together as a collection which can be split further down the tree by another variable. However, this heuristic, when combined with target encoding, leads to systematic bias

[3] Coppersmith, Hong & Hosking (1999) use the first principal component of the weighted matrix of class probabilities.

[4] Update 4.6-10 allows absent levels to be encoded if the categorical variable is an ordered factor. Categorical variables of type 'character' are converted to ordered factors, with the order determined alphabetically.

towards the first response class (Au, 2018). Furthermore, classifications for observations with absent levels can be influenced by interchanging the order of the response classes. Au (2018) therefore argued that observations with absent levels should be assigned randomly to a left or right branch as this reduces the systematic bias in prediction. There has been no documented investigation to date into the properties of this heuristic in the multiclass response case, however.

Here, various methods of dealing with high cardinality, nominal predictor variables in the context of random forest models and the absent-levels problem are examined. This chapter details how target-agnostic *versus* target-based encoding predictor variables with absent levels affects the accuracy of random forest models, and it presents two alternate methods for encoding predictor variables and/or absent levels. The prediction accuracy of these methods is then examined using a case study on source attribution of *Campylobacter* species using whole genome sequencing (WGS) data as predictors. The WGS data generates allele profiles based on unique nucleotide sequences for each gene in the chromosome.

More specifically, the aim of this chapter is to:

(i) assess the misclassification rate of multiclass random forest predictions when nominal predictor variables are target encoded and observations with absent levels are sent to the right side of a binary split, using real data from a published source-assigned case-control study;

(ii) compare the misclassification rate from (i) *versus* that of predictions when observations with absent levels are sent to a left or a right branch of a split according to the *a priori* hypothesis of equal class probability;

(iii) introduce the PCO-encoding method for ordinal encoding categorical predictors that makes use of ancillary information on the levels of predictor variables.

4.2 Methods

4.2.1 Random Forest

For a training set of N independent observations on P variables, where $x_n = (x_{n1}, x_{n2}, \ldots, x_{nP})$ is the vector of predictor variables for observation $n = 1, 2, \ldots, N$, and y_n is the corresponding response variable, classification and regression tree (CART) is a greedy recursive binary partitioning algorithm that successively partitions data (the parent node) into two smaller subsets (the left and right child nodes). Each partition is determined based on a decision rule for a single predictor variable to maximise predictive accuracy with respect to the response variable (Breiman et al., 1984). In a random forest, each individual tree is trained on a bootstrap resample of the training data ('bagging') using a randomly selected subset of the P predictors ('random subspacing'; Amit and Geman, 1997; Breiman, 1996; Ho, 1998), and is traditionally not 'pruned'. A classification can be predicted for a new observation by sending it down each

tree according to the decision rules until it arrives at a terminal node, then aggregating the tree predictions and taking the majority vote across the forest. Various control parameters can be set for random forest models, including the number of trees, the number of variables randomly selected as splitting candidates, and tree size (Wright and Ziegler, 2017).

4.2.1.1 Out-of-Bag Sample

Bootstrap aggregating, or bagging, in random forest sees each individual tree trained on a subset of the observations in the training set generated by subsampling with replacement. Correspondingly, for each tree there is a sample of observations that are not used for training - the out-of-bag (OOB) sample. Aggregating the predictions from the observations in the OOB sample can be used to generate an OOB prediction for each observation; the misclassification rate of OOB predictions for all training observations is the OOB error (Breiman, 2001). Breiman (1996, 2001) claimed that the OOB error alleviates the need for cross-validation or setting aside a separate test set, however, at least for two-class classification problems with numerical predictor variables, this is disputable (Janitza and Hornung, 2018; Mitchell, 2011).

4.2.2 Encoding of Categorical Predictor Variables

The method of encoding predictor variables can affect the performance of random forest (Au, 2018; Wright and König, 2019). For categorical features with a small number of levels, target-based encoding has been shown to achieve better results than one-hot encoding and integer encoding (Wright and König, 2019), however as a result of using the target variable, information leakage and overfitting is a concern. By using the probability of the target for encoding, there is information leakage from the target variable to the predictors. Further, if a predictor is encoded prior to splitting into training and testing sets, information from the target variable in the test set will leak to the predictors in the training set by way of the *a priori* encoding, which will impact cross-validation errors. When the encoding occurs prior to bagging (i.e., rather than each subsample undergoing encoding independently) the OOB errors will be similarly affected. A separate test set that is not used to inform the encoding will be a more reliable estimate of model performance.

For multiclass classification problems, three methods for encoding categorical predictor variables as ordered factors or continuous variables are considered:

4.2.2.1 Correspondence Analysis (CA) Encoding Method

The CA-encoding method is a target-based encoding method which performs a scaled correspondence analysis on the contingency table of counts of variable levels by class, following the approximation of Coppersmith, Hong, and Hosking (1999).[5] Each predictor variable is encoded according to the first principal component of the weighted matrix of class probabilities

[5]The `"ordered"` method in `ranger` performs a PCA on the weighted covariance matrix of class probabilities rather than on the weighted matrix of class probabilities, yet the results are equivalent.

and absent levels are encoded with a principal component score of infinity. This ensures all observations with an absent level branch as a group and always (i.e., at each node) in the same direction ('go right') (figure 4.1, a).

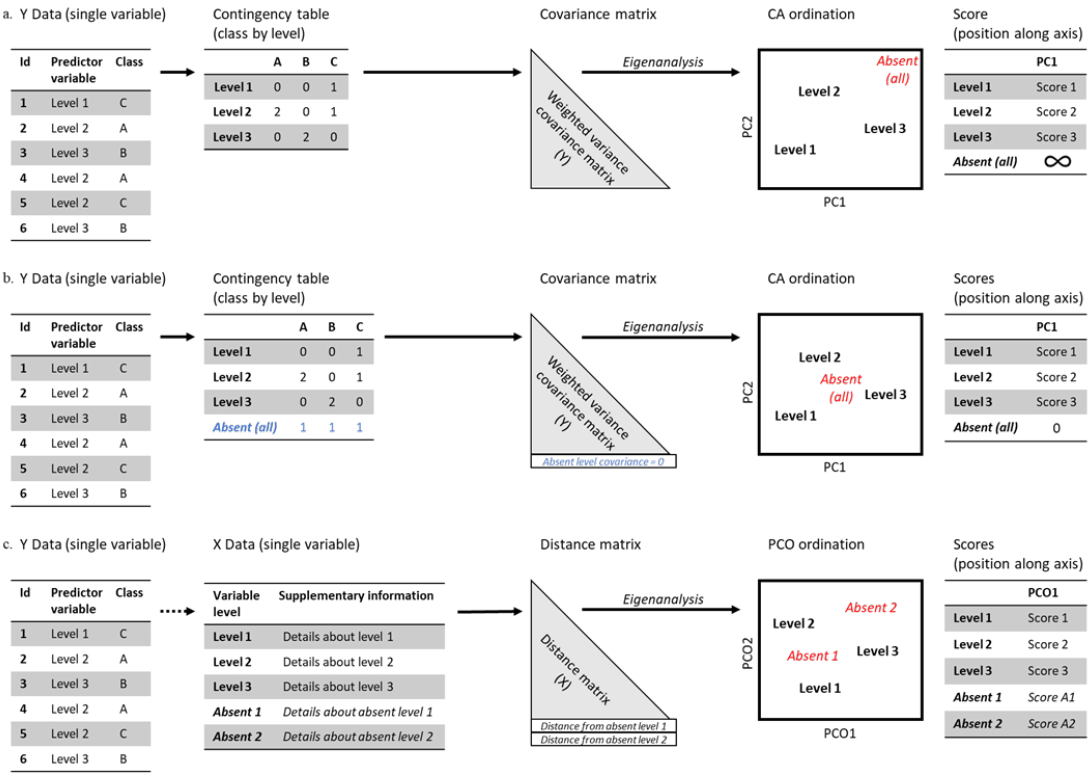

Figure 4.1: A visual description of the three methods described in this chapter (a) CA-encoding method - the levels of each predictor variable are ordered according to the first principal component of the class probabilities and absent levels are assigned a score of infinity; (b) CA-unbiased-encoding method - the levels of each predictor variable are ordered according to the first principal component of the class probabilities and absent levels are assigned a score of zero based on *a priori* equal class probabilities; blue text indicates conceptual information for an absent level; (c) PCO-encoding method - the levels of each predictor variable, including absent levels, are ordered according to their score for the first principal coordinate axis derived from ancillary pairwise distance information.

4.2.2.2 CA-unbiased Encoding Method

The difference between the CA and CA-unbiased methods of encoding lies in the treatment of absent levels. The novel CA-unbiased-encoding method encodes any absent level with a principal component score of zero (figure 4.1, b). This aligns with the assumption that any level of the predictor variable that is absent from the training data is *a priori* equally likely in any class and has equal class probabilities of $1/Y$, where Y is the number of classes. Because all absent levels will have equal class probability vectors, they can be combined into a single attribute value (Coppersmith, Hong, and Hosking, 1999). Then, because the class probabilities are not independent of each other, the sum of the principal component coefficients is zero and it follows

that the principal component score of an absent level with equal class probabilities will be zero. At some splits, the zero principal component score will fall on the left side of the splitting value and at other splits it will fall on the right side. In the unlikely case of a splitting value being exactly zero, all observations would be sent to the left. To account for this, the scores have a small degree of noise added so that observations will be randomly sent to either the left or right branch with equal probability in the case of a splitting value being exactly zero.

4.2.2.3 Principal Coordinates Analysis (PCO) Encoding Method

The PCO-encoding method is a target-agnostic ordinal encoding method which relies on ancillary information on the individual levels of predictor variables. For example, a categorical variable consisting of city names has ancillary information that includes latitude and longitude, as well as population-based information. The PCO-encoding method utilises the ancillary information, rather than the level names *per se*. The choice of ancillary variables used for the distance calculation will depend on how the association between levels should be defined, e.g., geographical *versus* social *versus* economic etc. This will determine the degree of similarity and how an absent level will be treated.

In the correspondence analysis methods above, the eigenanalysis step is performed on the weighted level by class contingency table and the score is the coefficient for the corresponding predictor level of the first principal component. In comparison, the eigenanalysis in the PCO-encoding method is performed on a distance matrix of the set of predictor levels extracted from ancillary information on the levels of predictor variables, and the score is the principal component score for the corresponding predictor level for the first principal coordinate (figure 4.1, c). The PCO-encoding method relies on ancillary information for each of the predictor variables, independently, in order to generate a set of matrices of dissimilarities. Principal coordinates analysis (PCO) (Gower, 1966) is then applied to this distance matrix, yielding a ρ-dimensional ordination of levels in Euclidean space. A single dimension (i.e., only the first principal coordinate) for each variable was chosen to maintain consistency between methods for comparison, however any number of dimensions could potentially be used. Using the method of Gower (1968), a new (absent) level can be interpolated into the ρ-dimensional space by virtue of the interpoint distances between this level and each of the levels in the training set. This then generates a score for each new level, and allows it to branch according to its resemblance to other levels in the training data.

4.2.3 Comparison of Encoding Methods

4.2.3.1 Source Attribution

The process of assigning cases of human zoonotic infectious diseases to their most likely origin is known as source attribution. Because of their role in human gastroenteritis, *Campylobacter jejuni* and *C. coli* have been the subject of a large number of source attribution studies using a variety of approaches, including epidemiological methods (Domingues et al., 2012; Pires et al.,

2010), comparative risk and exposure assessment (Pintar et al., 2017), expert knowledge elicitation (Hald et al., 2016; Havelaar et al., 2008), and microbiological methods (Arning et al., 2021; Brinch et al., 2023; Hald et al., 2004; Liao et al., 2019; Miller et al., 2017; Mullner et al., 2009a; Sheppard et al., 2009; Strachan et al., 2009). Microbiological methods of source attribution rely on comparing the phenotypic or genotypic profiles of human cases of infection with those of animal sources. Although many earlier studies have used just a small number of loci (targeted part of a gene in the bacterial chromosome) within the genome (< 10), the availability of next-generation sequencing (NGS) has greatly increased the number of loci available for analysis.

Models that use allelic-profile data arising from bacterial WGS have a high number of categorical predictors, which are often subject to the absent-levels problem. *Campylobacter* species are genomically very diverse and, although the allelic diversity (i.e., sequence variability within a gene) is inconsistent across the genome, some loci are highly variable (Parkhill et al., 2000; Sheppard and Maiden, 2015). *C. jejuni* and *C. coli* each have a circular chromosome, roughly 1.7 Mb long (Chen et al., 2013; Parkhill et al., 2000; Pearson et al., 2013; Taylor et al., 1992) which encodes for approximately 1700 genes (Parkhill et al., 2000). A core-genome multilocus sequence typing (cgMLST) scheme has been defined jointly for these species which contains a set of 1343 loci which are present in most ($\sim$ 95%) members of human *C. jejuni* and *C. coli* isolates (Cody et al., 2017). In any given dataset, an isolate will contain nearly all of the genes in this scheme, however the observed alleles of each gene are commonly found in only one or a few isolates. This means that there are many alleles across the genome which would be unique to individual collections of isolates from human and animal datasets.

4.2.3.2 Dataset

The Source Assigned Campylobacteriosis in New Zealand (SACNZ) study is a source-assigned case-control study of notified human cases of campylobacteriosis in the Auckland and MidCentral District Health Board regions, New Zealand, between 2018-2019 (Lake et al., 2021). *C. jejuni* and *C. coli* isolates were cultured from these human cases, as well as from poultry, sheep, and cattle processors serving the Auckland and MidCentral District Health Boards. WGS was carried out on the study isolates, with the microbiology and WGS procedures being described elsewhere (Lake et al., 2021). Following sequencing, draft genomes were assembled using the nullarbor2 pipeline[6] with default settings and cgMLST allele sequences were found by BLAST (Basic Local Alignment Search Tool) analyses (Altschul et al., 1990) against known alleles from the PubMLST *Campylobacter* database (Cody et al., 2017). Previously found and novel alleles were aligned using mafft (Katoh and Standley, 2013; Katoh et al., 2002) and an allele number assigned.[7]

The SACNZ dataset consists of 1211 isolates from four sources: cattle (n=168), chicken (n=205), sheep (n=187), and human (n=651). Each isolate has an allelic profile consisting of

[6]https://github.com/tseemann/nullarbor
[7]https://github.com/jmarshallnz/cgmlst

the pattern of alleles across 1343 genes. The allelic designation for each gene identifies the unique aligned sequence for a previously described allele or a novel allele sequence. More simply, the categorical predictor variables are genes with alleles as levels. The ancillary data is the sequencing information (i.e., the exact sequence of nucleotides, recorded as A, T, C, G, or missing) of each allele, for each gene. This ancillary nucleotide sequencing information is used to calculate a matrix of Hamming distances (Li and Jain, 2009) between each pair of alleles within each gene.

4.2.3.3 Cross Validation

The 651 isolates collected from humans were excluded from analysis because their true animal source was unknown, and the remaining 560 isolates were subject to ten-fold cross-validation for each of three methods (CA-encoding, CA-unbiased-encoding, and PCO-encoding) using the same random number seed. Across the methods, the forest consisted of 500 trees and the Gini index was used as the splitting criterion. For each method, ten independent random forest models were run (one on each of the ten folds) allowing each of the 560 isolates to be represented exactly once in testing data. Model performance was assessed by calculating the proportion of incorrect classifications on the set of test data for each fold and calculating the average and standard error, accounting for any variation between folds. Thus 560 isolates of known source were classified by a random forest model containing 500 trees resulting in 280,000 individual tree predictions for each method. To assess the effect of absent levels on classification success the number of absent levels used by each tree for prediction was recorded in addition to the individual tree predictions.

The order of analyses was as follows (see also figure 4.1):

1. create training and testing data
    - split the data into ten folds
    - select nine of the ten folds for a set of training data and the remaining tenth fold for a set of testing data
    - repeat until ten unique sets of training data and testing data have been created for each set and continue to 2.
2. prepare training data
    - create a level by class (i.e., allele by source) contingency table (CA-encoding, CA-unbiased-encoding methods)
    - encode each variable via principal component analysis (PCA) on the (weighted) contingency table (CA-encoding, CA-unbiased-encoding methods)
    - encode each variable via PCO on an ancillary set of data matched to the training data (PCO-encoding method)
3. fit the model on the prepared training data

4. prepare testing data
    - identify levels that are unique to the testing data (i.e., absent levels)
    - encode levels that are in the training data with the variable score from 2.
    - encode absent levels
        - with a score of infinity (CA-encoding method);
        - with a score of zero (CA-unbiased-encoding method)
        - with new scores via Gower's method (Gower, 1968) on ancillary data matched to the testing data (PCO-encoding method)
5. predict each test observation
    - identify individual tree predictions
    - identify trees that branched on an absent level

4.2.3.4 Code Availability

All analyses were carried out using R version 4.3.0 (R Core Team, 2024) and the `ranger` package ('RANdom forest GEneRator') (Wright and Ziegler, 2017). The R code used in this study is available at https://github.com/smithhelen/LostInTheForest. The sequence reads used for this study can be accessed from the NCBI Sequence Read Archive under BioProject accession number PRJNA675916.

4.3 Results

4.3.1 Genome Description

Of the 560 isolates, there were 558 distinct allelic profiles (i.e., only two isolates shared an identical set of alleles with another isolate and the remaining isolates differed by at least one allele across the core genome). The number of alleles per gene ranged from one to 222 (median 35) and the total number of alleles was 49,424. Across all 1343 genes, 25,317 alleles (51.2%) were seen in only a single source, and 17,575 alleles (35.6%) were seen in only a single isolate. 167/168 (99.4%) of the cattle isolates, 204/205 (99.5%) of the chicken isolates, and 187/187 (100%) of the sheep isolates contained alleles unique to their respective source. The unaligned sequence length of the genes ranged from 95 to 4554 nucleotides (median 816). The number of nucleotides that differed between any pair of alleles (the Hamming distance) in aligned sequences ranged from one to 2595 (median 42).

4.3.2 Random Forest Results

At least 90% of the random forest predictions, from any method, used at least one absent level for classification, and approximately one fifth (16% (PCO); 22% (CA and CA-unbiased)) of individual tree predictions used at least one absent level. The frequency of absent level use in

Table 4.1: Weighted average proportion and standard error of all tree predictions assigned to each of three host sources (cattle, chicken and sheep) for each of three methods of encoding categorical predictors.

Source	**Prediction**	**Method of encoding**		
		CA	CA-unbiased	PCO
Cattle	Cattle	0.605 ± 0.021	0.684 ± 0.021	0.622 ± 0.021
Cattle	Chicken	0.109 ± 0.044	0.127 ± 0.041	0.134 ± 0.046
Cattle	Sheep	0.310 ± 0.022	0.208 ± 0.022	0.267 ± 0.021
Chicken	Cattle	0.096 ± 0.023	0.118 ± 0.024	0.087 ± 0.023
Chicken	Chicken	0.797 ± 0.021	0.831 ± 0.021	0.842 ± 0.020
Chicken	Sheep	0.108 ± 0.020	0.064 ± 0.020	0.073 ± 0.019
Sheep	Cattle	0.179 ± 0.020	0.166 ± 0.019	0.172 ± 0.020
Sheep	Chicken	0.057 ± 0.031	0.065 ± 0.030	0.072 ± 0.030
Sheep	Sheep	0.777 ± 0.020	0.784 ± 0.021	0.772 ± 0.020

predictions varied considerably among individual trees and forests for all methods. The CA-unbiased-encoding methods used absent levels up to 24 times in a single tree, compared with 19 for the PCO-encoding method and 12 for the CA-encoding method. On average, a variable with absent levels was used for a single classification between 4.7 times (PCO) and 7.5 times (CA-unbiased) but fewer than 4% of trees, from any method, used a variable with absent levels more than once for a single tree prediction.

The ten most important predictor variables (genes) as measured by the permutation variable importance approach (Breiman, 2001) varied between methods. CA-encoding and CA-unbiased-encoding methods identified the same 10 genes, in identical order. Of these ten only one was identified by the PCO-encoding method.

4.3.3 Classification Accuracy

The CA-unbiased-encoding method had the lowest average misclassification error (23.2% ± 1.2%), followed by the PCO-encoding (25.2% ± 1.2%), and the CA-encoding (27.0% ± 1.2%) methods. The accuracy of predictions was dependent on the class being predicted (table 4.1, figure 4.2). Across the methods, isolates sourced from chicken were the most accurately classified (79.7% ± 2.1% − 84.2% ± 2.0%); isolates that were incorrectly classified were evenly distributed between sheep and cattle. Isolates sourced from sheep were the second most accurately classified for all methods (77.2% ± 2.0% − 78.4% ± 2.1%); incorrectly classified isolates were mostly assigned to cattle (16.6% ± 1.9% − 17.9% ± 2.0%) with fewer than 8% being assigned to chicken. Isolates sourced from cattle had the lowest classification success rates (60.5% ± 2.1% − 68.4% ± 2.1%), with most of the incorrect classifications predicted as sheep (20.8% ± 2.2% − 31.0% ± 2.2%) rather than chicken (10.9% ± 4.4% − 13.4% ± 4.6%).

4.3.4 Effect of Absent Levels

The class frequencies of predictions were similar across all methods when no absent levels were used for the predictions (figure 4.2). When absent levels were used for predictions, the predictions were not equally distributed across the three sources and the pattern of distribution depended on the method. For all methods, the class distribution followed the pattern of distribution for predictions made without absent levels, whereby incorrect chicken predictions were split between cattle and sheep; incorrect sheep classifications favoured cattle; and incorrect cattle classifications favoured sheep, but with a lower proportion of correct predictions in any class (figure 4.2). The accuracy of predictions also decreased as the number of absent levels in a tree increased (figure 4.3, see also appendix A.2).

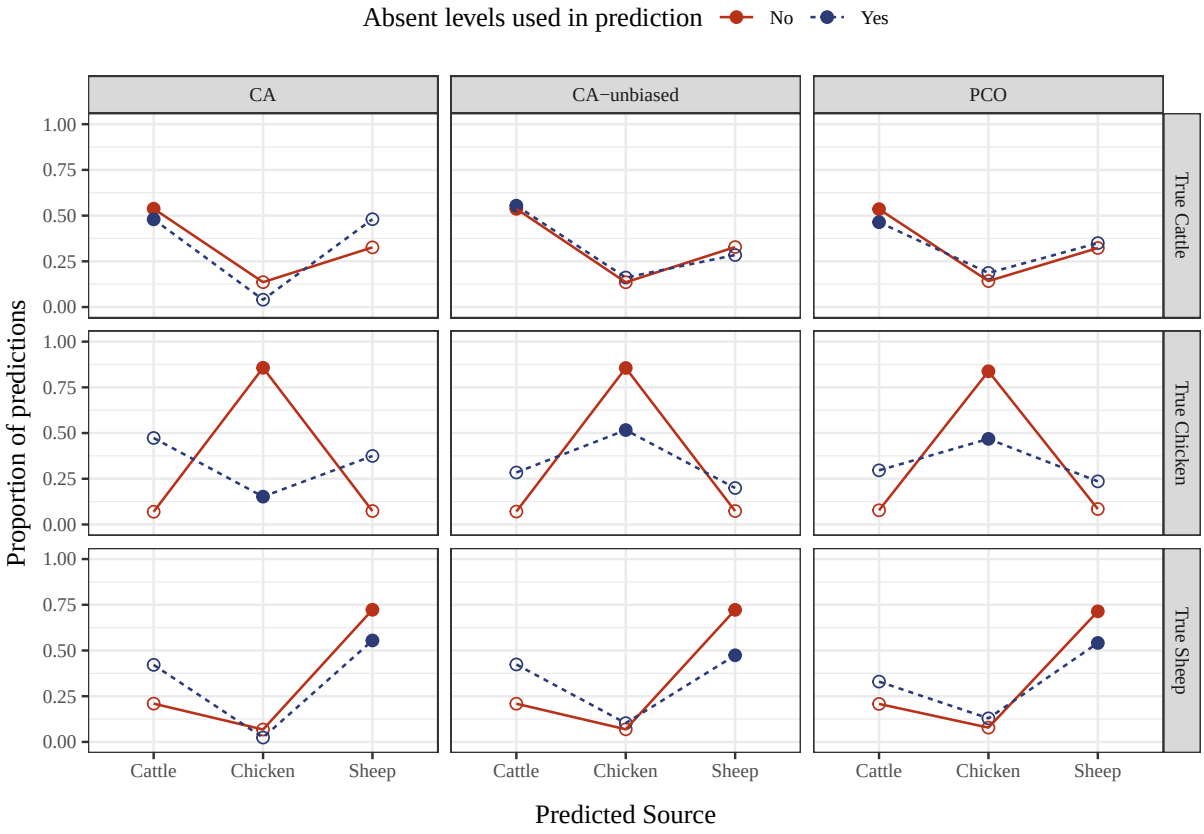

Figure 4.2: Proportion of tree predictions assigned to each of three host sources (cattle, chicken and sheep) when absent levels are used or not used in predictions. Open circles represent the proportion of cases for which the true class is predicted incorrectly; closed circles represent the proportion of cases for which the true class is predicted correctly.

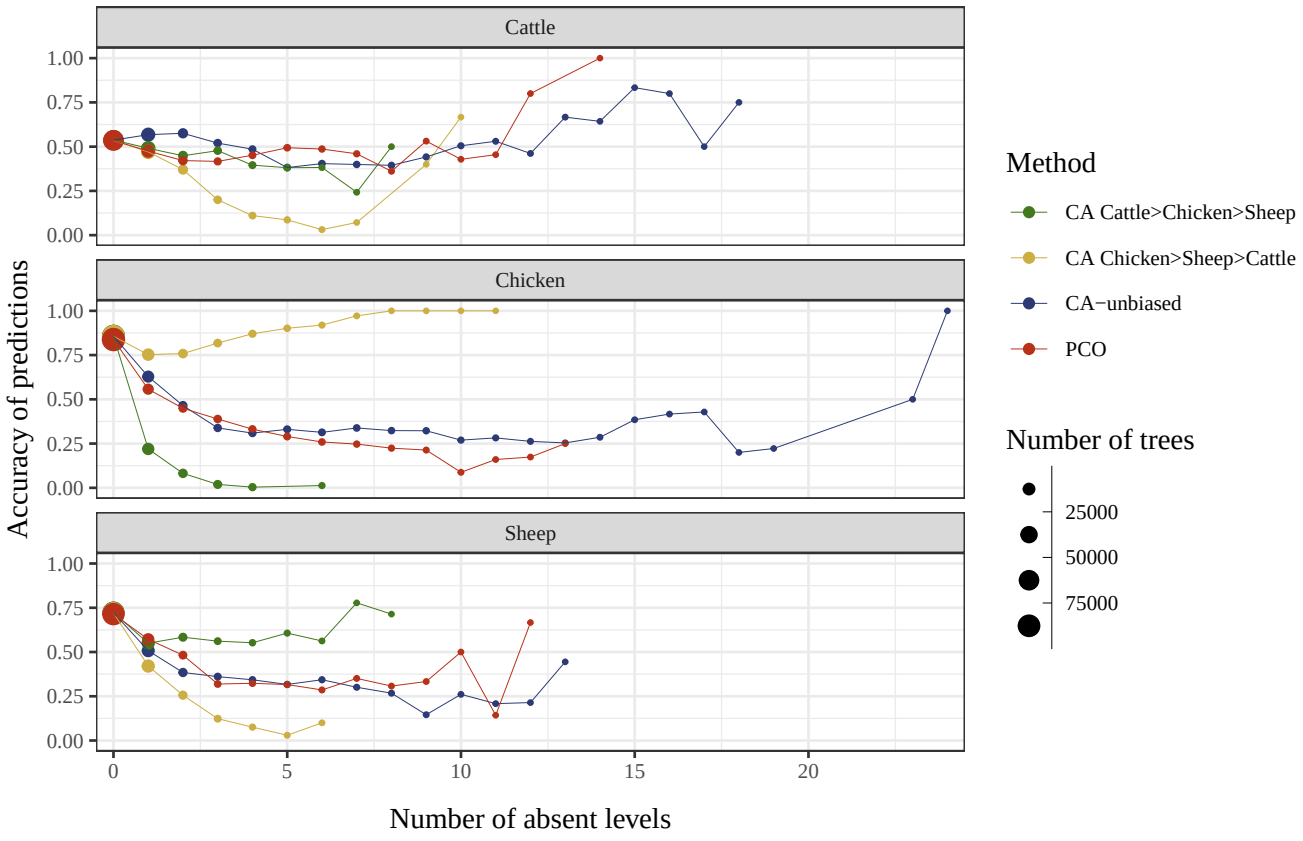

Figure 4.3: Proportion of predictions which were correct for trees with different numbers of absent levels and different methods and/or ordering of response class.

4.3.5 Effect of Response Class (Source) Order

The order of the response (source) levels also affected the success rates of predictions for the CA-encoding method when absent levels were used in prediction (figure 4.4). By default, most software treats the levels of categorical variables alphabetically, unless another ordering is specified explicitly. For the SACNZ data this equates to cattle < chicken < sheep. In the presence of absent levels, the CA-encoding method will encode any absent level with the highest rank and thus the observations will always be sent down the right branch of the tree. When the source levels were re-ordered as chicken < sheep < cattle, more observations with an absent level were assigned to chicken (the first response) than when the default ordering was used. This effect of class order did not occur with the CA-unbiased-encoding, or PCO-encoding methods (appendix A.1).

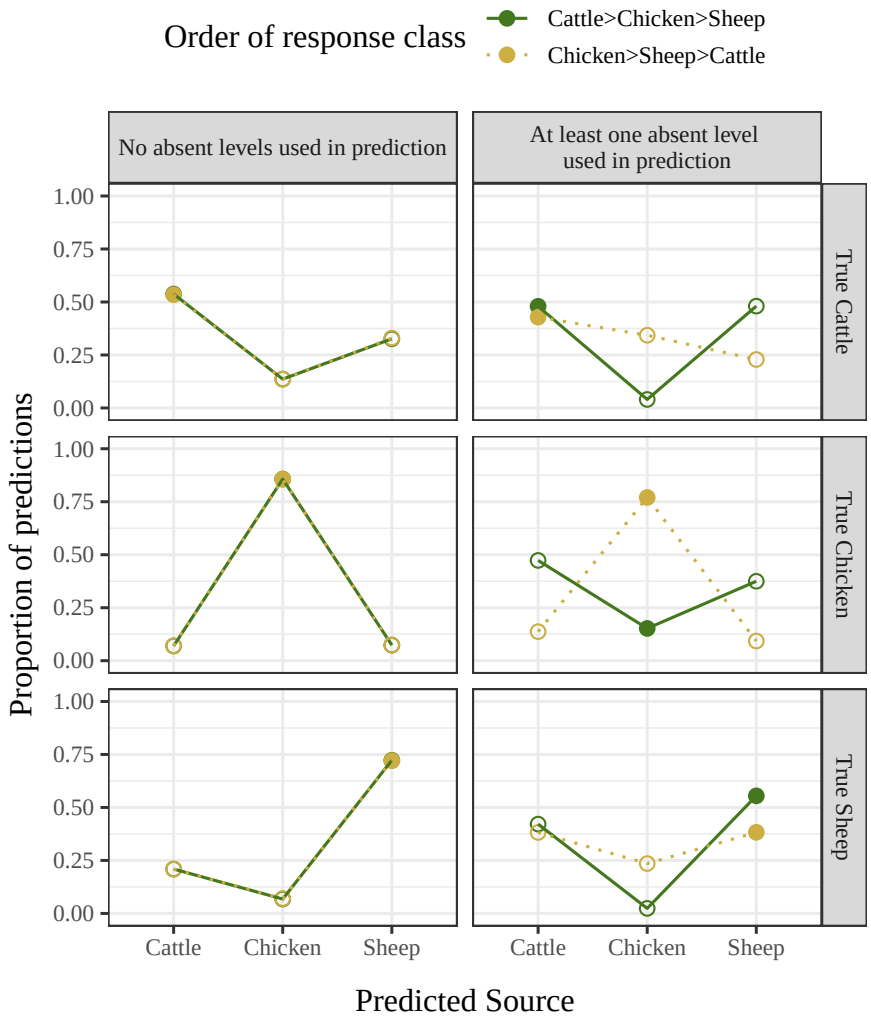

Figure 4.4: The effect of response class order on classification accuracy for the CA-encoding method. Open circles represent the proportion of cases for which the true class is predicted incorrectly; closed circles represent the proportion of cases for which the true class is predicted correctly.

4.4 Discussion

Data sets with large numbers of predictor variables and/or large numbers of categories create a significant challenge for modelling. Random forest is a compelling option for such cases, particularly suited to sets of high dimensional data of high cardinality. Random forest models trained with high cardinality variables, such as source attribution models utilising a core-genome MLST scheme, will almost certainly encounter absent levels when predicting for new data, and indicator encoding would lead to a prohibitively large number of binary variables - the cumulative number of unique alleles, across the genomes, from all the observations used to train the model.

Ordinal encoding can result in significant gains in efficiency of random forest models and

additionally bypasses any restrictions imposed on the number of levels.[8] Ordinal encoding also provides a means of classifying observations with absent levels as additional levels can be added sequentially. The ordinality induced by integer encoding is artificial, however, and may be detrimental to random forest predictions (Wright and König, 2019). It is particularly problematic if the alphabetical ordering of the levels (i.e., the labelling) has some degree of association with the class, which may occur with temporal labelling of predictor levels. For example, the open-access PubMLST database (Jolley et al., 2018) defines alleles numerically and in a sequential manner based on sequence deposition. In this instance, treating alleles as numeric would not be appropriate because allele "1" is not necessarily more related to allele "2" than it is to allele "500". However, it is likely that isolates have been added to the database in groups according to host source, so that their numeric order may partition into contiguous chunks by host. The numeric order thus provides information on likely host sources which is external to the data in a particular study, potentially biasing class assignment (chapter 7, appendix D.1).

This study found that, for random forest, different methods of encoding nominal variables had important implications for the accuracy of predictions when absent levels were encountered during prediction. When predicting using data with absent levels the CA-encoding method was biased towards the first response class. It also found that the systematic bias was affected by both the proportion of absent levels in the data as well as the level of association of the absent level with a response class (appendix A.2). For this method, the predictor levels are target encoded using their contribution to response class and an absent level is encoded with the highest rank. Changing the order of levels of the response classes can alter (reverse) the ranks of the predictor levels, however, the absent level will always retain the highest rank. Thus, the absent level will be next in rank to a level of a predictor associated with one response class in one ordering, but with the reverse ordering it will be next in rank to a different predictor level, potentially associated with a different response class. This option for encoding variable levels has previously been recommended when variables have a large number of levels and/or do not have an inherent order (Wright and König, 2019).

The first alternative method introduced in this chapter, the CA-unbiased-encoding method, is identical to the CA-encoding method except for the treatment of absent levels. The CA-unbiased-encoding method encodes all absent levels with a score of zero (rather than infinity) in line with the assumption of *a priori* equal class probabilities. This approach resolved the systematic bias towards the first response class caused by absent levels and showed a small improvement in overall classification accuracy (appendix A.2).

The second alternative method introduced in this chapter, the PCO-encoding method, used Gower's method of principal coordinates analysis on data that was independent of the class probabilities to inform the encoding of predictor variables, including absent levels (figure 4.1, c). This method assumes that an observation with an absent level is more likely to branch in

[8] When nominal encoding a categorical variable (e.g., the `"partition"` method in `ranger`), each binary node assignment is saved using the bit representation of a double integer, which limits this treatment to predictors with fewer than 54 levels (Wright and König, 2019).

the same direction as an observation whose corresponding level is 'similar' to the absent level. This requires information with which to quantify the similarity (or dissimilarity) of each pair of levels of a predictor variable. This study demonstrated the method using genomic sequencing data for each predictor variable, more specifically, the number of nucleotides shared by any two alleles (Hamming distance) for a given gene. In contrast to the CA-encoding and CA-unbiased-encoding methods, encoding using PCO was independent of the counts of levels of predictor variables in the training data, and thus also able to be applied to absent levels. In addition, rather than encoding all absent levels with the same score, the PCO-encoding method encoded each absent level individually. Using the Hamming distance between the absent allele and every other allele, the absent allele was encoded so that it was more similar to an allele with which it shared more nucleotides and less similar to an allele with which it shared few nucleotides. This is based on the assumption that isolates from one source would be more likely to have alleles which are similar in terms of their genome sequence, than isolates from another source (Pinheiro et al., 2005; Pérez-Reche et al., 2020). This method was not systematically biased, and had similar prediction accuracy to the CA-unbiased-encoding method.

The issue with absent levels will be less problematic for data where every level of every predictor variable in the set of observations to be classified is present in the training data, and more problematic for data containing variables with many levels. Previously, it was thought that no meaningful splitting decision can be made for observations with new levels at a splitting node and discussion has ensued regarding the advantages of keeping the observations with absent levels together *versus* assigning them randomly at a split (Wright and König, 2019). This chapter introduced two methods which do make meaningful splitting decisions for observations with new levels - the CA-unbiased-encoding method and the PCO-encoding method. Both of these methods produce competitive prediction results, resolve the systematic bias caused by absent levels, and avoid arbitrary splitting decisions for observations with absent levels. Although here only the first principal component/coordinate is used, it may be beneficial to increase the dimension to at least two principal components/coordinates. In addition, combining a target-based approach with ancillary information on the levels to inform variable ordering, particularly the placement of absent levels, may further improve classification success. These new methods would be suitable for any high cardinality predictors where a measure of level similarity could be determined. For example, a free text response field from a survey has a potentially infinite set of responses and absent levels would be almost inevitable. The string difference between responses could be used as a measure of similarity.

The success of a random forest classification model is often measured by the rate of misclassifications. Breiman (1996, 2001) claimed that the OOB misclassification rate (i.e., the rate of misclassification of cases that were not selected for training a particular tree) was as reliable as using an independent set of data for testing. When using a target-based encoding method (e.g., the CA-encoding or CA-unbiased-encoding methods), there is information leakage from the target variable to the predictors. The levels of each predictor variable are encoded according to the first principal component of the weighted matrix of class probabilities, calculated from

the entire (training) dataset before the analysis. Each observation in the set of training data is used to train approximately two thirds of the trees in the forest. The remaining third of trees can be used to generate an OOB prediction for that observation, which will be either correct or not. There is information leakage, however, because even when the observations are in the OOB set, the encoding of their corresponding levels was informed from the entire dataset (i.e., prior to the observations moving OOB) based on the correct response classes (i.e., the target); therefore, the OOB observations do not behave like fully independent test data. This leakage will impact OOB errors and they will likely underestimate the true misclassification rates. Potential solutions to this problem include re-ordering the levels at each split in the tree, re-ordering the levels of each bootstrap sample, or calculating the misclassification rate based on a fully independent test dataset. Target-agnostic encoding methods, such as the naïve alphabetical encoding and the PCO-encoding method, do not suffer the information-leakage problem because the response class (target) information is not used for the encoding. The PCO-encoding method will therefore not have this potential issue with incorrect OOB misclassification rates.

4.5 Conclusion

This chapter highlights potential pitfalls in the use of classification trees when an order is imposed on nominal predictor variables. These findings are applicable to random forest and other tree-based methods (e.g., boosted trees) when new levels of categorical predictor variables are encountered during prediction and/or where OOB misclassification rates are calculated. When levels of categorical predictor variables are target encoded using class probability information, and absent levels are integrated at the highest rank (effecting a consistent direction for them to branch at a split), predictions were systematically biased to the first response class. Target-based encoding of predictors using class probability information, and integrating absent levels according to the *a priori* hypothesis of equal class probability, is a potential and unbiased solution with good predictive properties. Target-agnostic encoding of predictors using information which quantifies the similarity between each pair of predictor levels, and integrating absent levels by virtue of their similarity to each of the other levels in the training data, is another potential solution which removes the need for arbitrary decisions on where to direct absent levels. This approach has good predictive properties, is not biased, and does not affect the OOB misclassification rate. The predictive performance of the PCO-encoding method depends on the ability to separate the levels according to class in the principal coordinate space and will depend on the ancillary information available. For high cardinality data, such as WGS data, it is almost certain there will be absent levels across the predictor variables, and that a large number of observations will be affected. Removing observations and/or variables with absent levels is, therefore, not a viable option. When there are no, or few, absent levels the different methods have similar predictive performance. However, as there can never be assurance of an absence of absent levels, there are no circumstances where the CA-encoding method should be used. As a result of this study, it can be recommended that, when ancillary information is available, such as with WGS data, the PCO-encoding method be used for random forest models and model performance can

be compared with the CA-unbiased-encoding method using misclassification rates calculated with an independent dataset. A reduction in bias for source attribution modelling will lead to a better understanding of potential risk factors in zoonotic infectious diseases to better inform public health decision making.

Supplementary Information

This chapter has two accompanying supplementary files -

Appendix A.1 - The effect of response class order on classification accuracy. An extension to figure figure four showing the effect of response class order on classification accuracy for the CA-encoding, the CA-unbiased-encoding, and the PCO-encoding methods.

Appendix A.2 - Bias resulting from treatment of absent levels. An illustrative simulation study showing the effect of increasing proportion of absent levels on classification accuracy for the CA-encoding and CA-unbiased-encoding methods.

Chapter 5

To CAP it Off - Further Methods of Encoding of Categorical Variables

5.1 Introduction

Encoding of categorical variables is a necessary preprocessing step for many machine learning algorithms. Ordinal encoding is an ideal method when variables have levels with a clear hierarchy or sequence (i.e., ordinal variables) as the transformed numeric values will preserve the inherent ordering of the levels. When the levels do not have a natural order (i.e., nominal variables), nominal encoding may be used. A nominal variable with a large number of levels (high-k), however, may be an issue for tree-based machine learning methods when transformed in this way, due to computational constraints.[1] One solution is to impose an order on the levels. Ordinal encoding both bypasses any limitations on number of category levels and reduces the number of potential partitions at each binary split from $2^{k-1} - 1$ to $k - 1$.

Two methods of ordering nominal predictors, the CA-unbiased-encoding method and the PCO-encoding method, have been shown to have good predictive performance and are unbiased in the presence of absent levels (i.e., levels of a predictor variable that are present in data for prediction that were not present when the random forest was trained) (chapter 4; Smith et al., 2024b). The CA-unbiased-encoding method is a target-based method (i.e., scoring uses information from the response variable) which performs a scaled correspondence analysis on the contingency table of counts of variable levels by class, following the approximation of Coppersmith, Hong, and Hosking (1999). Each predictor variable is encoded according to the first (or more) principal component(s) of the weighted matrix of class probabilities, and absent levels are encoded with a principal component score of zero in accordance with the *a priori* hypothesis of equal class probability (figure 5.1(a)).

The PCO-encoding method is a target-agnostic method (i.e., scoring is independent of the response variable) which performs a principal coordinates analysis (PCO) (Gower, 1966) on

[1] When nominal encoding a categorical variable, and when each binary node assignment is saved using the bit representation of a double integer, encoding is limited to predictors with fewer than 54 levels (Wright and König, 2019)

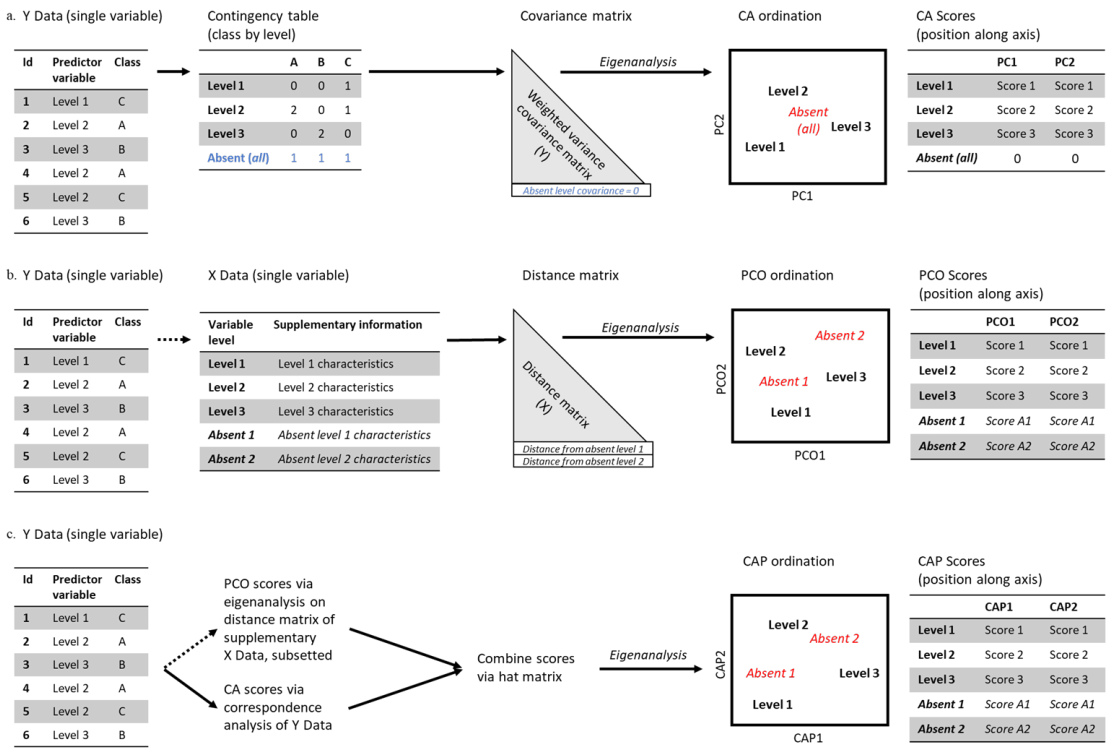

Figure 5.1: A visual description of the three methods described in this chapter (a) CA-unbiased-encoding method - the levels of each predictor variable are ordered according to the first two (or more) principal components of the class probabilities and absent levels are assigned a score of zero based on *a priori* equal class probabilities; blue text indicates conceptual information for an absent level; (b) PCO-encoding method - the levels of each predictor variable, including absent levels, are ordered according to their score for the first two (or more) principal coordinate axes derived from ancillary pairwise distance information; (c) CAP-encoding method - the levels of each predictor variable, including absent levels, are ordered according to their score for the first two (or more) canonical principal coordinate axes derived from a rotation of a subset of the PCO axes in (b) to maximally correlate with the PC axes in (a). Dotted line indicates the use of supplementary X data which is connected to, but not derived from, the Y data.

a matrix of distances, or dissimilarities, calculated between individual levels of predictor variables, where the distance is based on some characteristic of the variable, such as the string distance of level labels, or a supplementary dataset on the levels. Each predictor variable is encoded according to the first (or more) principal coordinate(s) and each absent level is scored, independently, based on its similarity to each of the other levels in the training data, using the method of Gower (1968) (figure 5.1(b)). Care must be taken with the choice of 'characteristic' by which the dissimilarity among category levels is defined (chapter 2). The encoding will be more effective when the dissimilarity between levels is meaningful (e.g., biologically) or likely to be associated with differences among classes. For example, differences between colour levels could be defined by alphabetical variation in the colour name, however, differences in RGB values would better capture any meaningful variation (figures 2.11, 2.12); and differences in countries may be defined, for example, geographically, culturally, or by Euclidean distances.

The choice of dissimilarity measure will impact the effectiveness of the encoding.

The goal of each of these methods is the same - to capture as much variation in the category levels as possible in the first eigenvector, with the new encoded value or score for each level being its position along the eigenvector. The position of each level along the eigenvector is the new encoded value or score. The difference in the methods lies in the multivariate space in which the variation among category levels is defined. The CA-unbiased-encoding method defines the variation in the class probability space, whereas the PCO-encoding method defines the variation in the chosen dissimilarity space of the category levels themselves. In the class probability space, there is no information available to differentiate new category levels from each other, and so the the CA-unbiased-encoding method encodes all absent levels as zero. This is unbiased, but is also uninformative. In contrast, in the multivariate dissimilarity space of the category levels, any new level may be placed according to its position relative to every other level, and so the PCO-encoding method will encode each absent level independently, and according to its position in the dissimilarity space. This allows an observation with an absent level to be treated as though it had a similar observed level, where the similarity is defined according to the chosen characteristic.

The ability of random forest to recursively partition any number of dimensions means that the difference in predictive performance of these encoding methods is most apparent in the treatment of absent levels. The CA-unbiased-encoding method is expected to perform well when levels are found in different relative frequencies across the classes, and when there are few absent levels. The PCO-encoding method is expected to perform well when the direction of greatest variation among category levels, according to the chosen characteristic, corresponds to the direction of greatest class separation. The issue then potentially lies in sets of data where there are many absent levels, and when the direction of greatest total variation is not the same as the direction of class separation in multivariate space. Under these conditions neither the CA-unbiased-encoding nor the PCO-encoding methods are expected to perform well for predicting observations, at least observations with absent levels. Rotating the set of points (i.e., the category levels) in the PCO space to emphasize class differences, so that the spread of points along the first eigenvector aligns with the direction of greatest class separation, is a potential solution to this problem. In this way, new levels may be incorporated in the PCO space, and therefore the rotated multivariate space, and be encoded using information on the levels while also aligning with class separation.

Canonical analysis of principal coordinates (CAP) (Anderson and Willis, 2003) is a method of constrained ordination, which rotates a cloud of points in multivariate space, as defined by a dissimilarity matrix, by reference to a specific *a priori* hypothesis. CAP performs a classical canonical analysis (e.g., canonical discriminant analysis, or canonical correlation analysis) on a set of new axes following an unconstrained ordination method, such as principal coordinate analysis (PCO; (Gower, 1966)) or nonmetric multidimensional scaling (nMDS; (Kruskal, 1964; Shepard, 1962)). Full details of the CAP method, including a mathematical description, are provided by Anderson and Willis (2003) and Anderson and Robinson (2003).

Measures of variable importance may be used to rank predictor variables according to their degree of influence on the predicted outcomes. Variable importance is often calculated from out-of-bag (OOB) samples. There are two broad measures of variable importance for random forest models - the Mean Decrease in Accuracy (MDA, or permutation importance) (Breiman, 2001); and the Mean Decrease in Impurity (MDI, or Gini importance) (Breiman, 2002). For both measures, a high value means that the variable has a positive impact on predictions. The CA-unbiased-encoding method and the CAP-encoding method are target-based methods of encoding and, therefore, a separate test set (i.e., rather than the OOB data) should be used to calculate both the out-of-bag error rate and measures of variable importance (chapter 6; Smith et al., 2024c).

This chapter introduces the CAP-encoding method - a novel method for encoding categorical predictor variables, including absent levels, which combines the key advantages of the CA-unbiased-encoding method and the PCO-encoding method, each described in chapter 4. First, the concept of canonical analysis of principal coordinates for encoding categorical variables is illustrated with a simulation study. Then the performance of this new approach is evaluated by assessing the misclassification rate of multiclass random forest predictions on three real-world datasets, each with high-cardinality categorical variables:

(i) survey data containing individual responses on regional self-identification in the Midwest, USA (Hickey, 2014);

(ii) data on traffic violation infringements issued by the Montgomery county police department, Maryland, USA (Montgomery, 2024);

(iii) a case study on source attribution of a foodborne pathogen, *Campylobacter jejuni* using whole genome sequencing (WGS) data as predictors (Lake et al., 2021).

5.2 Methods

The proposed CAP-encoding method encodes levels of categorical predictors by first performing a principal coordinates analysis on a symmetric matrix of distances between each pair of observed category levels, as per the PCO-encoding method. A subset of orthonormal principal coordinates (**Q**) are then rotated to maximally correlate with the direction of greatest variation in class probabilities, as determined by a correspondence analysis (CA) on a contingency table of category levels by outcome class. A hat matrix (**H**) is derived from the CA coordinates after centering on their column mean, and a classical canonical correlation analysis (CCorA) is performed on $\mathbf{Q}'\mathbf{H}\mathbf{Q}$ which generates a set of canonical eigenvalues **U** with associated eigenvectors. The set of CAP scores for the category levels is the product of **U** and **Q** (figure 5.1(c), appendix B.1).

5.2.1 Simulation Study

To illustrate the concept of CAP for encoding categorical variables, a set of data was simulated and analysed with random forest.

Table 5.1: Key attributes of three real-world datasets.

Dataset	Sample size	Target variable	Number of classes	Number of predictors Nominal	Other	Cardinality
Midwest survey	2421	Census region	9	1	24	767
Traffic violations	10000	Violation type	2	4	9	3-930
SACNZ	429	Source	3	1343	0	2-148

The simulated data consisted of a balanced frequency table, representing a two-class target variable and a 15-level categorical predictor variable. The position of each of the 15 predictor levels in a 2-dimensional PCO space was simulated such that the direction of greatest variation among the levels was along the first dimension, PCO1. Each level was then assigned to ten individuals, representing two classes, according to the probability that the level will belong to the first class, calculated from the inverse logit of each level's position along the second dimension, PCO2, scaled by β. β represents the magnitude of discrimination between the classes along PCO2. Each row sum of the class frequency table was ten, and the total number of individuals was 150. The dissimilarity matrix, **D**, was calculated from the Euclidean distance between each level in the simulated PCO space. In addition, ten of the category levels were recorded as 'observed' and five were recorded as 'unobserved' (i.e., absent levels).

The observations were split into two sets. A subset containing 75% of the observations from the 'observed' group of levels was used for training the random forest, and the remaining 25% of observations from the 'observed' group of levels and all the observations from the 'unobserved' group were used as the set of testing data.

The category levels were encoded according to each of three methods, the CA-unbiased-encoding method (chapter 4; Smith et al., 2024b), the PCO-encoding method (chapter 4; Smith et al., 2024b), and the new CAP-encoding method. Scores from the first dimension (i.e., the first eigenvector) were selected for each method to use in the random forest analysis.

A random forest was trained with ten trees and the Gini index splitting criterion. The misclassification rate of the observations from the 'observed' group and the 'unobserved' group in the testing data was calculated. The process was repeated 100 times, resulting in 1000 trees, for each of $\beta = 2$ and $\beta = 20$.

5.2.2 Analysis of Real-World Datasets

5.2.2.1 Midwest Survey

The Midwest survey data comes from a survey on regional identification of Midwesterners conducted for FiveThirtyEight in 2014.[2] The dataset contains responses to 25 questions from 2421[3] individuals representing nine census regions (classes) (table 5.1). There are 20 binary predictors - one for each Midwest state (IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD, WI, AR, CO,

[2]Midwest survey data available at https://github.com/fivethirtyeight/data/tree/master/region-survey

[3]357 rows with missing values for the target variable were removed prior to analysis.

KY, OK, PA, WV, MT, WY); four ordinal predictors - age group, education level, household income, and personal identification; and a single nominal predictor - 'In your own words, what would you call the part of the country you live in now?' (767 levels), referred to as 'Open response'. Observations with missing entries for the nominal predictor were removed prior to training the random forest. The distance measure used to differentiate levels of the nominal predictor was the Levenshtein distance (Levenshtein, 1966), which is the minimum number of single-character edits required to turn one response into the other.

5.2.2.2 Traffic Violations

This data comes from the set of all electronic traffic violations issued in the Montgomery county of Maryland.[4] The dataset contains a random subsample of 10,000 observations from two violation types ('Citation' and 'Warning'), following removal of incomplete observations (table 5.1). There is one numeric predictor (year); eight binary predictors - belts, property damage, fatal, commercial license, hazmat, commercial vehicle, alcohol, and work zone; and four nominal predictors - text description of the specific charge (930 levels), arrest type (18 levels), race (6 levels), and gender (3 levels). Full details of the data are published at https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q/about_data. A dissimilarity matrix of Levenshtein distances was calculated for each of the four nominal predictors.

5.2.2.3 SACNZ

The Source Assigned Campylobacteriosis in New Zealand (SACNZ) study is a source-assigned case-control study of notified human cases of campylobacteriosis in the Auckland and Mid-Central District Health Board regions, New Zealand, between 2018-2019 (Lake et al., 2021). The dataset contains 429 isolates of *Campylobacter jejuni* species from three sources (cattle (n=139), chicken (n=172), and sheep (n=118))[5] (table 5.1). There are 1343 nominal predictors (genes) which have cardinality of two to 187 levels (alleles). There were no loci without missing data, therefore observations with missing values were retained and missing values were treated as unique levels. Full details of the dataset are provided elsewhere (chapter 4; Smith et al., 2024b). A dissimilarity matrix of Hamming distances of the nucleotide sequencing information for each allele was calculated for each predictor.

5.2.2.4 Cross Validation

Each dataset was subject to ten-fold cross-validation for each of three methods (CA-unbiased-encoding, PCO-encoding, and CAP-encoding) using the same random number seed. Across the methods, the forest consisted of 500 trees and the Gini index was used as the splitting criterion. For each method, ten independent random forest models were run (one on each of the ten folds)

[4]Traffic violation data available at https://data.montgomerycountymd.gov/api/views/4mse-ku6q/rows.csv?accessType=DOWNLOAD

[5]The original data contained 1211 isolates from *Campylobacter jejuni* and *Campylobacter coli* species from four sources (humans, cattle, chicken, and sheep)

allowing each of the individual observations to be represented exactly once in testing data. Model performance was assessed by calculating the proportion of incorrect classifications on the set of test data for each fold and calculating the average and standard error, accounting for any variation between folds.

Variable importance was calculated using the independent holdout method (Smith et al., 2024c). This computes the mean decrease in accuracy (MDA, or permutation importance) by permuting values of the variable in a second cross-validation (test) fold which has been separated prior to encoding; and computing the difference in the error rate on the permuted test fold from the original test fold. The predictor variables were ranked according to the variable importance scores for their highest scoring dimension.

5.2.3 Code Availability

All analyses were carried out using R version 4.3.1 (R Core Team, 2024) and the `ranger` package ("RANdom forest GEneRator") version 0.15.1 (Wright and Ziegler, 2017). The R code used in this study is available at https://github.com/smithhelen/CAP. The sequence reads used for this study can be accessed from the NCBI Sequence Read Archive under BioProject accession number PRJNA675916.

5.3 Results

5.3.1 Simulation Study

In the simple case of two classes and balanced design, CAP successfully rotated the data in the PCO space so that the spread of points (predictor levels) along the first dimension was not according to greatest variation, as it is in the PCO space, but was according to greatest class separation (figure 5.2). In the PCO space, two dimensions are required to adequately separate the levels into class probability groups; whereas this is achieved with a single dimension in the CAP space.

The strength of the discrimination between the two classes for each level was set via the parameter β. For the trees which did not use absent levels for prediction, the average misclassification rates of the three encoding methods was the same and was lower for β=20 (3.7%) than for β=2 (30.8%). For the trees which used at least one absent level for prediction, the average misclassification rates of the CA-unbiased-encoding and the PCO-encoding methods both increased in line with random assignment to each class for both values of β (range 49.6% to 51.2%). The average misclassification rate of the CAP-encoding method for the trees which used at least one absent level for prediction, however, remained similar to the misclassification rates for the trees which did not use absent levels for prediction (33.5% when β=2, and 7.2% when β=20) (figure 5.3, appendix B.2).

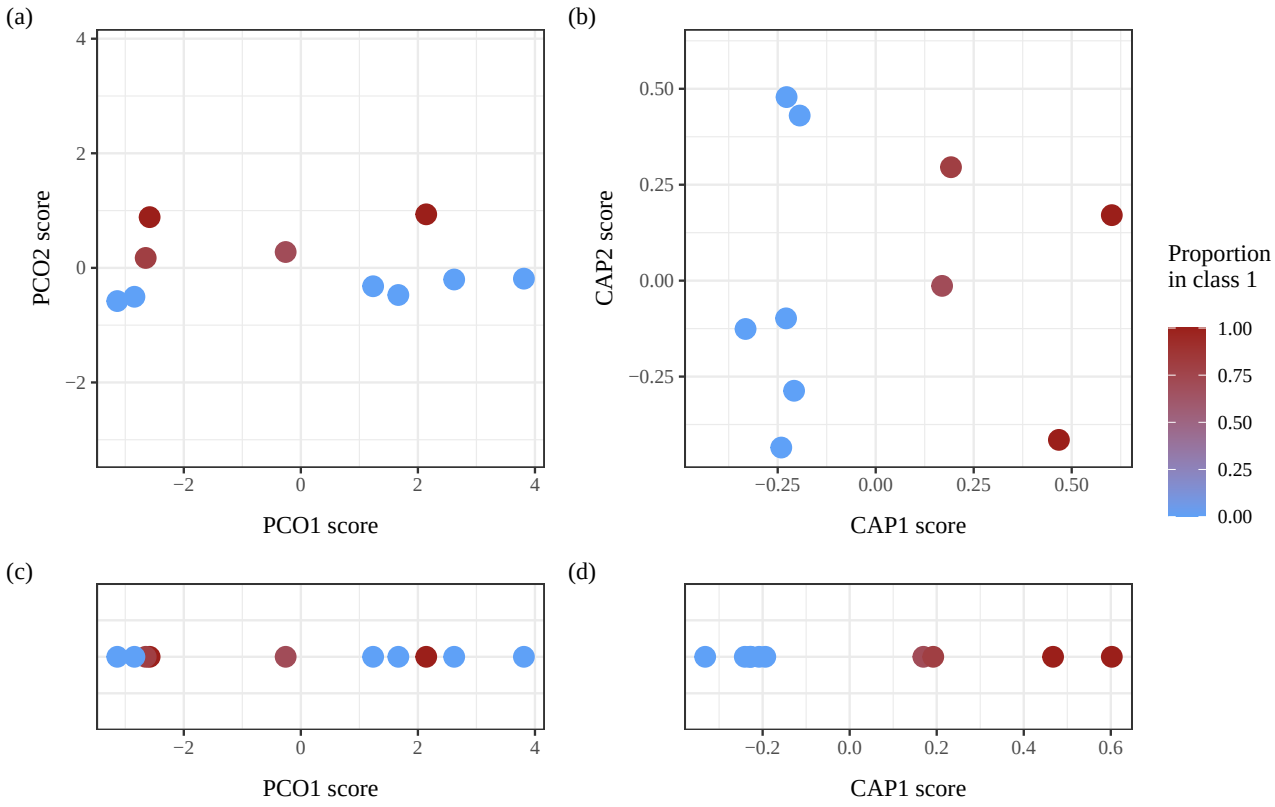

Figure 5.2: Placement of each predictor level in PCO space in (a) two dimensions and (c) one dimension and in the rotated CAP space in (b) two dimensions and (d) one dimension. Each point represents a predictor level. The colour represents the proportion of each level falling into class 1 *versus* class 2 when β=20. β represents the magnitude of discrimination between the classes along PCO2.

5.3.2 Real-world Datasets

For each dataset, the CAP-encoding method had the lowest average misclassification error, but the difference in accuracy between the methods was not consistent among datasets (figure 5.4, table 5.2). The improvement in prediction outcome following encoding is expected to be reflected in the variable importance scores for the encoded variable(s), and for each dataset the order of variable importance was determined by the method of encoding.

Midwest Survey Data

The misclassification rate of predictions for the Midwest survey data was lowest for the CAP-encoding method (38.2% ± 0.4%), followed by the CA-unbiased-encoding (44.2% ± 0.4%) and PCO-encoding (45.8% ± 0.4%) methods (figure 5.4(a)). The Midwest survey data has only a single categorical variable which was encoded ('open response'), therefore any difference in predictive accuracy between the methods can be attributed to this variable. The most important variable for the CA-unbiased-encoding and the CAP-encoding methods is the encoded nominal variable 'open response', however, the most important variable for the PCO-encoding method

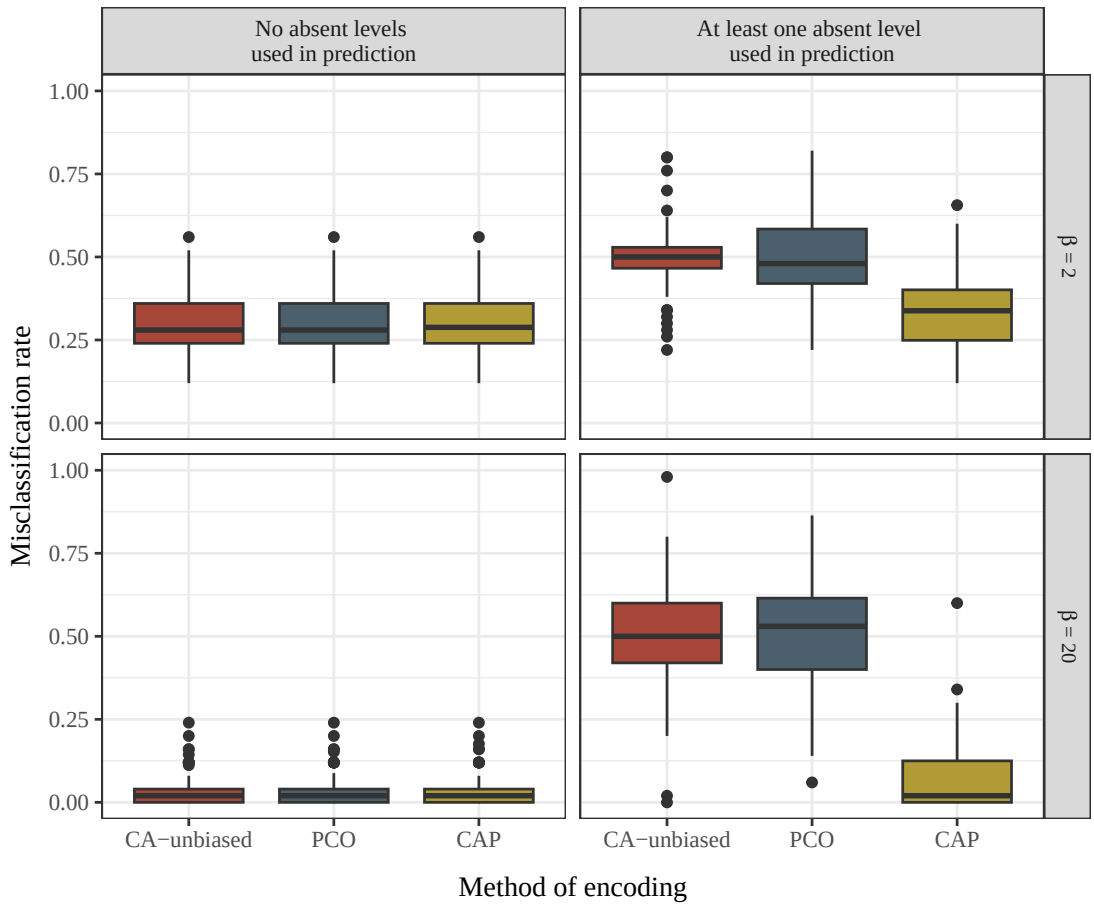

Figure 5.3: Misclassification rates of 1000 classification trees, from data simulated for ten individuals each with a single variable comprising 15 levels and assigned to two classes with probability proportional to β. Ten levels are observed in the training data and five are unique to the testing data. β represents the magnitude of discrimination between the classes along PCO2.

is the ordinal variable 'personal identification'.

For the 'open response' variable, the CA-unbiased-encoding method achieved very good separation of levels according to class membership for the known levels (figure 5.5(a)), however this method was limited by the large number of absent levels which were all scored as zero (figure 5.5(b)). The PCO-encoding method was less able to capture differences in levels according to class membership using the Levenshtein distance (figure 5.5(c)), and although absent levels were scored independently (figure 5.5(d), appendix B.3(a)), the scoring did not reflect class membership and was therefore not useful for prediction, as reflected in the lower variable importance ranking. The absent levels spread widely along the first PCO dimension, extending farther than the known levels (appendix B.3(b)), illustrating the nature of open ended text response fields (the extreme levels were those which were unique, highly verbose responses). The CAP-encoding method achieved both good separation of known levels (figure 5.5(e)) and of absent levels (figure 5.5(f)).

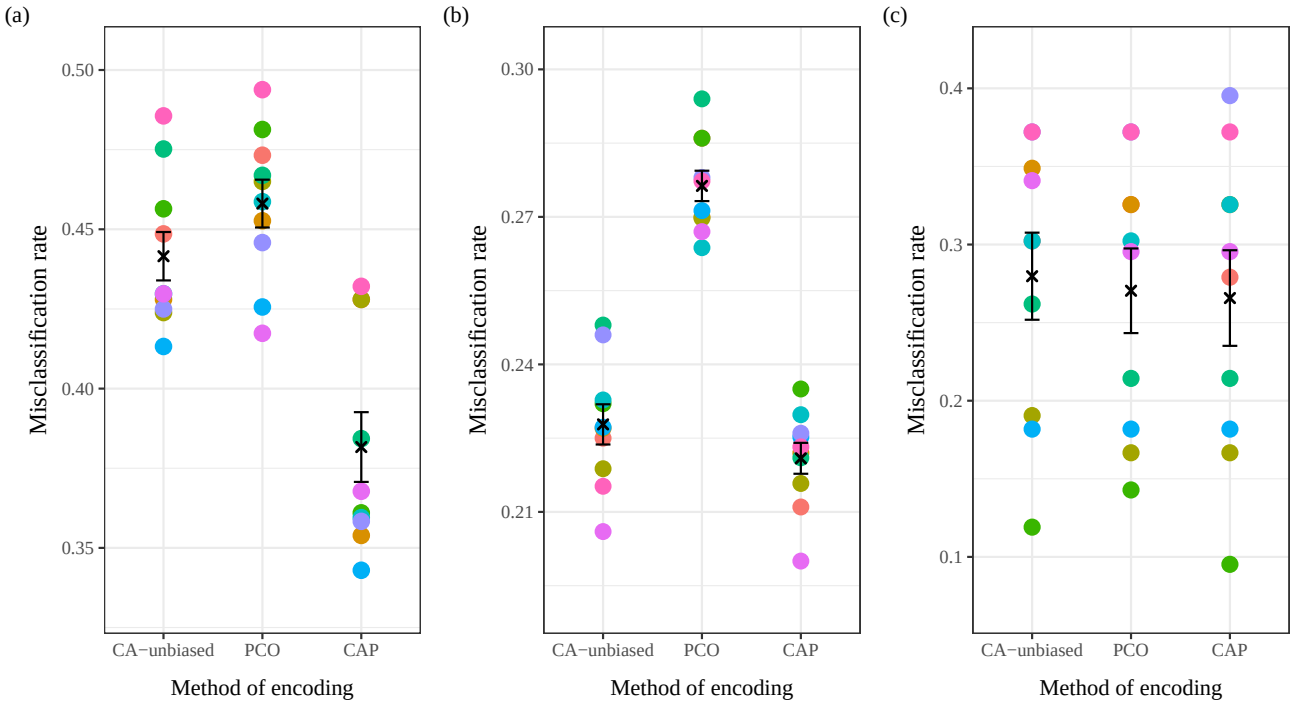

Figure 5.4: Misclassification rates for the (a) Midwest survey, (b) Traffic violation, and (c) SACNZ datasets for each method of encoding. Each coloured point represents the misclassification rate of a fold, and the weighted average and standard error are depicted with a black cross and error bars.

Traffic Violation Data

The misclassification rate of predictions for the traffic violation data was lowest for the CAP-encoding method (22.1% ± 0.1%), followed by the CA-unbiased-encoding (22.8% ± 0.1%) and PCO-encoding (27.6% ± 0.1%) methods (figure 5.4(b)). The traffic violations data has four nominal predictors which each underwent encoding (charge description, arrest type, race, and gender). There are two response classes, therefore there is only a single dimension for the CA-unbiased encoded variables, but two dimensions for the PCO-encoding and CAP-encoding methods. The most important variable for all three methods was 'charge description'.

For the 'charge description' variable, the CA-unbiased-encoding method achieved a clear gradient of scores according to class membership for the known levels (figure 5.6(a)). As with

Table 5.2: Weighted average and standard error of random forest misclassification rates for each of three methods of encoding categorical predictors.

Dataset	Method of encoding		
	CA-unbiased	PCO	CAP
Midwest survey	44.2% ± 0.4%	45.8% ± 0.4%	38.2% ± 0.4%
Traffic violations	22.8% ± 0.1%	27.6% ± 0.1%	22.1% ± 0.1%
SACNZ	27.9% ± 0.4%	27.0% ± 0.4%	26.5% ± 0.4%

the Midwest example, this method was limited by the large number of absent levels which were all scored as zero (figure 5.6(b)). The PCO-encoding method showed some clustering of levels from different classes but overall had difficulty capturing differences in levels according to class membership using the Levenshtein distance for both the known and the absent levels (figure 5.5(c, d)). The CAP-encoding method successfully rotated the PCO axes to achieve good separation of known levels (figure 5.5(e)). The separation of absent levels is not as clearly defined, but is still a significant improvement on the other methods (figure 5.5(f)).

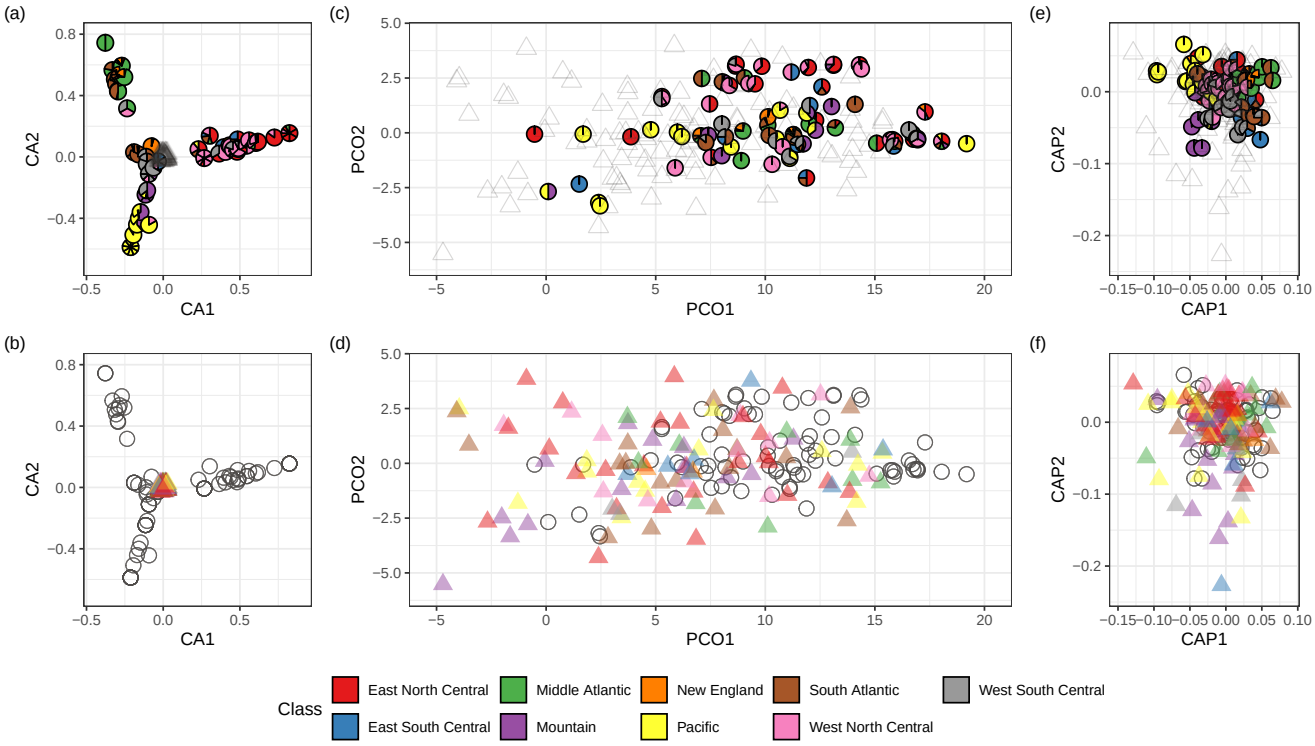

Figure 5.5: The first two dimensions of encoded scores (jittered) for the categorical variable 'open response' from the Midwest Survey dataset for the CA-unbiased-encoding (a, b), and CAP-encoding (e, f) methods, and a subset[6]of scores for the PCO-encoding method (c, d). Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels.

SACNZ Data

All 1343 variables in the SACNZ data are categorical (genes with alleles as levels) and were encoded. The misclassification rate of predictions for the SACNZ data were very similar for the three methods of encoding (range 26.5% (CAP-encoding) to 27.9% (CA-unbiased-encoding))

[6]The full set of scores are available in appendix B.3

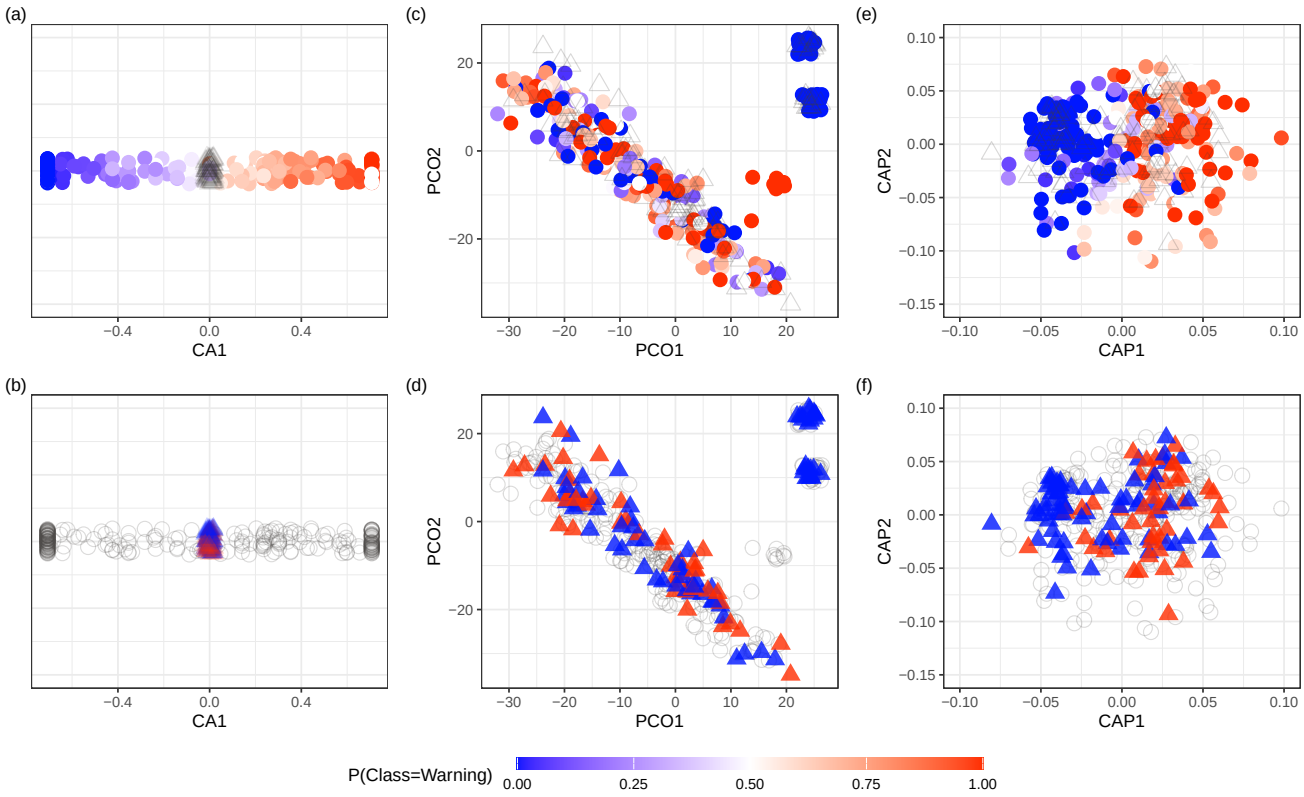

Figure 5.6: The first two dimensions of encoded scores (jittered) for the categorical variable 'charge description' from the Traffic Violation dataset for the CA-unbiased-encoding (a, b), PCO-encoding (c, d), and CAP-encoding (e, f) methods. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels.

(figure 5.4(c)). The most important variables were not shared between any of the encoding methods. The most important variable for the CA-unbiased-encoding method was 'CAMP0038'; for the PCO-encoding method was 'CAMP1162'; and for the CAP-encoding method was 'CAMP1179'.

All three methods were able to encapsulate some degree of class separation for each of these genes (figures 5.7(a,c,e), 5.8(a,c,e), 5.9(a,c,e)). These three genes each only had two absent levels, so the method of encoding was not strongly influential on the outcome (figures 5.7(b,d,f), 5.8(b,d,f), 5.9(b,d,f)).

5.4 Discussion

Random forest predictive models generally require, and benefit from, encoding of categorical variables (Liaw and Wiener, 2002; Pedregosa et al., 2011; Wright and Ziegler, 2017). The computational benefits of ordinal encoding categorical variables are well known and, more recently, methods of ordinal encoding which enable scoring of absent levels (i.e., levels of a predictor

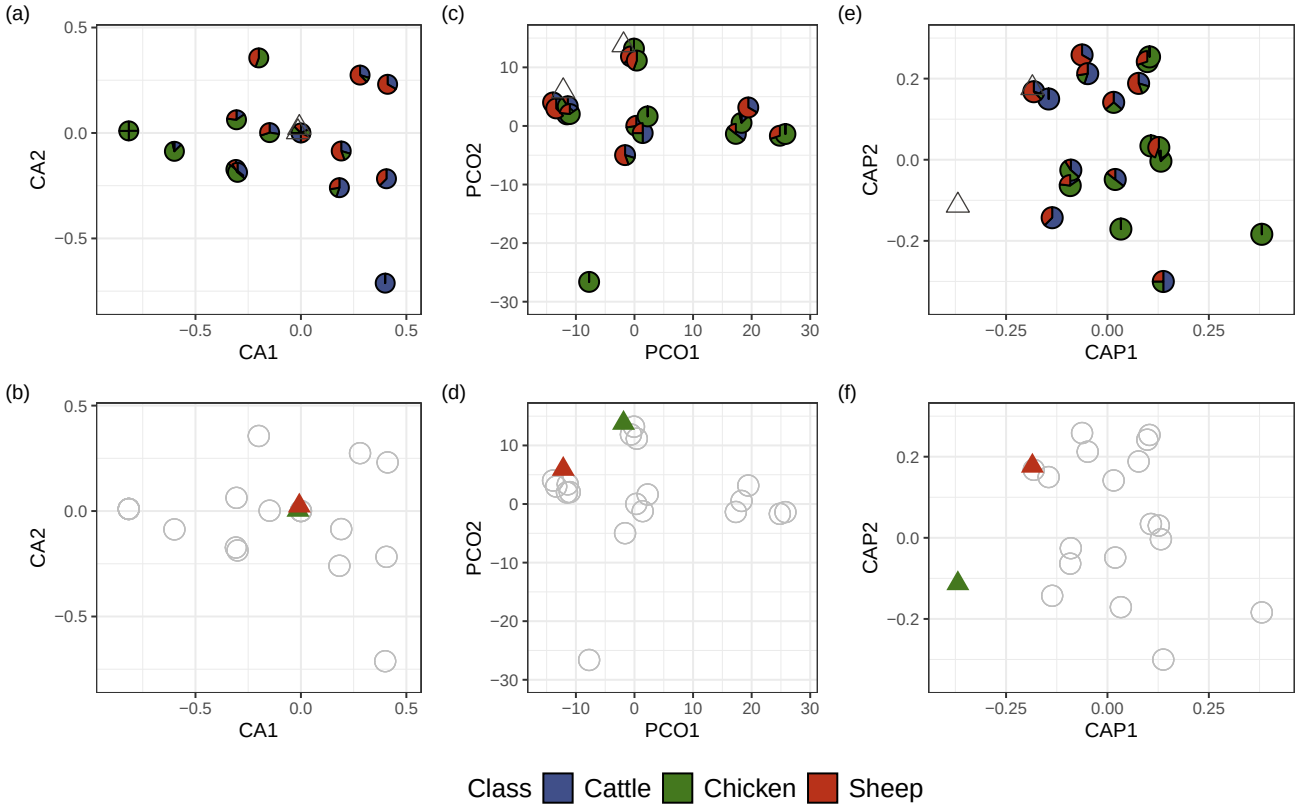

Figure 5.7: The first two dimensions of encoded scores (jittered) for the categorical variable 'CAMP0038' for the SACNZ source attribution dataset for the CA-unbiased-encoding (a, b), PCO-encoding (c, d), and CAP-encoding (e, f) methods. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels.

variable that were absent when a classification tree was grown but are present in new observations for prediction) have been described (chapter 4; Smith et al., 2024b).

Random forest has the ability to recursively partition each predictor and therefore the method of encoding is not highly influential on predictive outcome for observations which have a full complement of the predictor levels from the set of training observations, at least in unpruned trees. The distinguishing feature of different encoding methods is their mechanism of encoding absent levels. This chapter found that, for random forest, different methods of encoding nominal variables had important implications for the accuracy of predictions when absent levels were encountered during prediction.

The CA-unbiased-encoding method encodes all absent levels with a score of zero, in line with the assumption of *a priori* equal class probabilities. This is an unbiased method with good predictive properties when observations have a full complement of known levels; however, when

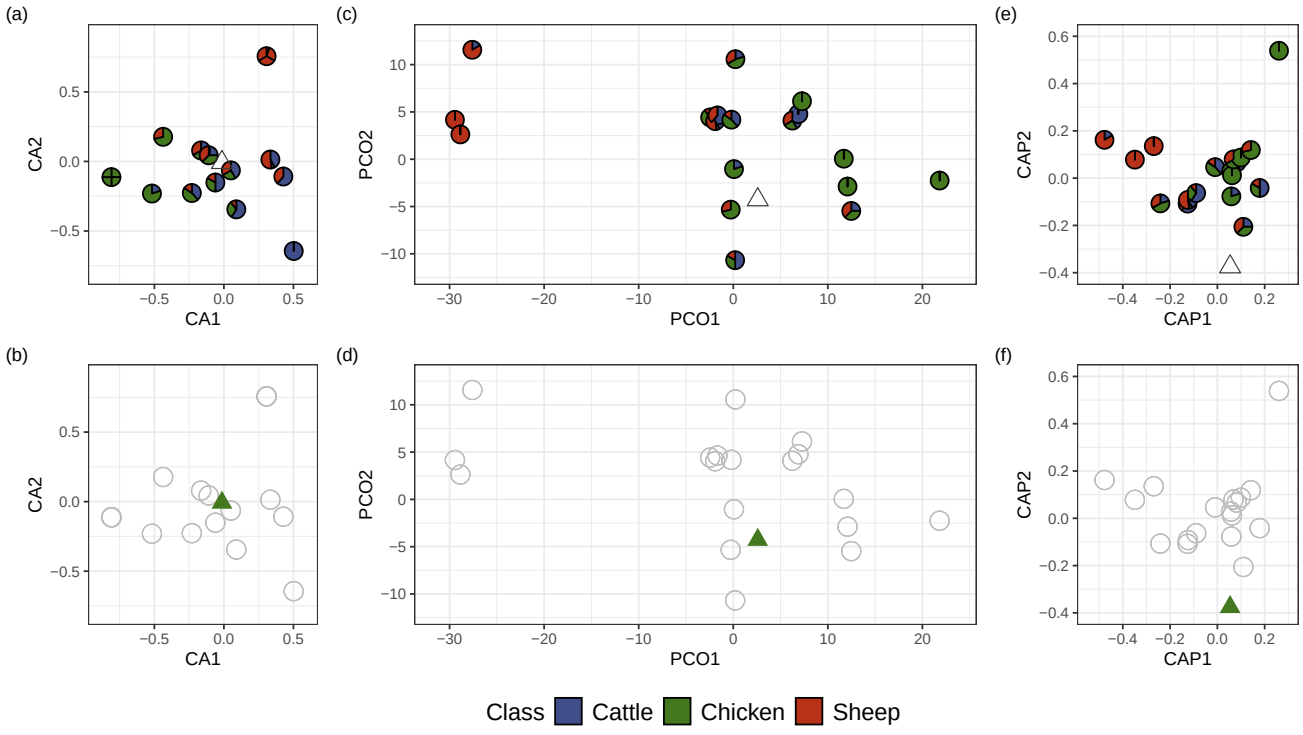

Figure 5.8: The first two dimensions of encoded scores (jittered) for the categorical variable 'CAMP1162' for the SACNZ source attribution dataset for the CA-unbiased-encoding (a, b), PCO-encoding (c, d), and CAP-encoding (e, f) methods. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels.

there are a high number of absent levels, this method is unable to separate observations based on these levels, reducing predictive performance. In addition, this method utilises the target variable to inform the encoding and, as such, is subject to information leakage to the OOB sample and therefore measures of accuracy and variable importance should be calculated using separate test data (chapter 6; Smith et al., 2024c).

The PCO-encoding method uses Gower's method of principal coordinates analysis to encode absent levels according to their similarity to each of the other levels in the training data based on some property of the variable. This method is unbiased and does not suffer information leakage because the encoding is truly independent of the target variable. This method, however, makes the assumption that an observation with an absent level is more likely to branch in the same direction as an observation whose corresponding level is 'similar' to the absent level, and requires an attribute with which to measure similarity (or dissimilarity). This method works particularly well when the measure of similarity is capturing a characteristic of the variable in which the variation between levels is correlated with the variation in class probabilities. When

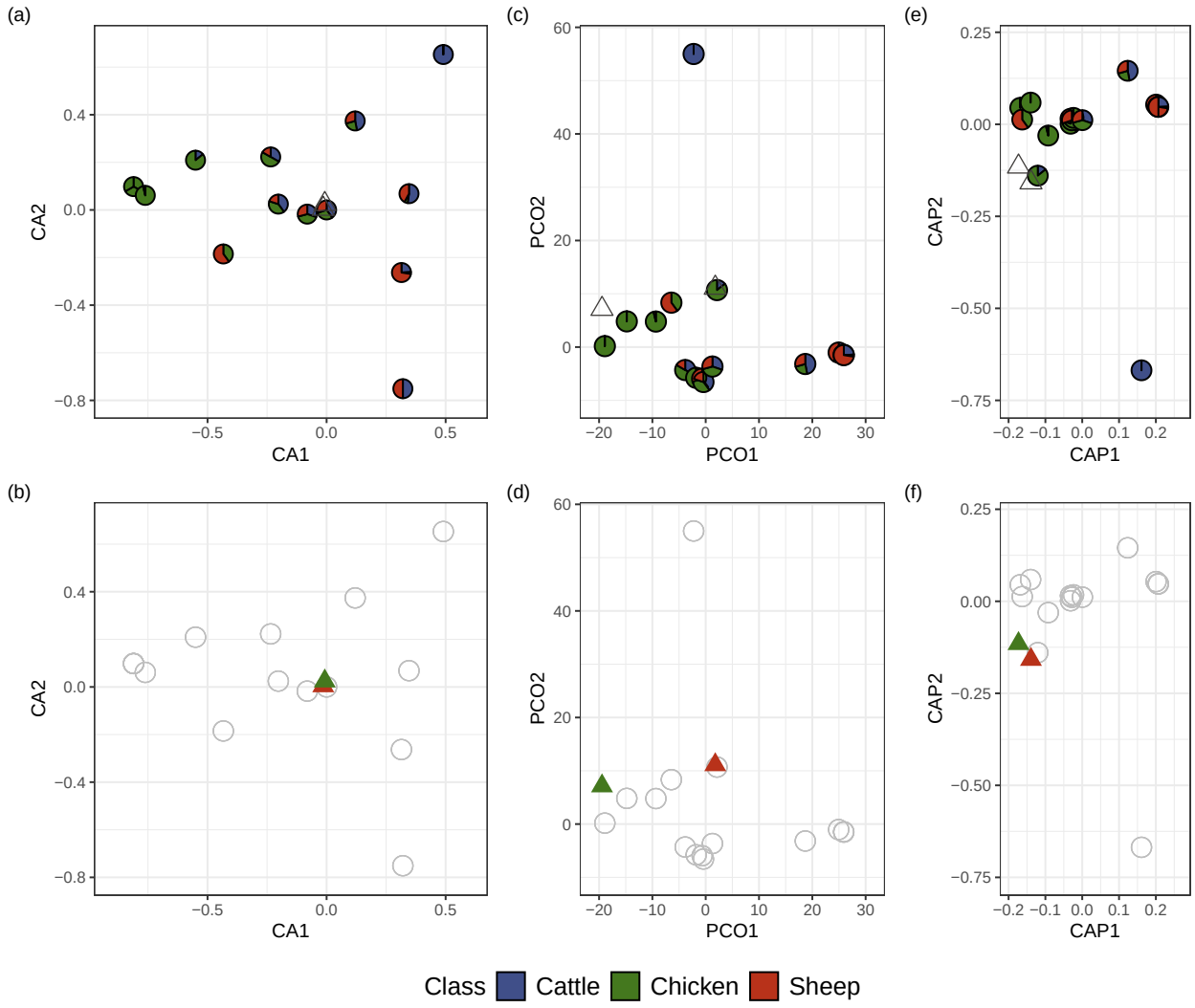

Figure 5.9: The first two dimensions of encoded scores (jittered) for the categorical variable 'CAMP1179' for the SACNZ source attribution dataset for the CA-unbiased-encoding (a, b), PCO-encoding (c, d), and CAP-encoding (e, f) methods. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels.

the direction of greatest variation among the levels does not overlap with the direction of greatest class separation then, although this method encodes absent levels independently of each other, the encoding will not be informative with regards to class membership.

This chapter developed a method that combines the strengths of the CA-unbiased-encoding method and the PCO-encoding methods using the method of canonical analysis of principal coordinates. The CAP-encoding method encodes absent levels by first performing a principal coordinates analysis, as per the PCO-encoding method, and then rotating a subset of the principal coordinates to correlate with the direction of greatest variation in class probabilities, as

per the CA-unbiased-encoding method. The CAP-encoding method performs at least as well as the CA-unbiased-encoding and the PCO-encoding methods, and is particularly advantageous when there is a high number of absent levels. This method utilises the target variable, therefore it is also subject to information leakage if care is not taken to use separate test data to evaluate predictive performance and variable importance.

The number of principal coordinates to include in the subset for rotation may affect prediction performance. The leave-one-out diagnostics described for the native application of CAP (Anderson and Robinson, 2003; Anderson and Willis, 2003) are less applicable here because observations with the same predictor category level may belong to multiple outcome classes. In addition, for datasets with multiple nominal variables, assessing them concurrently and in combination would be computationally intensive. The developers of CAP suggest retaining enough PCO dimensions to capture between 60% and 100% of the total variance in the dissimilarity matrix. This allows the majority of the information in the PCO space to be captured but ignores the random variation which is not beneficial for class discrimination. This study found, on the datasets that were examined, that predictive performance is optimal when the PCO subset is capturing 90-95% of the total variance in the dissimilarity matrix, across all variables (appendix B.4).

Of greater consequence is the number of dimensions to define as new numeric (encoded) variables for analysis by random forest. This study found that, for all methods, prediction performance improved when more than one dimension was used (appendix B.5) and that the optimal number of dimensions depended on both the method of encoding and the number of classes. For datasets with a higher number of classes, more dimensions were required before misclassification rates plateaued. The CA-unbiased-encoding method is limited to one fewer dimensions than the number of classes, and across the datasets improvement diminished after two dimensions. The PCO-encoding method particularly benefited from increasing number of dimensions, and misclassification rates continued to decrease up until one less than the number of classes. The CAP-encoding method performed universally better at all dimensions and achieved good prediction performance with two to three dimensions, depending on the number of classes (appendix B.5). A logical limit to the number of dimensions is one less than the number of classes, although, here, fewer were sufficient for datasets with many classes.

An additional dimension may be required for the PCO-encoding and CAP-encoding methods when there are missing values in the training data. A missing value is different to an absent level in that there is no information at all with which to represent the variable. Whether or not an observation has a value for a variable may, or may not, be informative for class assignment for that observation. When missing values are present in observations in the training data, they will be assigned a score as for any other predictor level. When missing values are only present in observations in the test data, they will be scored as absent levels. This causes no issues for the CA-unbiased-encoding method; however, for the PCO-encoding and CAP-encoding methods missing values may potentially dominate the first principal coordinate depending on how

they sit in the dissimilarity space. In this case, the first dimension will act as a binary measure of presence or absence of the variable (appendix B.6), and an additional dimension may be required.

There are several areas of this research that could be explored further in the future. The effect of class imbalance on predictive performance of encoding methods could be investigated to assess whether target-agnostic approaches offer any advantages over target-based approaches. In addition, given the importance of selecting a suitable characteristic and distance metric by which to define the categories for the distance based methods, this is an area for future consideration. Despite the vast array of available distance metrics, in some cases the patterns within classes may not be well captured and alternative metrics could be sought, for example accounting for genomic recombination when calculating the Hamming distance between alleles.

5.5 Conclusion

In this chapter, the CAP-encoding method was introduced, which combines the strengths of the target-based CA-unbiased-encoding and the target-agnostic PCO-encoding methods (chapter 4; Smith et al., 2024b), for encoding categorical predictor variables prior to analysis by random forest predictive models. Using three real-world datasets, it was empirically demonstrated that encoding nominal predictors using the CAP-encoding method performs uniformly better than the CA-unbiased-encoding and the PCO-encoding methods, which themselves have both been shown to be superior to the biased naïve approaches that are currently being employed (chapter 4; Smith et al., 2024b). For high-cardinality data, such as WGS data, and data with free text response fields, absent levels are likely to be present across the predictor variables and to be found in a large number of observations. Encoding these absent levels is a more satisfactory option than removing affected observations and/or variables. The distinguishing feature of different encoding methods, lies in their ability to separate absent levels according to class structure. The CAP-encoding method has good predictive properties and is not biased and should be the method of choice for encoding categorical predictors when ancillary information is available with which to define the similarity (or dissimilarity) of category levels. For this method, the size of the PCO-subset should initially be selected so that it is equal to the number of dimensions which capture 95% of the total variation in the dissimilarity matrix and initially selecting the number of CAP dimensions to be one less than the number of classes, and to adjust these parameters, if required, using misclassification rates calculated with an independent dataset. Although this chapter focused on encoding variables for random forest classification problems, the encoding can be applied much more broadly to any decision tree based method.

Supplementary Information

This chapter has six accompanying supplementary files -

Appendix B.1 - The CAP-encoding methodology.

Appendix B.2 - Simulation study comparing encoding methods when the direction of greatest variation in the category levels is along the first principal coordinate axis.

Appendix B.3 - Midwest survey 'open response' scores for the PCO-encoding method.

Appendix B.4 - The effect of different amounts of variation retained in the PCO subset on misclassification rates.

Appendix B.5 - The effect of increasing number of dimensions on random forest predictive performance.

Appendix B.6 - The effect of missing values on the first principal component of the variable 'CAMP1225' from the SACNZ dataset.

Chapter 6

Out of (the) Bag - Encoding Categorical Predictors Impacts Out-Of-Bag Samples

6.1 Introduction

Performance of random forest classification models is often assessed and interpreted using out-of-bag (OOB) samples. The method of bagging ('bootstrap aggregating') in random forest means that not every observation is included in every tree. For each tree, a bootstrapped sample which contains a specified proportion (typically two-thirds) of the observations in the training data is selected to train the model. The remaining one-third of observations, which are not included in the bootstrap sample are 'out-of-bag' and may serve as a test set for the tree (Breiman, 2001). The OOB sample may be used to estimate the predictive performance of the random forest and variable importance measures, amongst other things.

6.1.1 Out-of-Bag Error

An OOB prediction for an observation is obtained by aggregating the tree classifications for the observation from the OOB samples. The misclassification rate of the OOB predictions from all training observations is the OOB error (Breiman, 2001) (figure 6.1). OOB errors are popular because they are fast to compute, requiring only a single random forest to be computed, and have been reported to be a good estimate of the true prediction error (Adelabu et al., 2015; Lawrence et al., 2006; Mutanga and Adam, 2011). The OOB error may also be used to select appropriate values for tuning parameters, such as the number of predictor variables that are randomly drawn for a split.[1] Breiman (1996, 2001) claimed that the OOB error alleviates the need for cross-validation or setting aside a separate test set; however, it has been shown that, especially for small samples, the OOB error can over-estimate the true prediction error (Bylander, 2002; Janitza and Hornung, 2018; Mitchell, 2011). Methods to address the bias have been proposed (Bylander, 2002; Janitza and Hornung, 2018; Mitchell, 2011), although, when available, a large

[1] referred to as `mtry` in R packages `ranger`, `randomForest`, `randomForestSRC`, and the tidymodels framework; or `max_features` in Python's sklearn `RandomForestClassifier`.

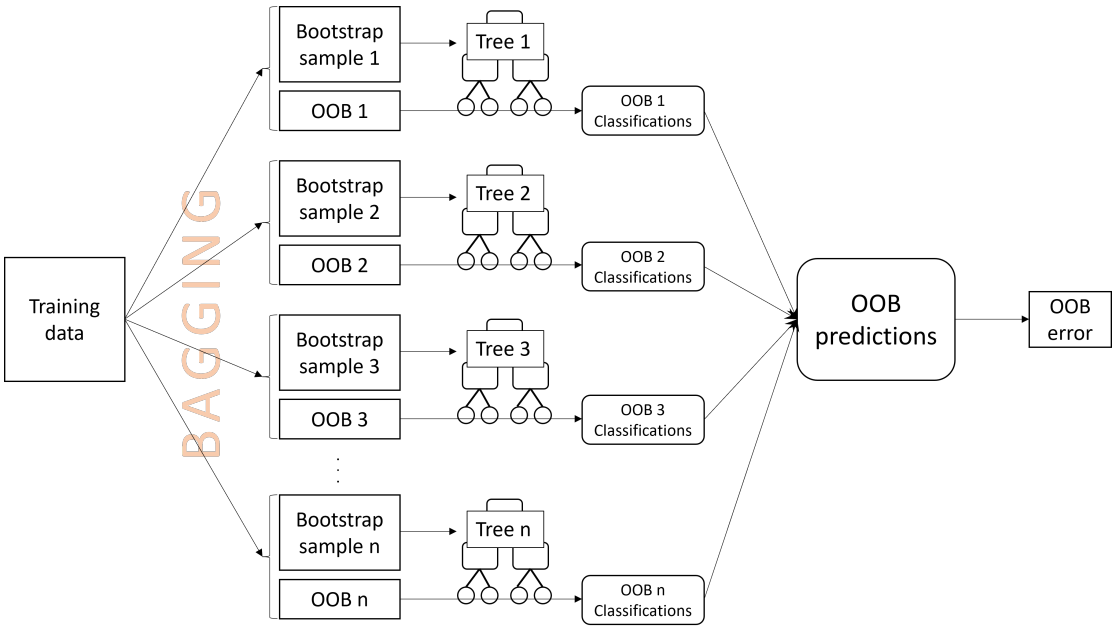

Figure 6.1: A visual description of the process of obtaining an out-of-bag (OOB) error estimate.

external validation data set will provide a more precise error estimate, serving as a gold standard (Hastie et al., 2009; Janitza and Hornung, 2018).

6.1.2 Variable Importance

OOB samples may also be used to calculate measures of variable importance. Variable importance can be used to rank predictor variables according to their degree of influence on the predicted outcomes. There are two broad measures of variable importance for random forest models - the mean decrease in accuracy (MDA, or permutation importance) (Breiman, 2001); and the mean decrease in impurity (MDI, or Gini importance) (Breiman, 2002). For both measures, a high value means that the variable has a positive impact on predictions.

MDA for a given variable is the mean decrease in prediction accuracy of the individual trees across the forest when the variable is not used for prediction. MDA is obtained by permuting values of the variable in the OOB sample and computing the difference in the error rate on the permuted OOB sample from the original OOB sample (figure 6.2). The idea is that permuting an important variable would result in a large decrease in accuracy while permuting an unimportant variable would have a negligible effect.

MDI is the weighted mean of the individual trees' decrease of impurity produced by a given variable. An important variable is expected to generate a larger decrease of impurity (i.e., more pure splits) than an unimportant variable. The decrease of impurity is measured as the difference between a node's Gini impurity and the weighted sum of the Gini impurity of the two child nodes, evaluated on the in-bag samples.

Several studies have highlighted issues with these importance measures and have proposed

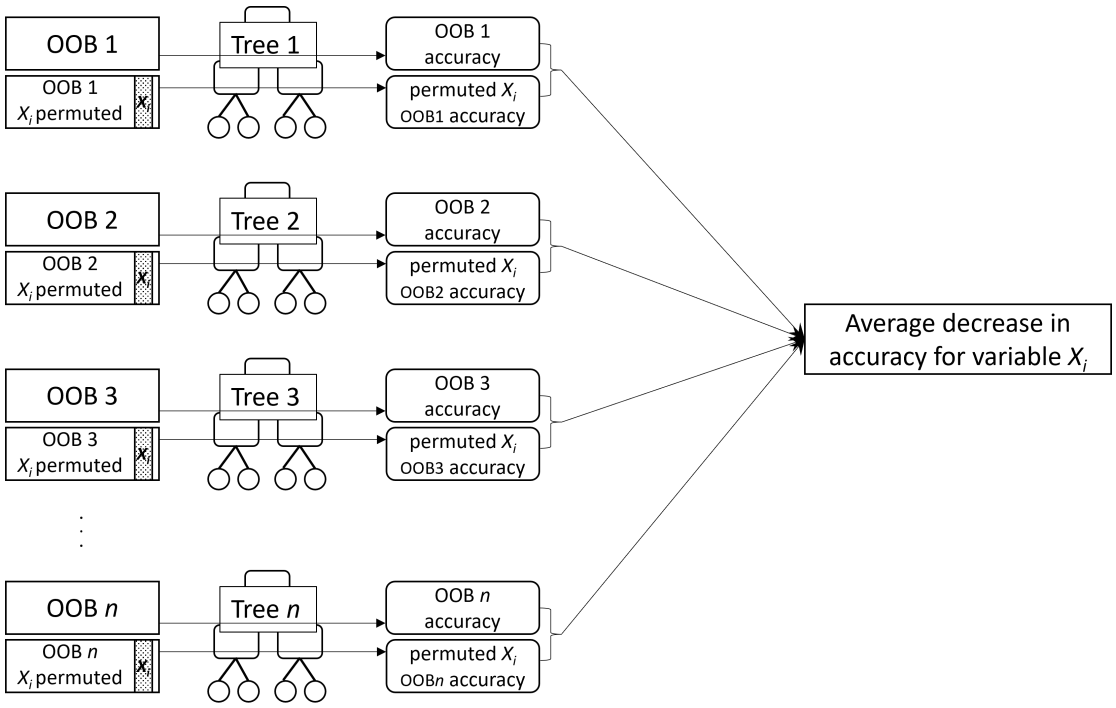

Figure 6.2: A visual description of the process of obtaining permutation importance (MDA) for variable X_i.

modifications which may overcome specific undesirable properties (Benard et al., 2022; Gregorutti et al., 2017; Janitza et al., 2018; Mentch and Zhou, 2022; Nembrini et al., 2018; Nicodemus, 2011; Nicodemus and Malley, 2009; Sandri and Zuccolotto, 2008; Strobl et al., 2007, 2008; Toloşi and Lengauer, 2011; Wallace et al., 2023; Williamson et al., 2023). Janitza et al. (2018) introduced the holdout variable importance method which computes MDA using a second cross-validation fold rather than the OOB data and this has been adopted as an option by the `ranger` and `randomForestSRC` packages. Also implemented by `ranger` is the actual impurity reduction (AIR) importance method (Nembrini et al., 2018; Sandri and Zuccolotto, 2008) which adjusts the original impurity by subtracting the impurity importance following random reordering of the variable (figure 6.3). There have been many other variable importance measures proposed (e.g., Dfuf et al. (2020); Epifanio (2017); Loecher (2022)), however they have not been widely adopted and MDA is generally considered the most efficient and accurate measure of variable importance (Boulesteix et al., 2012; Ishwaran, 2007; Nicodemus et al., 2010; Strobl et al., 2007; Szymczak et al., 2016; Ziegler and König, 2014).

6.1.3 Encoding Categorical Predictors

Categorical variables can, in theory, be used by random forest models in their raw state; however in practice, software will either require them to be numerically encoded (Pedregosa et al., 2011) or will encode them prior to processing (Liaw and Wiener, 2002; Wright and Ziegler,

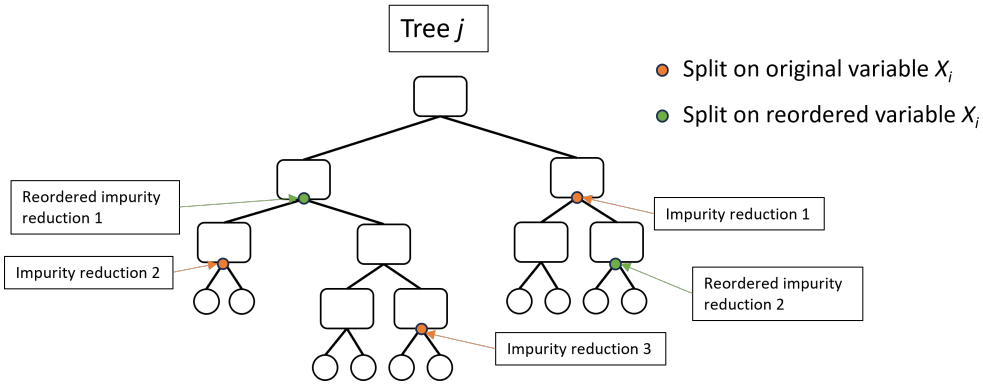

Figure 6.3: Illustration of the actual impurity reduction (AIR) calculation. The AIR for variable X_i for $\text{Tree}_j = \sum \text{Impurity reduction}_{X_i} - \sum \text{Impurity reduction}_{reordered X_i}$, where impurity reduction is the Gini impurity of the parent node minus the weighted sum of the Gini impurity of the two child nodes.

2017). There are several methods of encoding categorical variables. Ordinal encoding of categorical predictors has several benefits, including increased computational efficiency, evading restrictions on the number of predictor categories[2], and managing absent levels (chapter 4; Au, 2018; Smith et al., 2024b).The encoding method can be independent of the response variable (i.e., target-agnostic methods, such as one-hot encoding, integer encoding, and PCO-encoding (chapter 4; Smith et al., 2024b)) or can incorporate information about the target values associated with a given level (i.e., target-based methods, such as CA-encoding (Coppersmith et al., 1999; Wright and Ziegler, 2017) and CA-unbiased-encoding (chapter 4; Smith et al., 2024b)).

Encoding may be performed at different stages of the algorithm (figure 6.4). The most computationally efficient method is to encode the predictor variables prior to bagging (i.e., once on the entire dataset rather than each sub-sample undergoing encoding independently) (Wright and Ziegler, 2017). Encoding can also take place after bagging (i.e., on each sub-sample or at each split in the tree (Breiman, 1996; Liaw and Wiener, 2002)); however, this has a much higher computational cost.

Target-based encoding methods necessarily have information leakage from the target variable to the predictors. If a predictor is encoded prior to splitting into training and test sets, information from the target variable in the test set will leak to the predictors in the training set by way of the *a priori* encoding. In the same way, if a predictor is encoded prior to bagging, information from the target variable in the bootstrap samples will leak to the predictors in the OOB samples. The OOB observations will not, therefore, behave like fully independent test data. Target-agnostic encoding methods do not have this issue with information-leakage because the response class (target) information is not used for the encoding.

Treating the OOB samples like an independent test set is therefore only reasonable if a target-agnostic encoding method is used, or if a target-based encoding method is performed

[2]When nominal encoding a categorical variable, each binary node assignment is saved using the bit representation of a double integer, which limits this treatment to predictors with fewer than 54 levels (Wright and König, 2019).

subsequent to bagging. Otherwise, calculating misclassification rates and measures of variable importance on the OOB sample, or indeed the encoded variables, as in the case of the holdout variable importance (Janitza et al., 2018; Wright and Ziegler, 2017), is likely to underestimate the true error rate and overestimate the variable importance. The impact of method and timing of encoding has not been explicitly examined with regards to random forest OOB sample calculations.

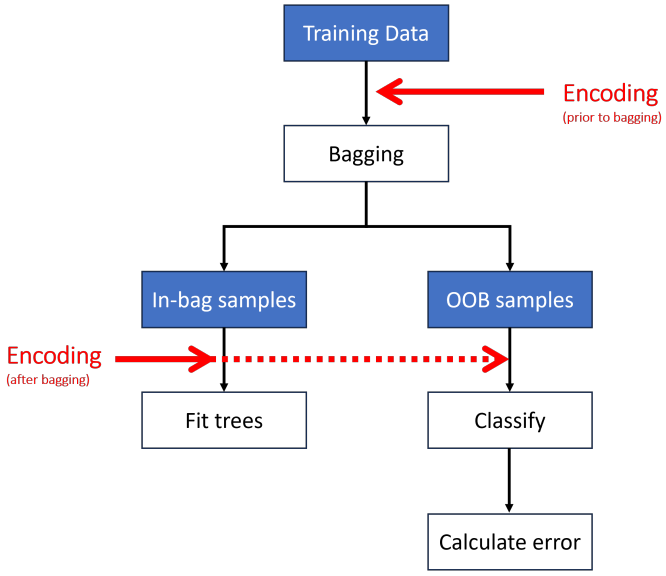

Figure 6.4: Encoding may take place prior to or after creating the out-of-bag (OOB) samples.

6.1.4 Study Aims and Objectives

Encoding of categorical variables is a necessary preprocessing step for many machine learning algorithms. The computational benefits of ordinal encoding categorical variables are well known. The potential leakage of target information to the OOB samples as a result of target encoding categorical variables prior to bagging is currently unreported. Current debates lie in the accuracy of OOB error estimates and/or variable importance measures, particularly for small sample sizes and unbalanced designs, but there appears to be no awareness that OOB samples may not be as 'good as an independent test set' and it remains a commonly held belief that OOB samples replace the need for separate test data.

For some popular random forest implementations (e.g., the R package `ranger` (Wright and Ziegler, 2017)), target encoding of categorical predictors prior to bagging is the recommended approach (Wright and König, 2019), and is performed internally within the method, in parallel with OOB error and variable importance calculations. This has potentially resulted in biased and even misleading results in a earlier studies.

In this chapter, the accuracy of OOB error estimates and variable importance measures are investigated when nominal categorical variables are ordinal encoded prior to bagging in random forest models. A random noise simulation study is performed that compares how target-based

versus target-agnostic encoding of categorical predictor variables can affect the OOB error and estimates of variable importance. This demonstrates that when target-based encoding is performed prior to bagging, OOB samples are biased due to information leakage from the target variable during the encoding process. It is therefore recommended to use a separate test set instead of the OOB sample, or else to perform the encoding after bagging.

Although this chapter focuses on random forest which incorporates bagging as a key component of the method, these results are generalisable to any applications which employ bagging (bootstrap aggregating), including other ensemble learning techniques, classification and regression tasks (Dfuf et al., 2020), outlier predictions (Mohandoss et al., 2021), feature selection (Calle et al., 2011; Deviaene et al., 2019; Díaz-Uriarte and Alvarez de Andrés, 2006), model tuning (Dauda, 2022), Gini-OOB index (Chen et al., 2023), and clustering (Bigdeli et al., 2022; Schumacher et al., 2016).

The aim of this study is to raise awareness of this simple, yet important and previously unreported, issue. Specifically, the goals are to:
(i) demonstrate why the common practice of using OOB samples instead of independent test data can lead to biased and potentially misleading results due to information leakage from the target variable during the process of encoding categorical predictors;
(ii) investigate via a short simulation study the accuracy of OOB error estimates and variable importance measures when nominal categorical variables are ordinal encoded prior to bagging in random forest models;
(iii) highlight the benefits of using independent test data for calculation of error estimates and variable importance measures; and
(iv) introduce the new 'independent holdout method' for calculating variable importance.

6.2 Methods

6.2.1 Implementation

There are many popular implementations of random forest, including over 20 packages in R[3] as well as the widely used Python machine learning library `scikit-learn` (Pedregosa et al., 2011). There is no single best implementation and most are optimised for some special property of the data (Wright and Ziegler, 2017). Algorithms do, however, differ in their treatment of categorical variables, including absent levels (i.e., levels of a predictor variable that are present in data for prediction that were not present when the random forest was trained) (table 6.1), which may impact predictions (chapter 4; Au, 2018; Smith et al., 2024b) and performance measures calculated from OOB samples.

An unordered (nominal) categorical predictor with k levels has $2^{k-1} - 1$ possible binary splits. A random forest algorithm may search the set of possible splits, either exhaustively (e.g.,

[3]https://koalaverse.github.io/machine-learning-in-R/random-forest.html#random-forest-software-in-r.

`randomForest`[4] (Liaw and Wiener, 2002) and `ranger`[5] (Wright and König, 2019)), or partially (e.g., `randomForestSRC`[6] (Ishwaran and Kogalur, 2023)). As each binary node assignment is saved using the bit representation of a double integer the exhaustive search option is limited to predictors with fewer than 54 levels. If the categorical predictor is defined as a character vector (i.e., rather than an unordered factor) it may, by default, be encoded alphabetically (e.g., `randomForest`) rather than converted to a factor (e.g., `ranger`). This is problematic if a separate data set (i.e., for prediction) has a different set of levels to those in the training set, in which case the ordinal encoding of the two sets will not match. This will occur if the observations for prediction contain only a subset of the levels from the training set, or if there are absent levels.

An ordered categorical predictor with k levels can be treated the same way as a numerical predictor with k unique ordered values and, at most, $k - 1$ possible split points. Again, care needs to be taken when the levels in the data to be predicted do not match exactly the levels in the training set as, for some algorithms (e.g., `randomForest`), the encoding of the levels may not match. For the case of two-class classification, a nominal predictor variable with k levels may be ordered by the proportion of observations with the second response class in each level. The ordering may occur at each split (e.g., `randomForest`[7]), or once prior to growing the forest (e.g., `ranger`[8]). Subsequently, treating these variables as ordinal leads to identical splits in the random forest optimisation as considering all possible 2-partitions of the k predictor levels (Breiman et al., 1984; Ripley, 1996). For multi-class classifications, an order may be imposed on a nominal variable alphabetically (e.g., `ranger`[9]), or according to the first principal component of the weighted covariance matrix of class probabilities, following Coppersmith, Hong, and Hosking (1999)[10] (e.g., `ranger`[11]). Ordering the variables once on the entire dataset prior to bagging, rather than at each split, is computationally efficient and negates the upper limit on the number of variable levels (Wright and König, 2019).

Some implementations of random forest require categorical variables to be one-hot encoded prior to analysis (e.g., Python's `scikit-learn`). This means a single predictor with k levels is replaced by $k-1$ indicator variables. Now there will be only a single possible split point at each node but from $k - 1$ indicator variables. Using this method, some of the category levels will be randomly ignored for each split, and so the original predictor will be represented by j binary predictors, where $j \leq k - 1$.

Treatment of absent levels also differs between implementations. Some algorithms are unable to process absent levels of unordered factors at all (e.g., `randomForest`). Some treat absent levels as missing values, or if there are no true missing values will map them to the child node

[4] This is the default option for `randomForest` in the case of multi-class classification or two-class classification with predictors which have fewer than 10 levels.

[5] when the argument `respect.unordered.factors` is set to `"partition"`

[6] https://www.randomforestsrc.org/articles/getstarted.html#allowable-data-types-and-factors

[7] this optimisation proceeds when the predictor variable has more than 10 unordered levels

[8] when the argument `respect.unordered.factors` is set to `"order"` or `TRUE`

[9] when the argument `respect.unordered.factors` is set to `"ignore"` or `FALSE`

[10] Coppersmith, Hong & Hosking (1999) use the first principal component of the weighted matrix of class probabilities.

[11] when the argument `respect.unordered.factors` is set to `"order"` or `TRUE`

that has the most samples (e.g., `scikit-learn`[12] and `randomForestSRC`). And some will send all observations with an absent level to a particular branch at any given node (e.g., `ranger`[13]) (chapter 4; Smith et al., 2024b).

The method of treatment of categorical variables, including absent levels, by four popular implementations of random forest is summarised in table 6.1.

Table 6.1: Implementation specific treatment of categorical variables.

Implementation	**Predictor type**	**Predictor treatment**	**Handles absent levels**	**Timing of encoding**	**Maximum levels**
ranger	character vector	converts to unordered factor	yes	-	-
	ordered factor	treats as ordinal	yes	-	-
	unordered factor	exhaustive partition	yes	-	53 levels
		orders alphabetically	yes	before bagging	-
		target encodes	yes	before bagging	-
randomForest	character vector	orders alphabetically	yes[1]	before bagging	-
	ordered factor	treats as ordinal	yes[1]	-	53 levels
	unordered factor	exhaustive partition	no	-	53 levels
		target encodes[2]	no	after bagging	53 levels
randomForestSRC	character vector	unable to process	-	-	-
	ordered factor	treats as ordinal	yes[3]	-	-
	unordered factor	partial partition	yes[3]	-	-
scikit-learn	character vector	one hot encoding	yes[3]	before bagging	-
	ordered factor	one hot encoding	yes[3]	before bagging	-
	unordered factor	one hot encoding	yes[3]	before bagging	-

[1] the absent levels need to be ordered last for consistency of encoding with the training set.
[2] optimisation is employed in the case of 2-class classification when there are more than ten levels of a predictor variable.
[3] treats absent levels as missing values.

6.2.2 Simulation Study

To investigate the accuracy of internally calculated misclassification rates and variable importance under null conditions, a set of data was simulated and analysed with random forest.

The simulated data consisted of n individuals, each with one predictor variable allocated uniformly and with replacement from k levels. One of three classification labels were randomly assigned to each individual. There was no relationship between the response and the predictors. A subset containing 80% of the observations was used for training the random forest, and the remaining 20% of observations were used as the set of testing data. The process was repeated for each combination of sample size $n \in \{20, 50, 100, 150, 200, 400\}$ and number of variable levels $k \in \{1, 5, 10, 35, 50, 100, 150, 200\}$.

For each random forest, the misclassification rate was calculated using each of two methods:
(i) the OOB sample; and
(ii) the misclassification rate of the observations in the testing data.

[12] https://scikit-learn.org/stable/modules/ensemble.html#random-forests
[13] https://github.com/imbs-hl/ranger/blob/master/R/predict.R#L167

In addition, for each random forest, the variable importance was calculated using each of five methods:
(i) the original MDI method, *sensu* Breiman (2002);
(ii) the original MDA method, *sensu* Breiman (2001);
(iii) the actual impurity reduction (AIR) importance (Nembrini et al., 2018; Sandri and Zuccolotto, 2008);
(iv) the holdout variable importance (Janitza et al., 2018); and
(v) the independent holdout method which is the holdout method but using cross-validation folds which have been separated prior to encoding.

For each combination of parameters, 99 sets of data were generated and a random forest was trained with 500 trees and the Gini index splitting rule. The levels of the predictor variables were integer encoded according to the alphabetical ordering of the levels and the average misclassification rate and variable importance were recorded for each method. The process was then repeated with the levels of the predictor variables being target encoded based on class probabilities.

The `ranger()` function from the R package `ranger` (Wright and Ziegler, 2017) offers both target-based and target-agnostic encoding options internal to the function and was used for the analysis; however, analysis by a different implementation using pre-encoded predictor variables would lead to equivalent results.

6.2.3 Code Availability

All analyses were carried out using R version 4.3.1 (R Core Team, 2024) and the `ranger` package ("RANdom forest GEneRator") version 0.15.1 (Wright and Ziegler, 2017). The R code used in this study is available at https://github.com/smithhelen/OutOfTheBag/releases/tag/v.1.0.0. This includes the code to generate the simulated data for reproducibility.

6.3 Results

6.3.1 Out-of-Bag Error

In the ideal case of balanced data with random assignment of individuals, the misclassification rate with simulated data was expected to be $\frac{2}{3} \approx 0.67$ regardless of the sample size, number of predictor levels, or method of encoding predictor variables. This was indeed the case when the misclassification rate was calculated for a fully withheld independent test set - except with a small sample size of 20. However, the internally calculated OOB error rate depended on the method used to encode the levels of the categorical predictor variables. When predictor levels were integer encoded based on alphabetical placement, the misclassification rate was 0.67, as expected; however, when the predictor levels were target encoded based on the first principal component of the weighted covariance matrix of class probabilities, the misclassification rate decreased with increasing numbers of factor levels, and this was compounded with smaller

sample sizes (figure 6.5). The bias was further exacerbated with increasing number of predictor variables (appendix C.1).

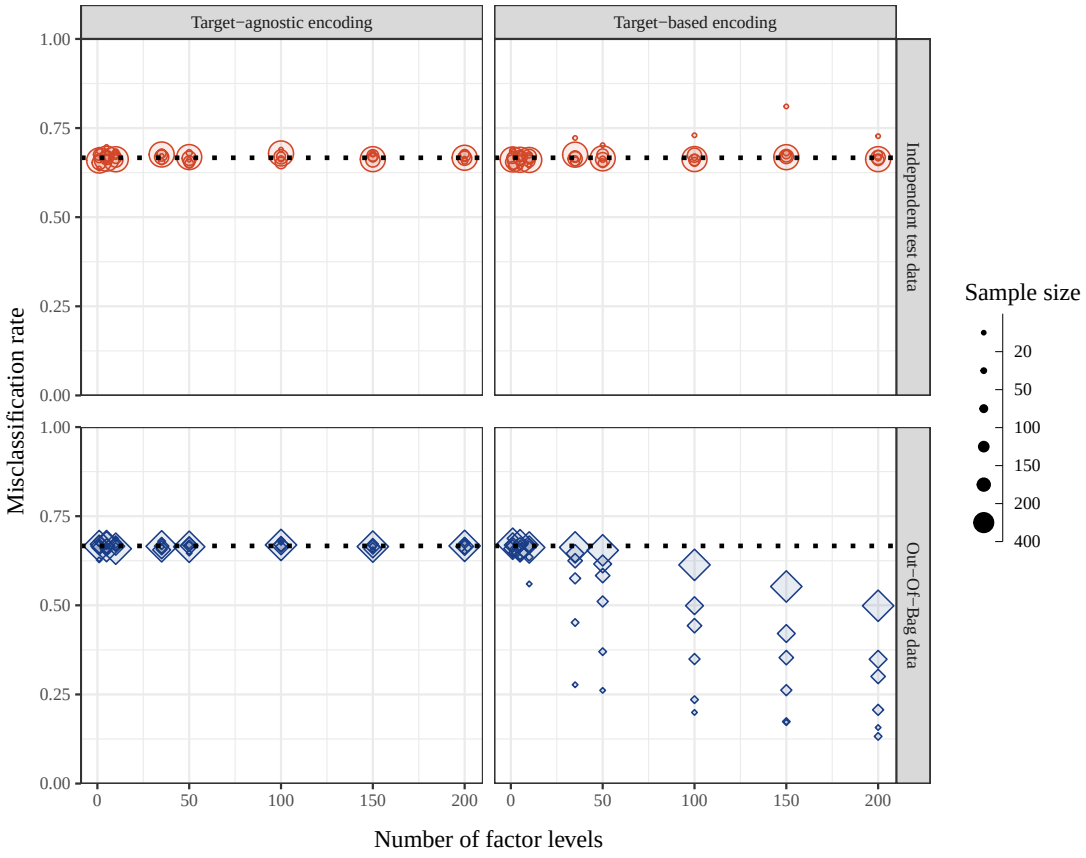

Figure 6.5: Misclassification rates of data simulated with balanced design and random assignment of individuals to one of three classes as calculated via independent test set (top panel, red circles) and internal OOB sample (bottom panel, blue diamonds) when the method of encoding predictor variables is target-agnostic (ordered (alpha)numerically, left panel) or target-based (ordered via principal component analysis (PCA) of class probabilities, right panel). The dotted line indicates the expected misclassification rate under the simulated null conditions.

6.3.2 Variable Importance Measures

The average variable importance was also expected to be impervious to the method of encoding predictor variables, and, under random assignment of variable levels and of individuals, the variable importance was expected to be zero. The independent holdout method was the only method that returned the expected outcome (i.e., zero importance for both target-agnostic and target-based encoding methods). The MDI measure, which is calculated on in-bag samples, was not affected by the choice of encoding; however, MDI increased with both sample size and number of variables for both target-agnostic and target-based encoding methods. Each of the other three variable importance measures were influenced by the choice of encoding method. Although the holdout method does not directly use OOB samples for its calculations, because it is performing the predictor encoding on the entire dataset, prior to splitting into cross-validation

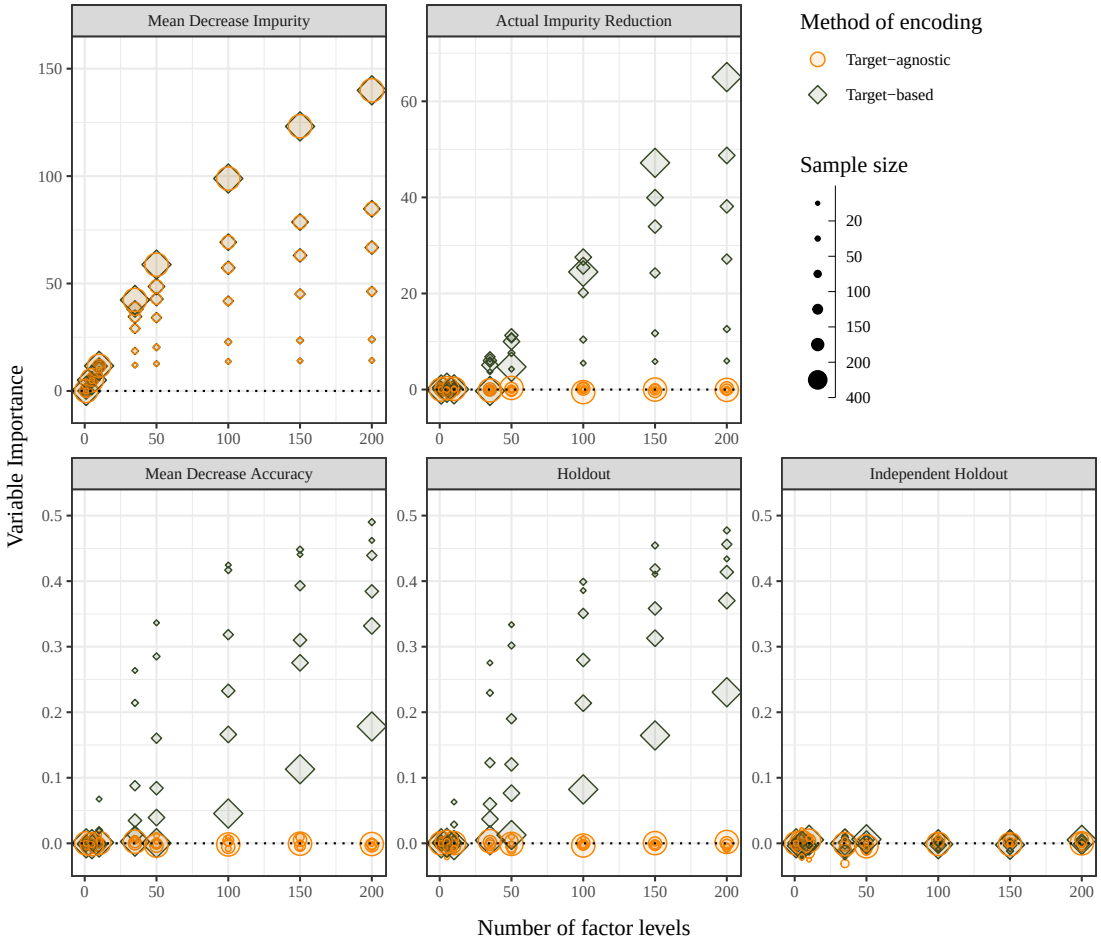

Figure 6.6: Average variable importance as calculated using the five methods when the method of encoding predictor variables is target-agnostic (circles; encoded as integers) or target-based (diamonds; encoded via principal component analysis (PCA) of class probabilities).

folds, it is affected in the same manner. When predictor levels were integer encoded (i.e., target-agnostic), the variable importance values were zero as expected; however, when the predictor levels were target encoded, the average variable importance increased with increasing numbers of factor levels. For the MDA and holdout methods, this was compounded with smaller sample sizes, but the opposite was true for AIR, which showed greater bias for larger sample sizes (figure 6.6). In contrast with the OOB misclassification rate, the positive bias diminished with increasing number of predictor variables (appendix C.2).

6.4 Discussion

Random forest predictive models are well suited to data sets containing a large number of categorical predictors and/or predictors containing many levels. Such data presents challenges for predictive models including absent levels and high computational demands. In these cases, one-hot encoding is not recommended as it frequently leads to a prohibitively large number of binary variables. Ordinal encoding, however, may improve both predictive performance and efficiency of models and offers a solution to the absent-levels problem (chapter 4; Smith et al., 2024b).

Methods of ordinal encoding of categorical predictors may be dependent or independent of the target variable. Target-based methods of encoding, including the two-class optimisation employed by `randomForest` and `ranger`, and ordering according to the first principal component of the weighted matrix of class probabilities, as implemented in `ranger`[14], use information from the target variable to inform the ordering. When the encoding is performed prior to bagging, there is leakage of information from the target variable to the observations in the OOB set. The leakage occurs because, even when the observations are out of bag, the encoding of their corresponding levels was informed from the entire training dataset (i.e., prior to the observations moving OOB) based on the correct response classes (i.e., the target). This means the OOB observations do not behave like fully independent test data.

Target-agnostic methods of encoding, such as the naïve alphabetical encoding, or ordering according to some characteristic of the data (e.g., the PCO-encoding method (chapter 4; Smith et al., 2024b)), are not subject to the issue of data leakage because the levels are encoded using data on the predictor variables only – the response class (target) information is not used. Therefore, in these cases, it is entirely appropriate to treat OOB observations like fully independent test data.

Breiman (1996, 2001) claimed that the OOB sample was as reliable as using an independent set of data for testing. This study found that, for random forest models, different methods of encoding nominal variables had important implications for the accuracy of calculations performed on OOB samples. It showed that the OOB misclassification rate, and the variable importance measures which utilise OOB samples (the MDA, holdout, and AIR measures), were biased when using a target-based encoding method due to 'data-leakage' during the *a priori* encoding

[14] also the CA-unbiased variation described in chapter 4 and Smith et al. (2024b)

of categorical predictors. When the encoding method is target-based, and the encoding is performed prior to bagging, the OOB data underestimates the true rate of misclassification, and overestimates true variable importance.

In all cases the bias increases with increasing number of factor levels, and is influenced by sample size. The effect of information leakage on OOB misclassification rates is more pronounced with smaller sample sizes and leads to lower misclassification rates (higher accuracy). The information leakage does not affect the permuted variable, as the relationship with the target is broken, and therefore MDA and holdout variable importance measures both increase with decreasing sample size. In contrast, when variable importance is measured using purity of splits (e.g., the AIR method), rather than misclassification rates, information leakage has a more pronounced effect when sample sizes are larger. Although the MDI measure is not affected by method of encoding, it is also dependent on sample size. For both the MDI and the AIR methods, increasing sample size results in better purity of splits leading to higher variable importance values. MDI is known to be biased in favour of variables with many possible split points (Strobl et al., 2007). Larger sample sizes represent a greater number of factor levels, and therefore split points, which is artificially inflating variable importance.

A potential solution to the problem of information leakage to the OOB sample is to order the levels of each bootstrap sample independently (i.e., rather than ordering once on the entire dataset prior to bagging) (figure 6.4). It is noted that there are currently no implementations of random forest which offer encoding after bagging for the multiclass case. Another option is to calculate misclassification rates and variable importance measures on truly independent test data.

The findings of this chapter have several important research and practical implications for machine learning practitioners. The aim is not to recommend a particular variable importance measure or error estimation technique, but rather to discard the belief that OOB samples are a replacement for independent test data in all instances. This is not an issue for numeric or ordinal data. But for nominal categorical predictors which are ordinal encoded using a target-based encoding method, it can be recommended to calculate misclassification rates from a separate, fully independent, test dataset; and to calculate variable importance via MDA using an independent test set as the holdout sample.

6.5 Conclusion

This chapter highlights how different methods of encoding of categorical predictors can bias OOB misclassification rates and variable importance measures. For datasets with a high number of variables and/or variable levels, absent levels are likely and ordinal encoding is a sensible approach for both speed of analysis and accuracy of predictions. When levels of categorical predictor variables are target encoded using class probability information and when encoding occurs prior to bagging, OOB samples suffer information leakage and are not a replacement for an independent test set. Using OOB data in place of an independent test set will lead to inflated measures of accuracy and variable importance. These findings are applicable to random forest

and other tree-based methods (e.g., boosted trees) where OOB misclassification rates and/or variable importance measures are calculated.

Supplementary Information

This chapter has two accompanying supplementary files -

Appendix C.1 The effect of increasing number of variables on the out-of-bag misclassification rate.

Appendix C.2 The effect of increasing number of variables on measures of variable importance.

Chapter 7

Source Attribution of *Campylobacter* Species using Whole Genome Sequencing Data.

7.1 Introduction

Next-generation sequencing (NGS) technologies and decreasing costs have transformed the approach to genetic-based research, and whole genome sequencing (WGS) data is increasingly becoming the preferred choice for genetic analysis (Bagger et al., 2024). WGS data provides a high-resolution, base-by-base view of the entire genome, allowing the smallest of variations to be detected. This unprecedented genomic detail provides opportunities for source attribution models as it allows fine-scale differentiation of closely related strains.

The growth of WGS data has seen a shift towards the development, and application of, source attribution models utilising machine learning (ML) algorithms. ML algorithms, such as network analysis, random forest, and logit boost, are well suited for analysing large and complex datasets, and have been successfully applied to the source attribution of a range of pathogenic foodborne bacteria, including *Listeria monocytogenes* (Mughini-Gras et al., 2025; Tanui et al., 2022), *Salmonella Typhimurium* (Guillier et al., 2020; Guzinski et al., 2024; Lupolova et al., 2019; Munck et al., 2020; Thystrup et al., 2024; Zhang et al., 2019), and *Campylobacter* species (Arning et al., 2021; Brinch et al., 2023; Harrison et al., 2021; Wainaina et al., 2022). ML algorithms work by 'learning' from a set of observed (training) data and subsequently making inferences about new unobserved (test) data. For source attribution models, the set of isolates from animal host sources is the training data with which the model learns, and the test data is the set of isolates from humans.

Modelling vast amounts of data presents several challenges, however. The pure volume of WGS data can significantly impact computational efficiency, resulting in slow performance and requiring high processing power, memory and storage. In addition, the large number of unique genotypes presents a number of issues for ML models (Brinch et al., 2023; Munck et al., 2020; Zhang et al., 2019). One limitation of using genes as input data (e.g., core-genome multilocus sequence typing (cgMLST) data) is the substantial number of missing alleles, which occurs

when no matches can be found (e.g., in PubMLST), requiring imputation (Brinch et al., 2023; Guzinski et al., 2024; Mughini-Gras et al., 2025; Munck et al., 2020; Tanui et al., 2022), removal (Arning et al., 2021; Harrison et al., 2021), and/or recoding (Arning et al., 2021; Harrison et al., 2021; Pascoe et al., 2024). Another issue is the presence of alleles in the set of human isolates that are not in the set of source isolates (i.e., absent-levels), a problem that escalates with increasing numbers of genes (figure 3.2).

Several approaches have been used to address these issues. Instead of using genes, models have been implemented from other levels of input data, such as single nucleotide polymorphisms (SNPs) (Guzinski et al., 2024; Zhang et al., 2019), *k*-mers (i.e., short sequence regions *k* bases long) (Arning et al., 2021; Brinch et al., 2023; Munck et al., 2020; Thystrup et al., 2024), unitigs (Pascoe et al., 2024), or protein variants (Lupolova et al., 2019; Zhang et al., 2019). One benefit of using genes as input data, however, is the ability to identify genes that have the best prediction potential for attribution of organisms to different sources (Tanui et al., 2022). Studies wanting to use genes as input and retain allelic information have, instead, used methods of feature reduction to decrease complexity of the data (Brinch et al., 2023; Munck et al., 2020; Tanui et al., 2022; Thystrup et al., 2024). For example, Tanui et al. (2022) reduced 1748 core genes to 736 by removing genes with low variability across the genome on the assumption that they were redundant variables and would not display strong host signals. However, when the feature reduction is extreme, for example Munck et al. (2020) reduced 3002 core genes to 17 and Pascoe et al. (2024) reduced 1343 core genes to five, this negates both the advantage of having increased data resolution and the ability to detect host segregating genes.

An alternative approach is to numerically encode the alleles for each gene. For example, Mughini-Gras et al. (2025) used one-hot encoding (i.e., new (binary) columns were created, indicating the presence of each allele) of core genes for the source attribution of *Listeria*; Arning et al. (2021) and Tanui et al. (2022) each used nominal encoding (i.e., the nominally assigned allele numbers were interpreted directly as numerical predictors) of core genes for the source attribution of *Campylobacter* and *Listeria* species respectively; and new methods of ordinal encoding were introduced in chapters 4 and 5 of this thesis (see also Singh et al. (2025); Smith et al. (2024a,b)) that encoded core genes for the source attribution of *Campylobacter* on the basis of allelic sequencing information. Ordinal encoding of alleles enables the full complement of genes, including accessory genes and absent levels, to be included in the model and is computationally more efficient than one-hot encoding. Ordinal encoding is also strongly preferred over nominal encoding, which may introduce bias when the pattern of allele numbering is inadvertently associated with host source, as may occur due to the numerical assignment of alleles in contiguous chunks by source (appendix D.1). This is evident in the data used in the chapter by Arning et al. (2021), but is not considered or commented upon.

The CAP-encoding method (Smith et al., 2024a), introduced in chapter 5, uses a principal coordinates approach to numerically encode each allele, independently for each gene, allowing the genes to be treated as numerical predictor variables and overcoming the problem of missing alleles in data for prediction. The CAP-encoding method is a two step procedure. Firstly,

Gower's method of principal coordinates analysis is used to position each allele in the principal coordinate space of the gene, according to its similarity to each of the other alleles in the training data, based on some property of the gene (e.g., sequencing information). Secondly, a subset of these principal coordinates are rotated in order to maximally correlate with the direction of greatest variation in class probabilities, as determined by a scaled correspondence analysis on the contingency table of counts of alleles by source. The CAP-encoding method is particularly advantageous when there is a high number of observations and/or genes with absent levels, such as with WGS data (chapter 5).

In this chapter, a source attribution analysis of *Campylobacter jejuni* and *Campylobacter coli* is performed using a random forest approach, based on WGS data collected from New Zealand between 2018 and 2019 as described in Lake et al. (2021). The CAP-encoding method (chapter 5; Smith et al., 2024a) is used to encode the alleles within each gene. This method uses a target-based approach to encoding, by using both the source frequency of whole-genome mulitlocus sequence typing (wgMLST) allele profiles and the Hamming distances of the nucleotide sequencing information between each pair of alleles. The self-attribution results (i.e., the misclassification rate of isolates from a known source) are compared when the model is trained using the traditional 7-loci mulitlocus sequence typing (7-loci MLST) genes *versus* the 1343 core-genome mulitlocus sequence typing (cgMLST) genes *versus* the full wgMLST pangenome, to determine the effect of increasing the resolution of the data and including accessory genes (i.e., genes which are not found in the majority of isolates). In addition the results are compared using two adjusted distance measures - (i) the residualised distance matrix, as described in section 3.2.3 of chapter 3, which removes the variation due to clonal complex (CC); and (ii) the recombinant-adjusted distance matrix, which down-weights each nucleotide position by its degree of recombination. Finally, the chapter aims to estimate the relative contributions of different sources of campylobacteriosis in New Zealand and compare the results against those obtained from the same dataset using the asymmetric island model on 7-gene MLST data (Lake et al., 2021). The 'independent holdout variable importance' method, introduced in chapter 6, is used to determine which genes are most influential to the attribution outcome.

7.2 Methods

7.2.1 Dataset

The data used for this analysis is from a source-assigned case-control study of notified human cases of campylobacteriosis in New Zealand, between 2018-2019, as presented in Section 3.2 of chapter 3 (see also Lake et al. (2021)). Briefly, the dataset consists of WGS data from 1211 *C. jejuni* and *C. coli* isolates cultured from humans (n=651), poultry (n=205), sheep (n=187), and cattle (n=168). cgMLST allele profiles were obtained as previously described. To identify the accessory genes, i.e., the non-core genes, the pangenome was analysed with pagoo (Ferrés and Iraola, 2021) to firstly identify all genes. All gene sequences were then compared, isolate by isolate, against known cgMLST alleles from the PubMLST *Campylobacter* database (Cody

et al., 2017) using BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) and the non-matched genes were classified as accessory genes. Allele numbers were assigned to each accessory gene based on unique sequences. Genes were then aligned using mafft (Katoh and Standley, 2013; Katoh et al., 2002).[1] This resulted in 5675 accessory genes being identified across the isolates, however 1360 of these were only found in single isolates, three had only a single allele across all isolates, and 291 had only a single allele across the full set of animal isolates, and so were excluded from the analysis. The final dataset therefore consisted of 5364 genes, made up of 1343 cgMLST genes and 4021 accessory genes.

7.2.2 Resemblance Measures

The nucleotide sequence of each allele, for each gene, was used to calculate a matrix of Hamming distances between each pair of alleles within each gene. At the genomic scale, *C. jejuni* and *C. coli* are highly diverse species (Dingle et al., 2001; Manning et al., 2003) and variation within a gene is not consistent across the genome (Parkhill et al., 2000; Sheppard and Maiden, 2015). The high genetic diversity results from clonal structure (i.e., relatedness), in addition to frequent intra- and inter-species recombination (Dingle et al., 2001; Manning et al., 2003; Suerbaum et al., 2001; Yu et al., 2012). The shared MLST nucleotide sequence-based typing method for *C. jejuni* and *C. coli* assigns a unique allele identification number to each unique variant for each locus when using the PubMLST allele sets as a reference (Dingle et al., 2005). The original MLST scheme considered only the nucleotide sequences of internal fragments of seven housekeeping genes accounting for a total of approximately 0.1% of the genome (Dingle et al., 2001). The corresponding allelic profile from the seven loci define the sequence type (ST) of each isolate and subsequently the clonal complexes (CCs). CCs are groups of isolates based on 7-gene MLST sequence variation that represent lineages presumed to have derived from a common ancestor (Dingle et al., 2002; Manning et al., 2003). CC are assigned a number based on a central genotype derived from the BURST (based upon related STs) clustering algorithm and are named after the central genotype. The original MLST scheme has been expanded to a core-genome scheme (cgMLST), defined as a set of 1343 loci which are present in most members of *C. jejuni* and *C. coli* (Cody et al., 2017), and a whole-genome scheme (wgMLST) (Cody et al., 2013; Maiden et al., 2013), in which all the loci of a given isolate are compared to equivalent loci in other isolates. In all MLST schemes, the allele is the unit of comparison, representing single genetic events, and the number of different nucleotides between strains is ignored. This is in contrast with the Hamming distance, which is directly proportional to the number of nucleotides that differ between strains. A large Hamming distance could, therefore, reflect a large number of single point mutations, or a single large scale evolutionary event, such as a homologous recombination event.

An attempt was made to account for the nature of clonal structure and recombination within the matrix of Hamming distances in two ways. In the first instance, a dissimilarity-based multivariate analogue to a residual plot, from which the variation due to clonal structure is removed,

[1] https://github.com/jmarshallnz/cgmlst

was calculated for each gene following the method described in Anderson (2017), and detailed in section 3.2.3. The premise is that the variation remaining in the residualised distance matrix can be directly attributed to differences associated with host source (figure 3.8).

In the second instance, the calculation of the Hamming distance was adjusted to down-weight each nucleotide position according to its degree of recombination. Regions of recombination were detected using Gubbins (Croucher et al., 2015), and the degree of recombination was estimated by counting the number of times each nucleotide position was part of a recombinant region (Hadfield et al., 2017).

7.2.3 Random Forest

7.2.3.1 Cross Validation

To assess the effect of increasing the number of genes in the analysis, the 560 isolates collected from animals were subject to ten-fold cross-validation for each set of data (7-gene MLST, 1343-gene cgMLST, and 5364-gene wgMLST data) using the same random number seed to ensure the same isolates were in each fold for each method. Likewise, to assess the effect of adjusting the distance measure to account for clonal structure and recombination, the 560 isolates collected from animals were subject to ten-fold cross-validation for each of the three sets of distance matrices (Hamming, Hamming residualised on CC, and recombinant-adjusted Hamming distance). The residualised distance matrices were calculated independently for each fold as the relative proportions of CCs is fold dependent.

In all cases, ten independent random forest models were run (one on each of the ten folds) allowing each of the isolates to be represented exactly once in testing data. Each forest consisted of 500 trees and used the Gini index splitting rule. Prior to training the forest, each gene was numerically encoded using the CAP-encoding method and the distance matrix of choice. As there are three classes, each nominal predictor (i.e., gene with alleles as levels) was replaced with two numerical predictors (encoded CAP scores). The subset of PCO dimensions for rotation was selected such that it captured at least 95% of the variation in the multivariate data cloud. When the residualised distance matrix was used, the residualised variable (CC) was included as an additional predictor variable and encoded using the CA-unbiased-encoding method (chapter 4; Smith et al., 2024b) so that any absent levels were scored as zero, in line with the *a priori* assumption of equal class probability.

The scores from the genes in the training set were transferred to the genes in the test set, and any new alleles (absent levels) were encoded based on their proximity to known alleles in the rotated principal coordinate space. Model performance was assessed by calculating the proportion of incorrect classifications on the set of test data for each fold and calculating the average and standard error, accounting for any variation between folds.

7.2.3.2 Attribution

To predict the source of the 651 human isolates, a random forest was trained using the full set of 560 source isolates, and 5364 wgMLST genes with their corresponding matrices of Hamming distances. Following training, the random forest model was used to predict the original host source for each of the human isolates. Estimates of uncertainty were calculated using a probability forest with the same set of parameters as the original random forest. For each tree in the forest, the probability of each human isolate being attributed to each source was calculated. The mean of these probabilities over the set of human isolates gives an average probability of attribution to each source for each tree. The 2.5% and 97.5% quantiles were then determined from this set of mean probabilities to give a 95% uncertainty interval.

Variable importance was calculated using the independent holdout method described in chapter 6 and Smith et al. (2024c). This computes the mean decrease in accuracy (MDA, or permutation importance (Breiman, 2001)) by permuting values of the variable in a second cross-validation (test) fold which has been separated prior to encoding and computing the difference in the error rate on the permuted test fold from the original test fold.

7.2.4 Code Availability

All analyses were carried out using R version 4.3.1 (R Core Team, 2024) and the packages `ranger` ('RANdom forest GEneRator') version 0.15.1 (Wright and Ziegler, 2017) and `tidymodels` (Kuhn and Wickham, 2020). The R code used in this study is available at `https://github.com/smithhelen/SourceAttribution`. The sequence reads used for this study can be accessed from the NCBI Sequence Read Archive under BioProject accession number PRJNA675916.

7.3 Results

7.3.1 Accessory Genes

The total dataset consisted of 5364 genes which were found in at least two isolates and had at least two alleles. This included 4021 accessory genes, in addition to the 1343 cgMLST genes. 20 of the cgMLST genes were found in fewer than 95% of the isolates, and 20 of the accessory genes were found in more than 95% of the isolates, including 12 that were found in every isolate (figure 7.1). Thus the core genome consisted of 1343 core genes, 98.5% of which are represented by the PubMLST cgMLST scheme.

7.3.2 Effect of Increasing Number of Genes

Overall, the cgMLST scheme and the wgMLST scheme had very similar misclassification rates (24.1% ± 1.4% and 24.5% ± 1.5% respectively), and both schemes had lower misclassification rates than the 7-gene MLST scheme (36.8% ± 1.5%). The difference between schemes was particularly dramatic for *C. coli* isolates from cattle (figure 7.2 (a)).

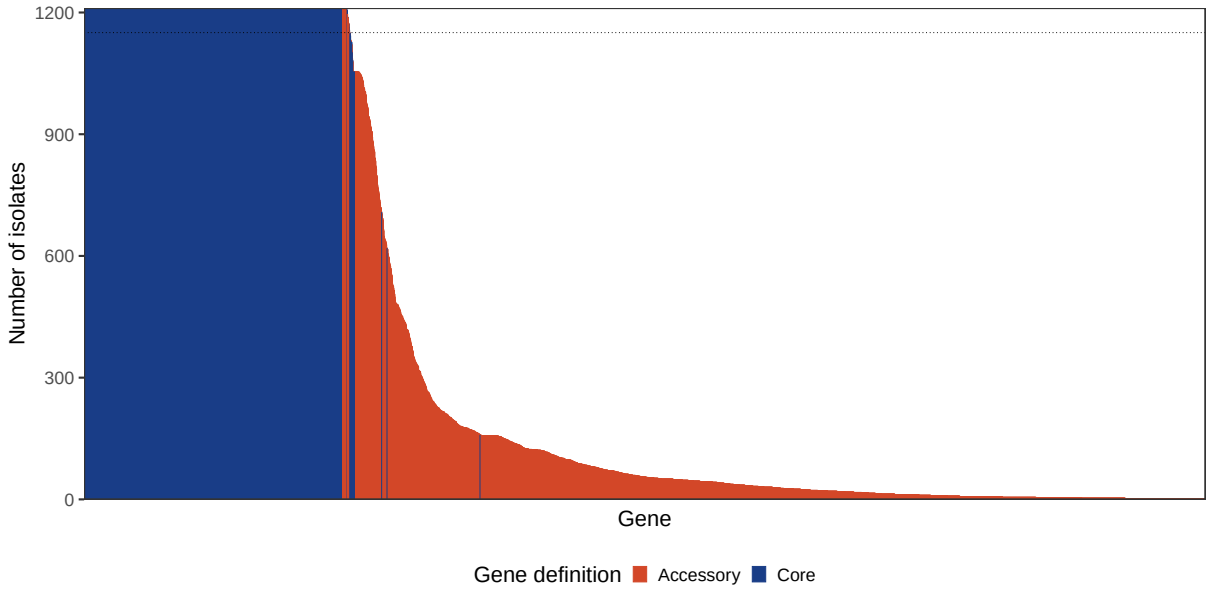

Figure 7.1: The number of isolates with each gene. Colour indicates the cgMLST (blue) and non-cgMLST (red) genes. The horizontal line marks 95% of the isolates.

Consistent with earlier studies, the accuracy of predictions was dependent on the class being predicted (figure 7.2). For all schemes, isolates sourced from chicken were the most accurately classified (74.1% – 86.4%), followed by isolates sourced from sheep (61.3% – 77.2%), and isolates sourced from cattle had the lowest classification success (52.6% – 61.6%).

In addition, the accuracy of predictions was dependent on the species of *Campylobacter* (figure 7.2(a)). For chicken and sheep isolates *C. coli* had better classification success than *C. jejuni*, however for cattle isolates *C. jejuni* had better classification success than *C. coli*.

7.3.3 Effect of Adjusting for Recombination

Overall, there was very little effect on misclassification rates of adjusting the Hamming distance for either the variation inherent in CC or for recombination (figure 7.3) (range 23.8% ± 0.6% – 24.3% ± 0.6%). Although the residualised distances resulted in the lowest overall misclassification rate, the adjustment was only advantageous to the classification of ruminant isolates and was detrimental to the classification of chicken isolates.

7.3.4 Attribution

The highest proportion of human cases were estimated to be attributed to chicken (49.9% – 64.7%), followed by cattle (17.5% – 30.7%), and the lowest proportion were estimated to be attributed to sheep (13.1% – 25.1%) (figure 7.4). Attribution of the same isolates using 7-gene MLST data and the asymmetric island model also estimated a higher degree of attribution to chicken followed by cattle, however this model estimated few isolates (0.08% – 2.6%) to be attributable to sheep (Lake et al., 2020). When a higher number of genes were used in the same

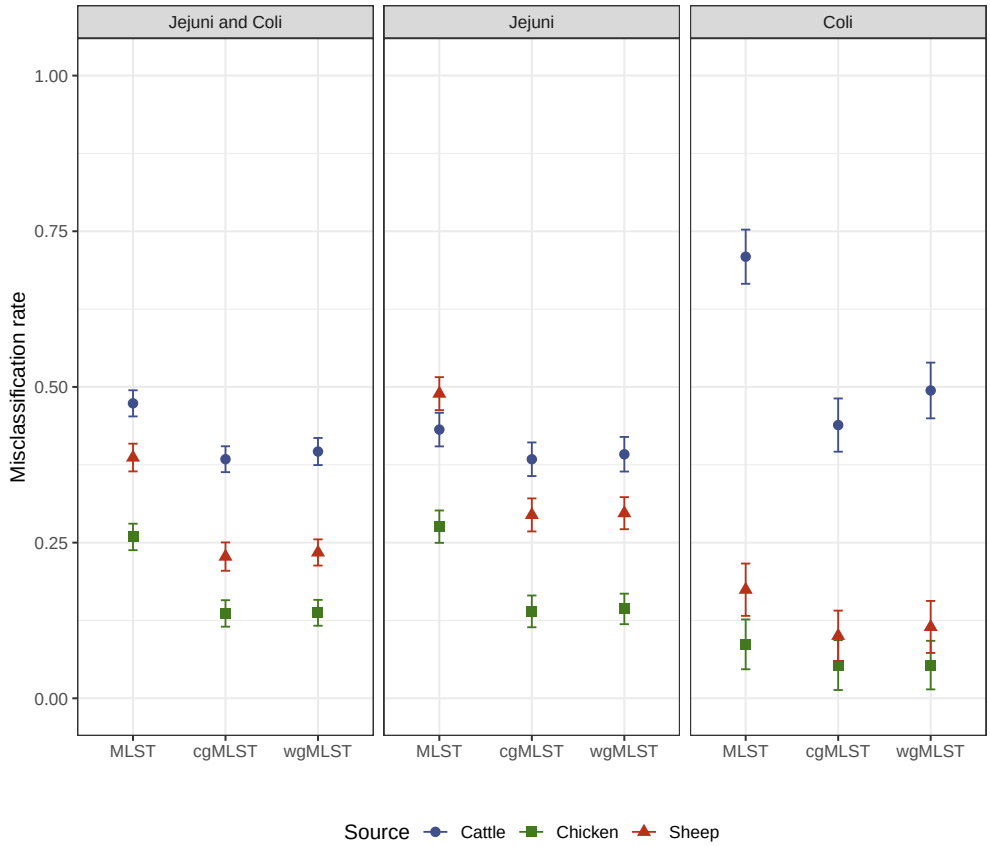

Figure 7.2: The effect of the number of genes on the accuracy of the random forest model for all isolates combined and for isolates separated by species.

model, a greater number of isolates were attributed to unknown sources (not sheep) (Lake et al., 2021).

7.3.4.1 Variable Importance

The independent holdout variable importance method aims to calculate the decrease in classification accuracy that results from replacing each variable in turn with random noise. The ten genes with the highest importance scores included a mix of core and accessory genes (table 7.1), however all genes had low importance scores, ranging from below zero to 0.0041. In other words, removal of the most 'important' gene only decreased the classification success of the model by 0.41%. In addition, selecting different subsets for the calculation of variable importance resulted in a different set of genes being identified as 'important' (appendix D.2). This suggests a high level of correlation between the genes. Although this was not investigated here, the order of importance could be considered alongside potential groupings of correlated genes.

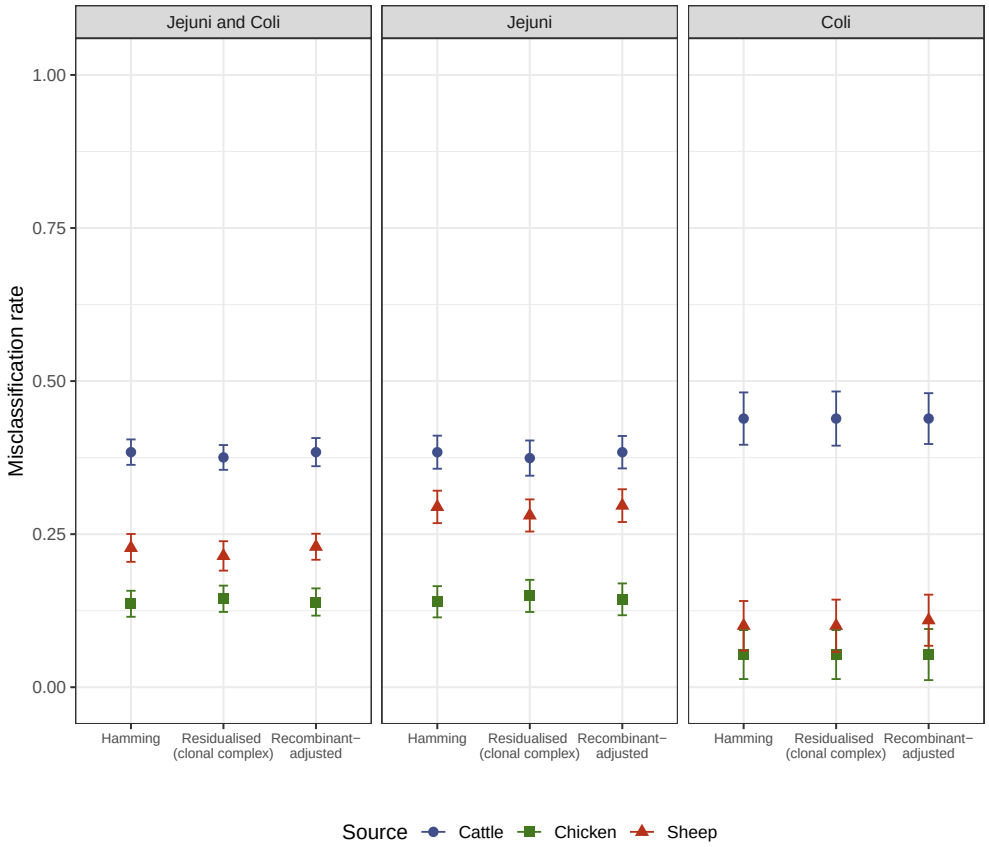

Figure 7.3: The effect of the distance measure on the accuracy of the random forest model for all isolates combined and for isolates separated by species.

7.4 Discussion

Whole genome sequencing data allows fine-scale differentiation of closely related strains and is rapidly becoming the preferred data choice for source attribution modelling. Several machine learning methods of source attribution have been developed for WGS data, including network analysis, random forest, and logit boost. These methods suffer from issues inherent with large volumes of categorical data such as computational issues and the absent-levels problem (i.e., unique categories in the data for prediction). This has led to source attribution studies applying methods of extreme feature reduction prior to analysis, which reduces the ability of models to harness the benefits of WGS data. Converting sequence data to numerical predictors is an alternative approach to avoid these issues (chapter 4; Smith et al., 2024b). *k*-merisation involves counting short sequence regions *k* bases long (*k*-mers) along the genome (Zielezinski et al., 2017) resulting in counts of potentially vast numbers of *k*-mers often necessitating preliminary feature reduction. Models using the *k*-merisation approach have shown accurate results at the expense of reduced interpretability and the opportunity to investigate the role of individual genes in the attribution. Integer encoding involves interpreting allele numbers as true numeric values and treating genes as continuous numerical predictors. As alleles have no inherent order (i.e.,

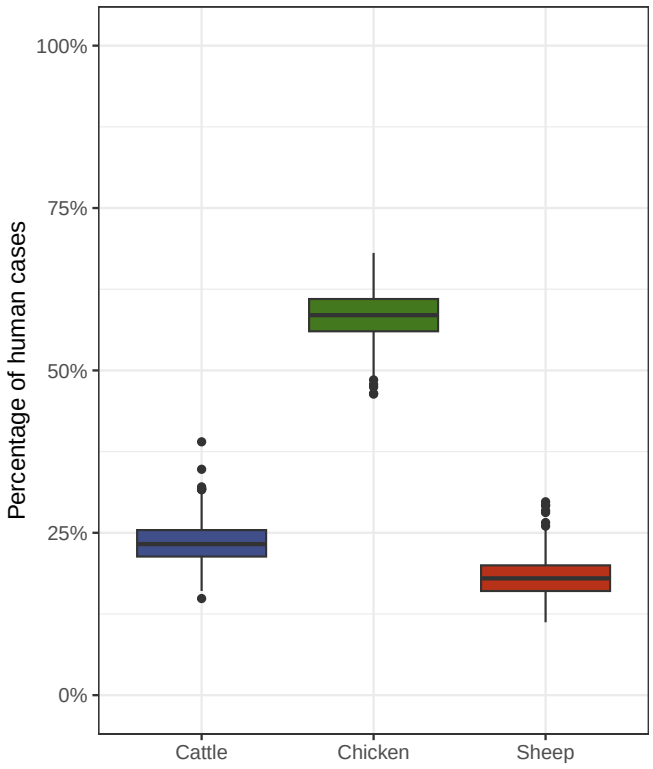

Figure 7.4: Source attribution for cases in a source-assigned case-control study of campylobacteriosis in New Zealand. Lower and upper tails are 2.5th and 97.5th percentiles of individual tree mean isolate probabilities (i.e., 95% uncertainty interval).

allele '1' is no more similar to allele '2' than it is to allele '202'), enforcing an order in this way is likely to introduce bias due to the method of allele designation in contiguous chunks by host. Recently new methods of encoding categorical variables have been developed which allow random forest models to use an unlimited number of variables and are unbiased. These methods involve ordering the categories (alleles) based on either their similarity to each other, or according to their relative host source probability. One such method is the target-based CAP-encoding method (chapter 5; Smith et al., 2024a) that uses a combination of sequence similarity and relative host probabilities. Alleles unique to the set of human isolates are encoded according to their similarity to each of the alleles in the set of animal isolates. Each of these methods is linked to the concept of genomic signatures and the hypothesis that isolates from the same source will be more related than isolates from different sources.

In line with this concept, the distance measure used for encoding should define alleles from the same source as having smaller distances to each other than alleles from a different host source. The Hamming distance measures the number of nucleotides that differ between sequences, on the assumption that alleles that share many nucleotides have similar host preferences. One complication is that most alleles are found in multiple sources, and the alleles that are found in single sources are more often than not only found in single isolates. The CAP-encoding method effectively adjusts the Hamming distances to account for the number

Table 7.1: The ten most important variables according to the independent holdout variable importance measure.

Locus	**Variable Importance**	**Function**
CAMP1076 (dnaX)	0.007143	DNA polymerase III, replicative synthesis (Bateman et al., 2025; Gomes et al., 2024)
CAMP1165 (uvrC)	0.005357	DNA repair (Bateman et al., 2025; Jehanne et al., 2020; Truglio et al., 2006)
CAMP0289	0.005357	No known function (Dai et al., 2017)
CAMP0751*	0.005357	No known function (Cody et al., 2017; Kovanen et al., 2014)
CAMP0975 (mutS)	0.003571	Putative mismatch repair protein, suppression of homologous recombination (Bateman et al., 2025)
CAMP1235 (psel)	0.003571	Generation of pseudaminic acid (Pse) (McDonald and Boyd, 2021)
CAMP1331*	0.003571	Methyltransferase, capsule biosynthesis (Fan et al., 2022; Revez et al., 2014)
tsdA*	0.003571	Tetrathionate reductase, energy metabolism (Liu et al., 2013)
ycaD 1*	0.003571	Amino acid transport/metabolism (Karki et al., 2023)
CAMP0031	0.003571	Type IIS restriction/modification enzyme (Anjum, 2013; Revez et al., 2014)

* accessory gene (i.e., not part of the cgMLST scheme).

of sources that alleles are found in, as it rotates the principal coordinates to align with the direction of greatest variation in host probabilities. Another issue with the Hamming distance is that single recombinant events result in large nucleotide differences between alleles and therefore dominate the matrix of Hamming distances. Accounting for this by down-weighting the contribution of nucleotides from recombinant regions did not, however, improve predictive performance of the source attribution model. Likewise, groups of closely related alleles that have evolved concurrently, will have smaller Hamming distances between them than between them and distantly related alleles, regardless of host preference. Adjusting the Hamming distance to remove distances between groups of isolates from different CCs effectively residualises the matrix such that the remaining variation contains differences due to host source, however this adjustment also did not improve predictive performance. This suggests that the ability of random forest models to repeatedly partition variables allows for the smaller differences resulting from host signature to be discerned, amongst the stronger patterns of clonal structure and recombination, removing the need to separate out clonal structure or recombination prior to attribution with random forest.

A major benefit of the CAP-encoding method is that it allows an unlimited number of genes to be included in the random forest model. Previous studies have indicated that WGS data enhances the ability to distinguish genetic variations and more accurately determine the origin

of infection-causing isolates. This study found that classification accuracy was greatly improved by the addition of the 1343 core genes over and above the seven MLST genes but that the addition of accessory genes did not further improve predictive accuracy. In addition, it found that there was no group of consistently important genes and that multiple genes shared the highest importance values (both core genes and accessory genes), presumably due to a high level of correlation between the genes.

This chapter used a random forest predictive model, with WGS data and the CAP method of encoding, based on Hamming distances between alleles of each gene, to attribute cases of *C. jejuni* and *C. coli* infections in New Zealand to putative livestock sources. The PubMLST defined core genome has been validated against isolates from the United Kingdom, Europe, and North America, but the level by which it is transferable to New Zealand has not been fully explored. The set of isolates from this source-attributed case-control study of campylobacteriosis in New Zealand is well represented by the cgMLST scheme with 97.5% of the cgMLST genes being part of the New Zealand core genome and only 2.5% of the core genome (20 genes) consisting of genes that are not defined by the cgMLST scheme.

In this chapter, chickens were estimated to be the main source of infection, followed by cattle and then sheep. These results largely agree with those of Lake et al. (2021) which were based on a reduced set of genes from the same set of isolates and used the asymmetric island model. However, the study by Lake et al. (2021) attributes an even greater proportion of cases to poultry and attributes very few cases to sheep. This chapter, based on wgMLST data, places a higher emphasis on cattle and sheep than previously described. Poultry meat has long been identified as a significant pathway for campylobacteriosis across developed countries (Cody et al., 2019), including New Zealand (ESR, 2023), where interventions in broiler meat production have successfully reduced incidence of infection (Muellner et al., 2011; Sears et al., 2011). This chapter highlights that cattle and sheep, either directly as foods or via their production environments, also contribute to human campylobacteriosis in New Zealand.

The independent holdout measure of variable importance provides an unbiased estimate of importance for each gene and may help to identify host-associated genes. While the importance analysis did not consistently highlight genes that had an association with host, including those that have previously been identified in other settings (e.g., pantothenate (vitamin B_5) (Sheppard et al., 2013b)) and used for source attribution (e.g., (Thépault et al., 2017)), one of the most 'important' genes, uvrC, has been identified as a potential host-segregating marker for *C. coli* (Jehanne et al., 2020). It is, however, unlikely that importance scores for individual genes will capture the collective importance of groups of correlated genes. Genes with a role in host segregation that are also highly correlated with other genes may not generate a high importance score because the effect of gene removal is negated by the presence of other (correlated) genes that serve a similar function. In contrast, a gene that has a minimal role in host segregation and that is not correlated with other genes may generate a higher importance score. This issue is clearly a complication of WGS data and is of less concern when datasets have fewer genes, or independent variables.

This is the first time that wgMLST allelic profiles have been used directly for source attribution of *Campylobacter*. This study has shown that the CAP-encoding method is an effective method for processing MLST data prior to analysis by machine learning methods. The CAP-encoding method is unbiased and is strongly preferred to encoding using allele numbers directly (chapter 7, appendix D.1). Because the CAP-encoding method utilises the target variable, calculations of error rates and variable importance must be made using an independent set of data and not the OOB data. This study found that the Hamming distance between pairs of alleles provides a good measure of similarity between alleles and captures the variation between host sources without the need for adjusting for clonal structure or recombination. It also found that there is no additional benefit of including accessory genes and that cgMLST data is a good resolution of data for analysis.

Despite the success of this method, limitations of using machine learning methods with WGS data are recognised. ML methods are not robust to imbalance in the training data and algorithms may not correctly classify minority sources. This may be approached by downsampling or upweighting the dataset (Lupolova et al., 2019; Munck et al., 2020) and this should be investigated further. In addition, classification is restricted to the sources represented in the training data, and all isolates will be attributed to one of these sources. Furthermore, finding an isolate in a host source does not necessarily imply that the host is the main preferred habitat of the isolate. These issues highlight the need for comprehensive and ongoing sampling strategies. The improvement, development, and generalisation of these methods is encouraged to continue to advance understanding and control of *Campylobacter* transmission. It is hoped there will be further exploration of, for example, alternative distance measures, the incorporation of parameter tuning (e.g., the proportion of genes considered at each split), and the inclusion of additional predictive features such as geographic location and timing.

7.5 Conclusion

Host associated genetic variation in *Campylobacter* species can be observed in both core and accessory genes. Including WGS data in source attribution models can provide better evidence to inform policy development and prioritise intervention strategies to control campylobacteriosis. Using WGS data with machine learning methods presents practical considerations which are offset by numerical encoding of wgMLST data. The CAP-encoding method, together with supplementary Hamming information, is the preferred method for encoding *Campylobacter* genes prior to analysis by random forest. While chicken continues to be identified as the most important reservoir of *C. jejuni* and *C. coli* in New Zealand, cattle and sheep also contribute to cases of human campylobacteriosis.

Supplementary Information

This chapter has two accompanying supplementary files -

Appendix D.1 The effect of nominal encoding of categorical variables.

Appendix D.2 Another ten most important variables according to the independent holdout variable importance measure.

Chapter 8

General Discussion

8.1 Out of the Woods

Campylobacteriosis

The principal aim of this thesis was to develop new tools for source attribution that could be applied to whole genome sequencing (WGS) data, motivated by campylobacteriosis. Foodborne diseases, such as campylobacteriosis, represent a significant risk to public health and contribute significantly to the global burden of disease. Estimating the relative contributions of the main sources of human infection of *Campylobacter* in New Zealand will help to inform public health decision making. *Campylobacter jejuni* and *Campylobacter coli* are the most common causes of bacterial gastroenteritis among human populations in developed countries, including New Zealand where campylobacteriosis is consistently the highest occurring notifiable disease and is the largest contributor to the economic costs of foodborne disease (Sears et al., 2011). Infection typically results in acute gastroenteritis with associated fever, abdominal cramping, diarrhoea, vomiting and headaches. Most infections are self-limiting; however, more serious complications can occur, such as Irritable Bowel Syndrome (IBS) and Guillain-Barré Syndrome (GBS).

Campylobacter is ubiquitous in the environment and forms part of the commensal microbiota of numerous wild, farmed and companion animals. Due to the ubiquity of *C. jejuni* and *C. coli* in animal intestines, and its shedding into the environment, transmission can occur via numerous routes; however, the principal source of human infection in developed countries is contaminated retail meat products, especially chicken meat (Nohra et al., 2020; Pascoe et al., 2024).

Molecular typing

Molecular typing methods for *Campylobacter* have been used to differentiate between species, and to identify clusters of strains within species. The most common typing method is multilocus sequence typing (MLST). The original MLST scheme for *C. jejuni* and *C. coli* targets seven housekeeping loci with each unique sequence variant being assigned an allele number

from the PubMLST database (Dingle et al., 2005; Jolley et al., 2018). The corresponding 7-loci allelic profile defines the sequence type (ST) of each isolate and STs are further clustered into related groups called clonal complexes (CC). MLST can also be assigned from whole-genome sequence information and applied gene-by-gene (Jolley and Maiden, 2010; Sheppard et al., 2012). With advances in technology and decreasing costs, WGS datasets are becoming increasingly available. The *Campylobacter* core-genome multilocus sequence typing (cgMLST) scheme identifies 1343 genes which are present within most (95%) members of the population[1] (Cody et al., 2017), and the whole-genome multilocus sequence typing (wgMLST) scheme includes all genes, even those which may be missing from a large proportion of other strains in the population (Maiden et al., 2013). Using data from the Source Assigned Campylobacteriosis in New Zealand (SACNZ) study (Lake et al., 2021), that contains 1211 isolates of *C. jejuni* and *C. coli* from three animal sources (cattle, chicken, and sheep) and from humans, This thesis show that cgMLST is also representative of *C. jejuni* and *C. coli* in New Zealand (chapter 3). 99.5% of the cgMLST genes are part of the New Zealand core genome and only 2.5% (20 genes) of the core genome of the New Zealand isolates consists of genes that are not defined by the cgMLST scheme.

Source attribution

Molecular typing methods can be used to attribute cases of human *Campylobacter* infection to possible animal sources. This is essential for understanding, preventing, and monitoring the spread of campylobacteriosis. Early molecular source attribution methods were developed to use with the relatively low-resolution 7-loci MLST scheme and are limited by the diversity of the loci in the model. As a result, these methods tend to show greater discriminatory power between major groups of sources (e.g., poultry *versus* ruminant) than within these major groups (e.g., cattle *versus* sheep within ruminants). The high resolution of WGS data allows fine-scale differentiation of closely related strains. As large whole genome sequenced reference datasets (sequenced from isolates of known origin across possible sources) become available, attribution using cgMLST or wgMLST data are likely to improve discrimination within source groups and overall attribution accuracy. At such high resolution however, the quantity of data increases by several orders of magnitude, and the counts of each unique genotype decrease with each additional gene until many genotypes are found in single isolates only. This also means that a large proportion of the human isolates will have a unique genotype (i.e., they are distinct from those found in the animal sources). For these reasons, existing source attribution models are generally poorly scalable to the whole genome level.

Random Forest

The growth of WGS data has seen a shift towards the development of machine learning (ML) methods for source attribution. For example, network analysis, random forest, and logit boost have been used for source attribution of *Listeria monocytogenes* (Mughini-Gras et al., 2025; Tanui et al., 2022), *Salmonella Typhimurium* (Guillier et al., 2020; Guzinski et al., 2024; Lupolova

[1] cgMLST has been validated against isolates from the United Kingdom, Europe, and North America

et al., 2019; Munck et al., 2020; Thystrup et al., 2024; Zhang et al., 2019), and *Campylobacter* species (Arning et al., 2021; Brinch et al., 2023; Harrison et al., 2021; Wainaina et al., 2022). These models 'learn' from the set of isolates in the animal sources and subsequently make inferences about the set of isolates from humans. Random forest is a method of supervised machine learning that creates an ensemble of decision trees and uses bagging (bootstrap aggregating) and random subspacing to prevent overfitting. Random forest is well suited to sets of data with a high number of variables and/or high numbers of variable levels, such as WGS data.

Absent levels

Due to the variable nature of alleles and the continual evolution of new genetic variants, genomic data is subject to the absent-levels problem (Au, 2018). Absent levels are levels that are present in new observations for prediction but were not present in the set of observations with which the model was trained. In this thesis, the extent of absent levels in data from the SACNZ study is evaluated (chapters 3, 4). This shows that almost all of the 1343 core genes contain absent levels (i.e., unique alleles), collectively affecting almost every isolate. Removing isolates and/or variables (i.e., genes) with absent levels is, therefore, not a viable option. There is currently no established solution for dealing with absent levels in random forest models (Au, 2018).[2]

Encoding categorical variables

The computational benefits of ordinal encoding categorical variables with large numbers of levels are well known. Ordinal encoding categorical variables also addresses the problem of absent levels by allowing absent levels to be integrated amongst the known levels and used for prediction. Coppersmith et al. (1999) first introduced the idea of ordering variables in decision trees and this method was adapted for random forest in the `ranger` package ('RANdom forest GEneRator') (Wright and Ziegler, 2017) for `R`, along with the option to order variables alphabetically. This method is a target-based approach which orders levels based on class probability information and designates absent levels the lowest rank. Using a case study on source attribution using data from the SACNZ study with cgMLST genes as predictors, the effect of absent levels on random forest predictive models when treated this way is evaluated and shows that predictions are systematically biased (chapter 4). In this thesis, three methods for ordinal encoding categorical predictor variables are developed that not only allow absent levels to be used by the random forest algorithm, but are unbiased and benefit from the information inherent in the level itself.

CA-unbiased-encoding

The first encoding method developed, in chapter 4, is an adaptation of the ordering method by Coppersmith et al. (1999) and implemented in `ranger`. The CA-unbiased-encoding method is a target-based encoding method which performs a scaled correspondence analysis on the contingency table of counts of variable levels by class. Each predictor variable is encoded according to the first principal component of the weighted matrix of class probabilities. Absent

[2] https://github.com/imbs-hl/ranger/issues/94

levels are encoded with a principal component score of zero. This assumes that any level of the predictor variable that is absent from the training data is *a priori* equally likely in any class and has equal class probabilities. This ensures all observations with an absent level branch as a group but not necessarily in the same direction across all nodes.

PCO-encoding

Two novel methods were subsequently developed in chapters 4 and 5 for ordinal encoding of categorical predictor variables which utilise a set of supplementary information on the category levels themselves (i.e., the sequence information of alleles) to inform the ordering. The first new method is the PCO-encoding method (chapter 4). This is a target-agnostic encoding method that encodes the levels (known and absent) according to their similarity or dissimilarity to each other via the method of principal coordinates analysis (PCO) (Gower, 1966). The PCO-encoding method performs an eigenanalysis on a matrix of Hamming distances calculated from the nucleotide sequence information of the alleles. The score is the principal component score for the corresponding predictor level for the principal coordinate(s). Using the method of Gower (1968), a new (absent) level can be scored by virtue of the interpoint distances between this level and each of the known levels. This then generates a score for each new level, and allows new levels to branch independently of each other, being informed by their resemblance to other levels in the training data.

CAP-encoding

The second novel method for encoding categorical predictor variables is the CAP-encoding method (chapter 5). The CAP-encoding method combines the advantages of the target-based CA-unbiased-encoding method and the target-agnostic PCO-encoding method. Based on the method of canonical analysis of principal coordinates (CAP) (Anderson and Willis, 2003), principal coordinates are first defined by the dissimilarity measure on the supplementary information, and then a classical correspondence analysis (CCorA) is performed which rotates a subset of the principal coordinates to correlate with the direction of greatest variation in class probabilities. The score is the principal component score for the corresponding predictor level for the rotated principal coordinate(s). As with the PCO method, the method of Gower (1968) is used to score new levels by virtue of its similarity to each of the known levels.

Method comparison

The goal of each of these methods is the same – to capture as much variation in the category levels as possible in the first eigenvector, with the new encoded value or score for each level being its position along each eigenvector. The difference in the methods lies in the multivariate space in which the variation among category levels is defined. The CA-unbiased-encoding method defines the variation in the class probability space; the PCO-encoding method defines the variation in the chosen dissimilarity space of the category levels themselves; and the CAP-encoding method defines the variation in the rotated PCO-space constrained by the variation in the class probability space.

The ability of random forest to recursively partition any number of dimensions means that the difference in predictive performance of these encoding methods is most apparent in the treatment of absent levels. In the class probability space, there is no information available to differentiate new category levels from each other, and so the the CA-unbiased-encoding method encodes all absent levels as zero. This is an unbiased method with good predictive properties when observations have a full complement of known levels; however, when there are a high number of absent levels, this method is unable to separate observations based on these levels, reducing prediction performance. The CA-unbiased-encoding method is expected to perform well when levels are found in different relative frequencies across the classes, and when there are few absent levels. In contrast, in the multivariate dissimilarity space of the category levels, any new level may be placed according to its position relative to every other level, and so the PCO-encoding and CAP-encoding methods will encode each absent level independently, according to its position in the dissimilarity space. This allows an observation with an absent level to be treated as though it had a similar observed level, where the similarity is defined according to the chosen characteristic. These methods make the assumption that an observation with an absent level is more likely to branch in the same direction as an observation whose corresponding level is 'similar' to the absent level, and requires an attribute with which to measure similarity (or dissimilarity). The PCO-encoding method works particularly well when the measure of similarity is capturing a characteristic of the variable in which the variation between levels is correlated with the variation in class probabilities. The CAP-encoding method is expected to perform well regardless of the direction of greatest variation among category levels. CAP ensures that the direction of greatest variation among the levels overlaps with the direction of greatest class separation. This gives the best of both worlds.

Using the same case study on source attribution of *Campylobacter* species using the SACNZ data with cgMLST genes as predictors, the implications of using different methods of encoding categorical variables are evaluated for the accuracy of random forest models (chapter 5). This shows that all three methods are unbiased solutions to the absent-levels problem and have good predictive properties. In addition, three real-world datasets were used to empirically demonstrate that encoding categorical predictors using the CAP-encoding method performs uniformly better than the CA-unbiased-encoding and the PCO-encoding methods, which themselves are both superior to the biased naïve approaches that are currently being employed (chapter 5). Therefore, the CAP-encoding method should be the preferred method of encoding categorical predictor variables for decision tree-based methods when supplementary information is available with which to define the similarity of category levels, and when supplementary information is not available the CA-unbiased-encoding method should be used.

Out-Of-Bag Data

An extension to this research focussed on potential issues when using out-of-bag (OOB) data following variable encoding (chapter 6). Performance of random forest models is often assessed and interpreted using OOB samples. The method of bagging in random forest means that not every observation is included in every tree. The observations that are excluded from each tree are

defined as 'out-of-bag'. Predictions from the OOB observations are frequently used to estimate error rates and measures of variable importance.

It is a common belief that the OOB samples alleviate the need to set aside a separate test set (Breiman, 1996, 2001). The aim of this additional work was to investigate how target-based *versus* target-agnostic encoding of categorical predictor variables for random forest can bias performance measures based on OOB samples. The CA-unbiased-encoding and the CAP-encoding methods utilise the target variable to inform the encoding and are, therefore, target-based methods, whereas the PCO-encoding method is target-agnostic because it is truly independent of the target variable. This thesis shows that, when categorical variables are encoded using a target-based encoding method, and when the encoding takes place prior to bagging, OOB samples suffer information leakage and are not a replacement for an independent test set. Under these circumstances, the OOB sample can underestimate the true misclassification rate, and overestimate variable importance. These findings are applicable to random forest and other tree-based methods (e.g., boosted trees) where OOB misclassification rates and/or variable importance measures are calculated.

In addition, a new variable importance measure is developed which is an adaptation of the holdout variable importance method of Janitza et al. (2018) (chapter 6). The original method computes the mean decrease in accuracy (MDA (Breiman, 2001)) using a second cross-validation fold; however current implementations[3] give biased calculations with target-based encoding methods because the folds are separated subsequent to encoding. The independent holdout method separates observations into two cross-validation folds prior to encoding and then computes the mean decrease in accuracy following random permutation of each variable in turn. This method is unbiased and is suitable for use with all methods of encoding. Using a separate test data set is recommended when evaluating predictive performance of tree based methods that utilise a target-based encoding method, and using the independent holdout method is recommended to estimate variable importance.

Optimal number of genes

Having developed and tested a robust method for encoding categorical variables, the next step was to use the CAP-encoding method with random forest to attribute human cases of *Campylobacter* to animal sources. This naturally led to the question of how many genes should be included in the predictive model and what is the best measure with which to compare alleles. Although it is accepted that accuracy of attribution will be improved by using WGS data, and several studies have shown better results with cgMLST data (over 7-gene MLST data) (Arning et al., 2021; Mughini-Gras et al., 2025; Wainaina et al., 2022), almost all studies have used either extensive preliminary feature reduction (Arning et al., 2021; Brinch et al., 2023; Munck et al., 2020; Pascoe et al., 2024; Tanui et al., 2022; Thystrup et al., 2024) or integer encoding of

[3]The holdout variable importance calculation is available in the R packages `ranger` and `randomForestSRC`

variables[4] (Arning et al., 2021; Tanui et al., 2022).[5] Having found a solution to the absent-levels problem, it is now possible to directly compare results of models using a different number of genes as input variables. This chapter examines the effect of using low-resolution 7-gene MLST data *versus* cgMLST data *versus* wgMLST data in a random forest model for source attribution of *Campylobacter* (chapter 7). A marked improvement in accuracy is shown with the addition of core genes, however, the addition of accessory genes does not further improve predictive performance although it does not deteriorate it either. Furthermore, host associated genetic variation in *Campylobacter* species was observed in both core and accessory genes and there was a high level of correlation between genes.

Distance measures

The premise behind molecular source attribution is that isolates from the same source will be more related than isolates from different sources. The distance measure used for encoding should ideally be able to detect this genomic signature and define alleles from the same source as being more similar to each other than alleles from a different source. Initially, a matrix of Hamming distances was calculated between each pair of alleles within each gene using the nucleotide sequencing information of each allele, on the assumption that alleles that share many nucleotides have similar host preferences. However, *C. jejuni* and *C. coli* are highly diverse species and show evidence of both clonal relatedness through substitutions and panmixis through frequent recombination events. The Hamming distance measures the number of nucleotides that differ between sequences - a large Hamming distance could, therefore, reflect a large number of single point mutations, or a single large scale evolutionary event, such as recombination. Two methods were therefore used to adjust the Hamming distance (chapter 7). Firstly, the Hamming distance was adjusted for each gene using a dissimilarity-based multivariate analogue to a residual plot, so that the variation due to clonal structure was removed. Secondly, the Hamming distance was adjusted for each gene by down-weighting each nucleotide position according to its degree of recombination. There was very little effect on misclassification rates of adjusting the Hamming distance for either the variation inherent in CC or for the variation due to recombination. This suggests that the ability of random forest to recursively partition variables allows for the smaller differences resulting from host signature to be discerned, amongst the stronger patterns of clonal structure and recombination. It was therefore concluded that the Hamming distance is a suitable measure for comparing alleles for the purpose of source attribution.

Attribution of *C. jejuni* and *C. coli* in New Zealand

Finally, this thesis applied the CAP-encoding method to a random forest predictive model for source attribution of whole genome sequencing data from the SACNZ study (chapter 7). wgMLST level data and the associated set of Hamming distances between alleles of each gene

[4]Integer encoding is likely to introduce bias in MLST-based data due to the process of allele assignment in contguous chunks by source.

[5]Mughini-Gras et al. (2025) used one-hot encoding with random forest analysis of *Listeria monocytogenes* using cgMLST data.

were used to attribute cases of *C. jejuni* and *C. coli* infections in New Zealand to putative livestock sources. Using these approaches the study estimates that approximately (49.9% – 64.7%) of *C. jejuni* and *C. coli* infection was linked to chickens, (17.5% – 30.7%) was linked to cattle, and (13.1% – 25.1%) was linked to sheep. These results largely agree with those of Lake et al. (2021) which were based on a reduced set of genes from the same set of isolates and used the asymmetric island model, though the study by Lake et al. (2021) attributes an even greater proportion of cases to poultry and attributes very few cases to sheep. This chapter, based on wgMLST data, places a higher emphasis on cattle and sheep than previously described. Poultry meat has long been identified as a significant pathway for campylobacteriosis across developed countries (Cody et al., 2019), including New Zealand (ESR, 2023), where interventions in broiler meat production have successfully reduced incidence of infection (Muellner et al., 2011; Sears et al., 2011). This chapter highlights that cattle and sheep, either directly as foods or via their production environments, also contribute to human campylobacteriosis in New Zealand. The consistently high incidence of campylobacteriosis in New Zealand and the range of possible sources motivates the need for ongoing surveillance studies and refinement of source attribution methods to continue to monitor infection, quantify the role of host sources, and inform policies for disease control.

8.2 Wider Applications of the Models

This research focuses on the development of methods of encoding categorical predictor variables in order to allow for WGS data to be used with ML models for the source attribution of *Campylobacter*. This work also has the potential to be used for the attribution of other disease causing organisms, such as *Salmonella typhimurium* and *Listeria monocytogenes*, which have genes with alleles as categorical predictor variables. It is not necessary for pre-defined loci (e.g., PubMLST) to be used, as any scheme for classifying gene content can be applied to raw read data, assembled contigs, or draft genomes, and allele numbers assigned to unique sequences for each gene. In addition, the encoding methods are well suited to other types of data consisting of large numbers of variable levels, such as free text response fields of survey data. As the encoding is applied predictor by predictor, the methods are able to be used in conjunction with other relevant covariates, either numeric or categorical.

8.3 Future Work

The work presented in this thesis can be extended in several directions. The effect of class imbalance on predictive performance of encoding methods should be further explored, including assessing whether target-agnostic approaches offer any advantage over target-based approaches. As a general rule, ML methods that are trained on imbalanced data show poor predictive performance for the minority classes. This may negatively impact attribution studies especially when sparsely sampled host sources, such as wild birds, are included in the model. In addition, the effect of gene correlation on measures of variable importance has not been explored. Potentially,

groups of correlated genes, such as clusters of orthologous genes (COGs), may collectively be more important than each of the genes individually and removal of single genes to assess the effect on predictive accuracy may not be the best assessment of their importance. These methods may also inflate the importance of genes which are not particularly important, but which are correlated with few (or no) other genes. Performing the analysis at the amino acid level (i.e., only considering non-synonymous substitutions) and, as better bioinformatic tools become available, including translation and post-translation features such as epigenetic DNA methylation and acetylation, structural variations, available tRNA pools, variations in promoter regions, homopolymeric tracts causing slip-strand mispairing, and pseudogene formation, may also be areas worth exploring.

In terms of implementing the methods, there are several potential areas for refinement. Firstly, although the encoded variables may be used in conjunction with other covariates, this is currently a manual process. Development of a method that allows for independent variables to be encoded using variable specific methods of encoding and/or distance measures would be a valuable extension. In addition, there is potential to add tuning steps into the algorithm. This could include adjusting the number of genes considered at each split, and optimising the number of principal components/coordinates to retain for each variable. Finally, the importance of selecting a suitable characteristic and distance metric by which to define the categories for the distance based methods should be highlighted. The success of the distance-based encoding methods is directly related to the ability of the distance measure to capture differences between observations from different classes. Despite the vast array of available distance metrics, in some cases the patterns within classes may not be well captured and alternative metrics could be sought or developed. One approach may be to consider the effect of clonal structure and recombination jointly, by, for example, residualising the distance matrices following down-weighting for recombination.

8.4 Concluding Remarks

Source attribution of *C. jejuni* and *C. coli* using WGS data is constrained by the intensive computational demands of vast numbers of genes and alleles within genes, as well as the presence of unique genotypes (absent levels) in the human isolates for prediction. This thesis describes the extent of absent levels in *Campylobacter* data in New Zealand and quantifies the impact they have on tree-based predictive models, which has not previously been addressed in the published literature. This work primarily explores methods of ordinal encoding of categorical predictor variables. It introduces three methods of encoding that are computationally efficient, and allow for unique genotypes in prediction, thus allowing for wgMLST genes to be used, directly and without feature reduction, in tree-based methods for source attribution, such as random forest. Combatting the issue of absent levels has allowed the unique opportunity to compare the predictive performance of random forest models using 7-gene MLST *versus* cgMLST *versus* wgMLST data and found that cgMLST data is a good level of resolution for attribution of *Campylobacter* species. This thesis goes on to disprove the common belief that out-of-bag data is as reliable as a

separate test set and highlights the importance of using an independent set of data for assessing error rates and variable importance. It also demonstrates that the Hamming distance is a suitable measure with which to compare alleles for the purpose of source attribution, although it does not discount potential advantages of novel distance measures and raises this as an area for future research. This thesis concludes by performing attribution of human cases of campylobacteriosis from the SACNZ study using a random forest predictive model with wgMLST level data as input following encoding with the CAP-encoding method informed by the matrix of Hamming distances between pairs of alleles for each gene.

The model estimates that chicken are the primary source of infection with *C. jejuni* and *C. coli* in New Zealand and that cattle and sheep have a larger role than previously estimated using 7-gene MLST level data and the asymmetric island model. These results emphasise the need to explore additional public health measures focusing on cattle and sheep. Methods of source attribution are undoubtedly improved by the addition of WGS data and methods of ordinal encoding. For random forest source attribution models it is recommended to use cgMLST level data, the CAP-encoding method, and sets of Hamming distances for each gene. The data in this thesis was balanced, and the methods have not been explored with unbalanced data; this is another area for future research. Previous obstacles presented by high-resolution WGS data have been successfully overcome in this thesis, enabling more accurate source attribution estimates, which will provide insight into potential risk factors in the spread of foodborne diseases and better inform public health decision making.

Appendix A :

Supplementary Files for Chapter 4 - Lost in the Forest

A.1 The effect of response class order on classification accuracy

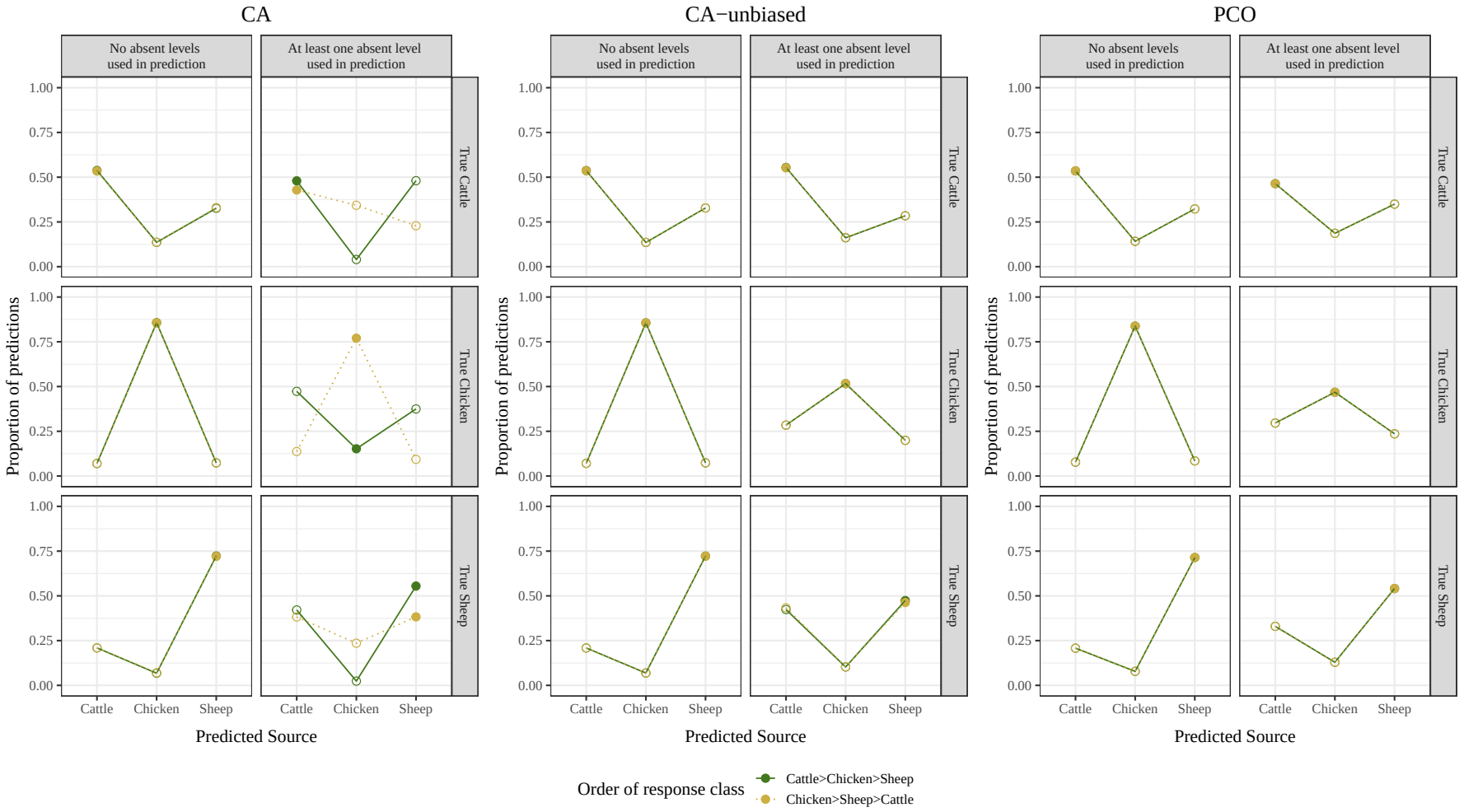

Figure A.1: The effect of response class order on classification accuracy. Open circles represent the proportion of cases for which the true class is predicted incorrectly; closed circles represent the proportion of cases for which the true class is predicted correctly.

A.2 Bias resulting from treatment of absent levels

To examine how the bias resulting from treatment of absent levels is affected by the proportion of absent levels, a set of data was simulated and analysed with random forest following encoding via the CA-encoding and the CA-unbiased-encoding methods. The simulated dataset consisted of n observations coming from one of two classes (A and B), each with a single predictor variable that took one of three levels (a, b, or c). Levels a and b were perfectly associated with class A and B respectively, while level c was associated with class A with probability p_A which varied from $p_A \in 0.2, 0.5, 0.8$.

The training set consisted of the predictor with levels a and b leading to perfect separation; the test set consisted of the absent level c with probability $p_c \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, with levels a and b assigned with probability $(1 - p_c)/2$.

In this simple scenario, the missclassification rate can be assessed exactly. Let m be the probability of the observation being sent to the right branch of the split, where

$$m = \begin{cases} 1, & \text{for CA-encoding} \\ 0.5, & \text{for CA-unbiased-encoding,} \end{cases}$$

then the expected misclassification rate of class A is

$$\frac{\frac{1-p_c}{2} + mp_Ap_c}{\frac{1-p_c}{2} + p_Ap_c};$$

the expected misclassification rate of class B is

$$\frac{\frac{1-p_c}{2} + (1-m)p_c(1-p_A)}{\frac{1-p_c}{2} + p_c(1-p_A)},$$

and the expected total misclassification rate is

$$1 - p_c + mp_Ap_c + (1-m)(1-p_A)p_c.$$

Here, the simulations matched the expected probabilities (figure A.2). When absent levels are scored as infinity, as per the CA-encoding method, the random forest model is biased towards the first response class (Class A) and this bias gets worse with increasing proportion of absent levels and increasing association of absent level with class B. In addition, the overall misclassification rate for the CA-encoding method is affected by the level of association of the absent level with the response class. When absent levels are scored as zero, as per the CA-unbiased-encoding method, the random forest model favours the response class with the greatest association with the absent level but it is not affecting the overall misclassification rate which is independent of the level of association of the absent level with either response class.

Different methods of encoding nominal variables have important implications for the accuracy of error rates when absent levels are present in the data. The CA-encoding method is

biased when classifying observations with absent levels. When there are no, or few, absent levels both methods have similar predictive performance, however the CA-encoding method never out-performs the CA-unbiased-encoding method.

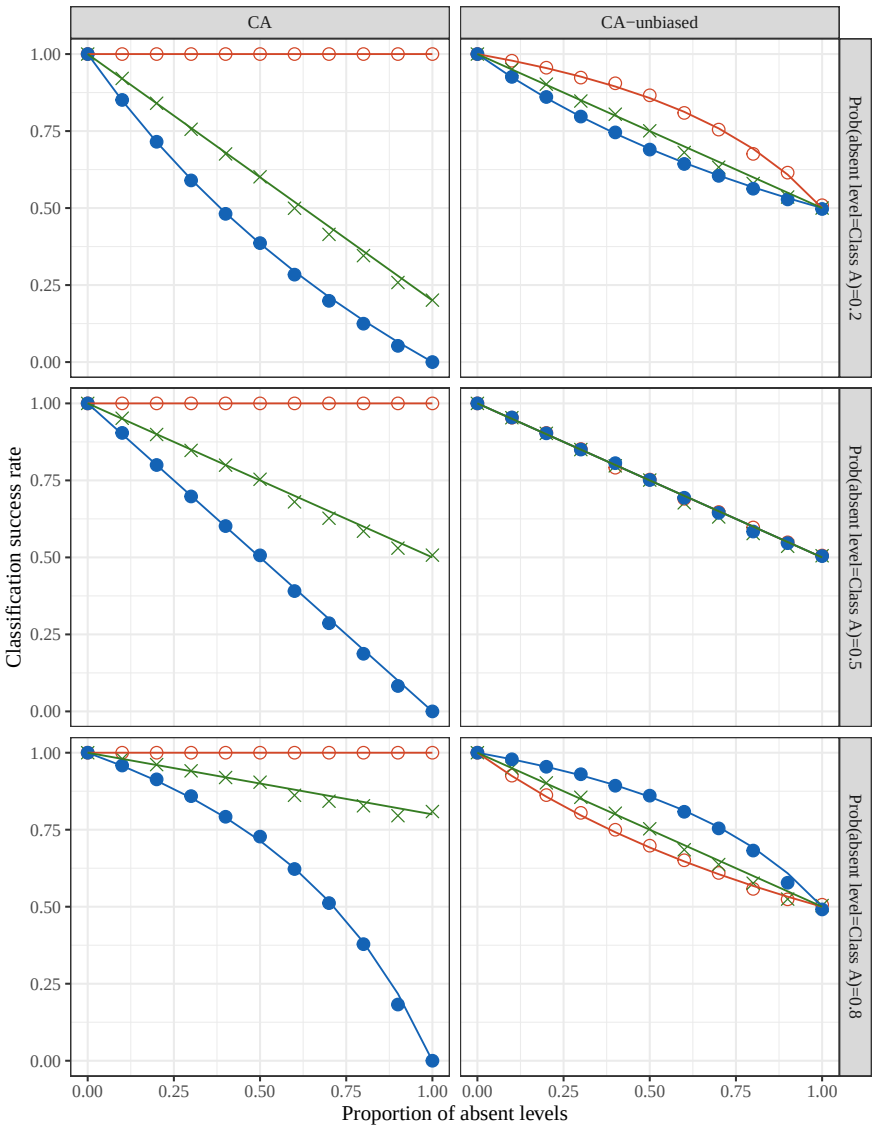

Figure A.2: The effect of absent levels on classification accuracy for the CA-encoding and CA-unbiased-encoding methods. Open circles represent the proportion of cases for which Class A is predicted correctly; closed circles represent the proportion of cases for which Class B is predicted correctly; crosses represent the proportion of total cases which were predicted correctly; the lines represent the expected probabilities.

Appendix B :

Supplementary Files for Chapter 5 - To CAP it Off

B.1 The CAP-encoding methodology

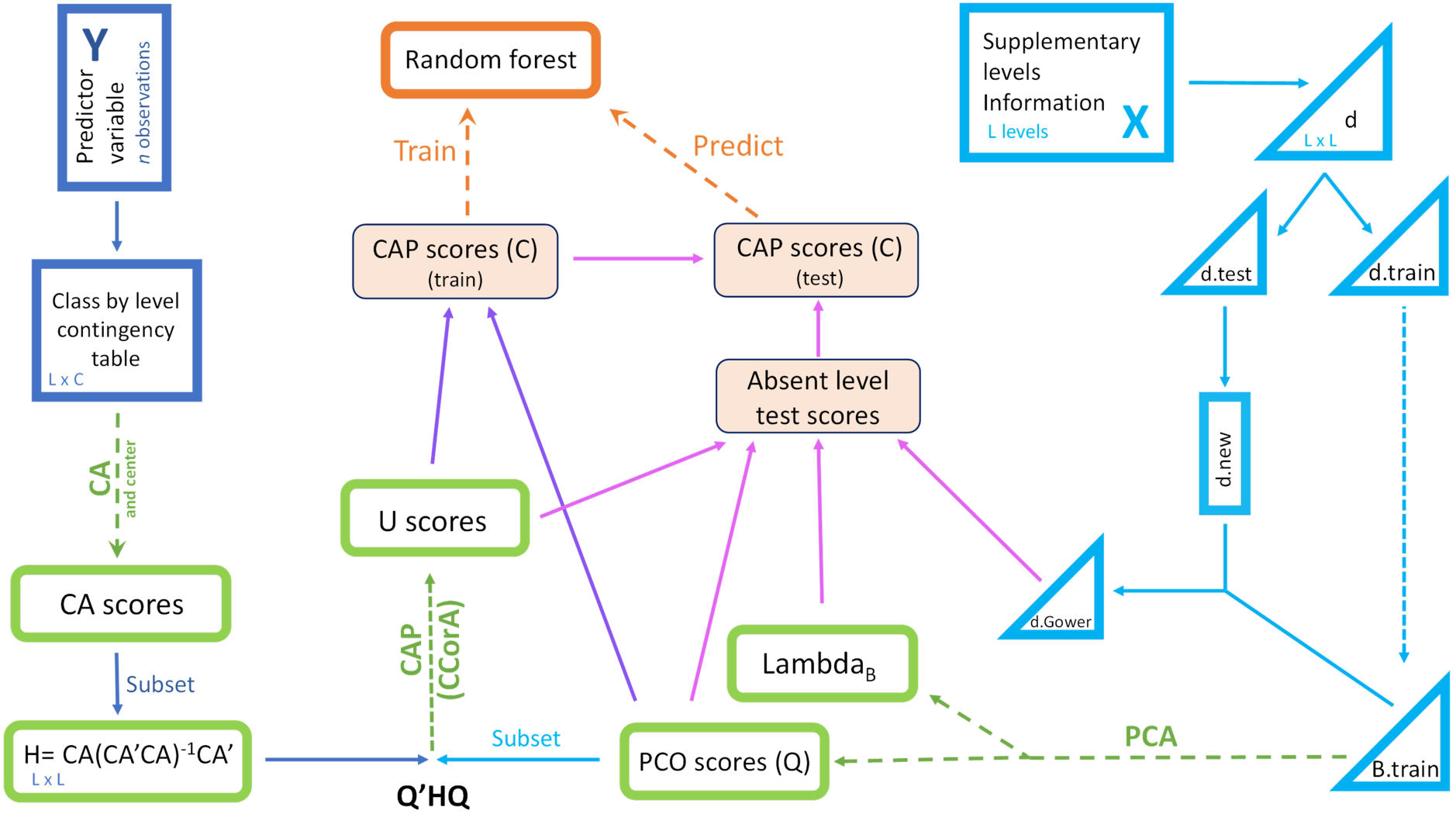

Figure B.1: Schematic illustration of the CAP-encoding methodology.

Let:

- $Y = [y_i]$ be a vector of n observations, where each observation y_i comes from the set of L observed category levels, and belongs to a single class C;
- **CA** be an $L \times p$ matrix that contains the row scores from a correspondence analysis of an L by C contingency table of the observations in Y which has been centered on its column means; p is the minimum of $C - 1$; the number of positive eigenvalues; or a pre-specified number of dimensions;
- **H** be the L by L orthonormalised "hat" matrix $\mathbf{H} = CA[CA'CA]^{-1}CA'$;
- **d** be an L by L symmetric matrix of distances among the observed category levels $l \in \{l_1, ..., l_L\}$, calculated using any distance metric of choice;
- **Q** be the set of orthonormal principal coordinates obtained from the dissimilarity matrix **d**, and subsetted according to some criterion, such as the number of axes resulting in minimum misclassification error, a defined percentage of variation of the levels in PCO space, or the number of non-zero eigenvalues.
- **U** be the set of canonical eigenvectors obtained from a classical canonical correlation analysis (CCorA) performed on $\mathbf{Q'HQ}$;
- **C** be the set of CAP (training) scores for the category levels calculated as the product of **U** and **Q**.

B.2 Simulation study comparing encoding methods when the direction of greatest variation in the category levels is along the first principal coordinate axis

A set of data was simulated and analysed with random forest as per section 2.1. The simulated data consisted of a balanced, 15 levels by two class frequency table, representing a single predictor variable. The position of each of the 15 predictor levels in a 2-dimensional PCO space was simulated such that the direction of greatest variation among the levels was along the first dimension, PCO1. Each level was then assigned to ten individuals, representing two classes according to the probability that the level will belong to the first class. In contrast with section 2.1, the probability of assignment is calculated from the inverse logit of the position of each level along the first (not second) dimension, PCO1, scaled by β. β now represents the magnitude of discrimination between the classes along PCO1, not PCO2. All other steps are identical to those described in section 2.1.

When the direction of class separation is the same as the direction of greatest variation in the predictor levels, the PCO-encoding method is able to separate the levels into class probability groups from a single dimension and rotation of the axes is not beneficial. For the trees which did not use absent levels for prediction, the average misclassification rates of the three encoding methods was the same and was lower for β=20 (0.6%) than for β=2 (8.0%). For the trees which used at least one absent level for prediction, the average misclassification rate of the CA-unbiased-encoding method increased in line with random assignment to each class for both values of β (51.6% when β=2, and 49.0% when β=20). However, the average misclassification rates of the PCO-encoding and CAP-encoding methods for the trees which used at least one absent level for prediction remained similar to the misclassification rates for the trees which did not use absent levels for prediction (10.1% to 11.4% when β=2, and 7.2% to 8.6% when β=20) (figure B.2).

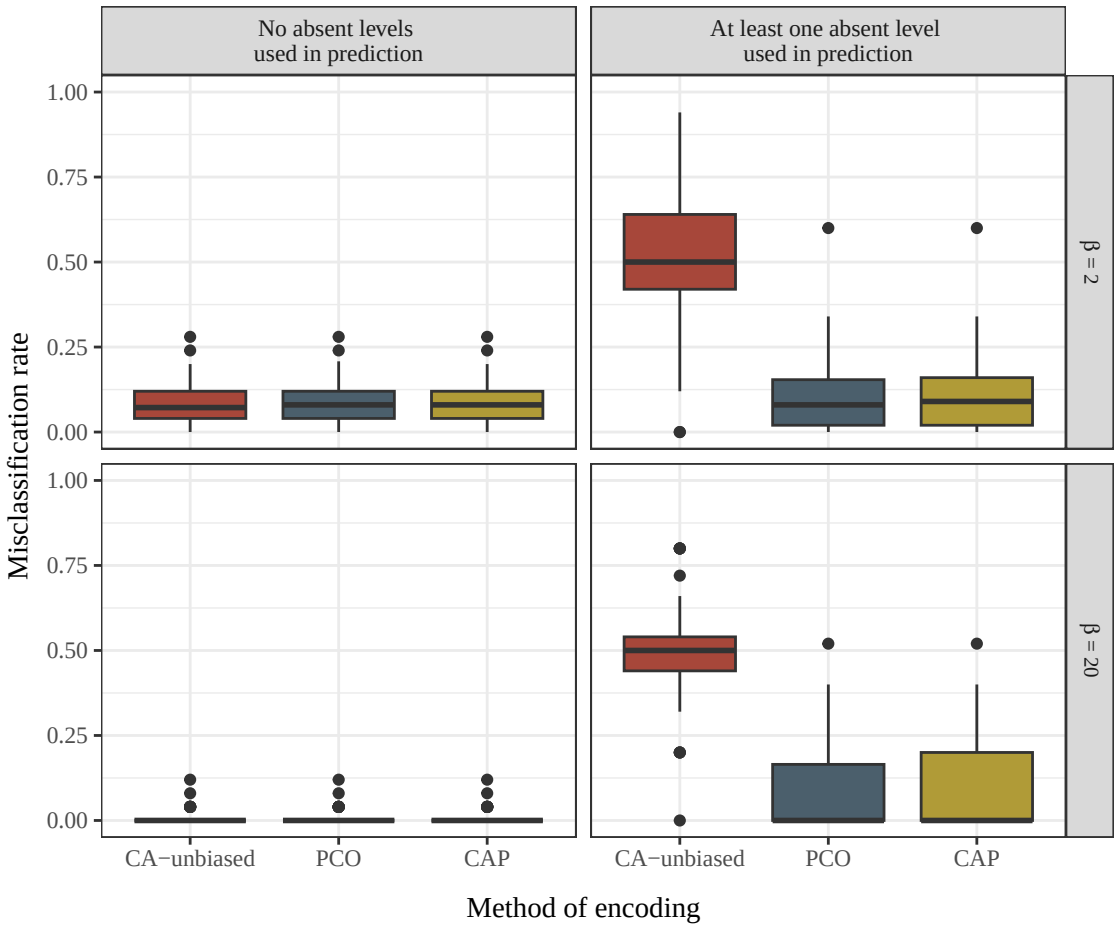

Figure B.2: Misclassification rates of 1000 classification trees, from data simulated for ten individuals each with a single variable comprising 15 levels and assigned to 2 classes with probability proportional to β. Ten levels are observed in the training data and five are unique to the testing data. β represents the magnitude of discrimination between the classes along PCO1.

B.3 Midwest survey "open response" scores for the PCO-encoding method

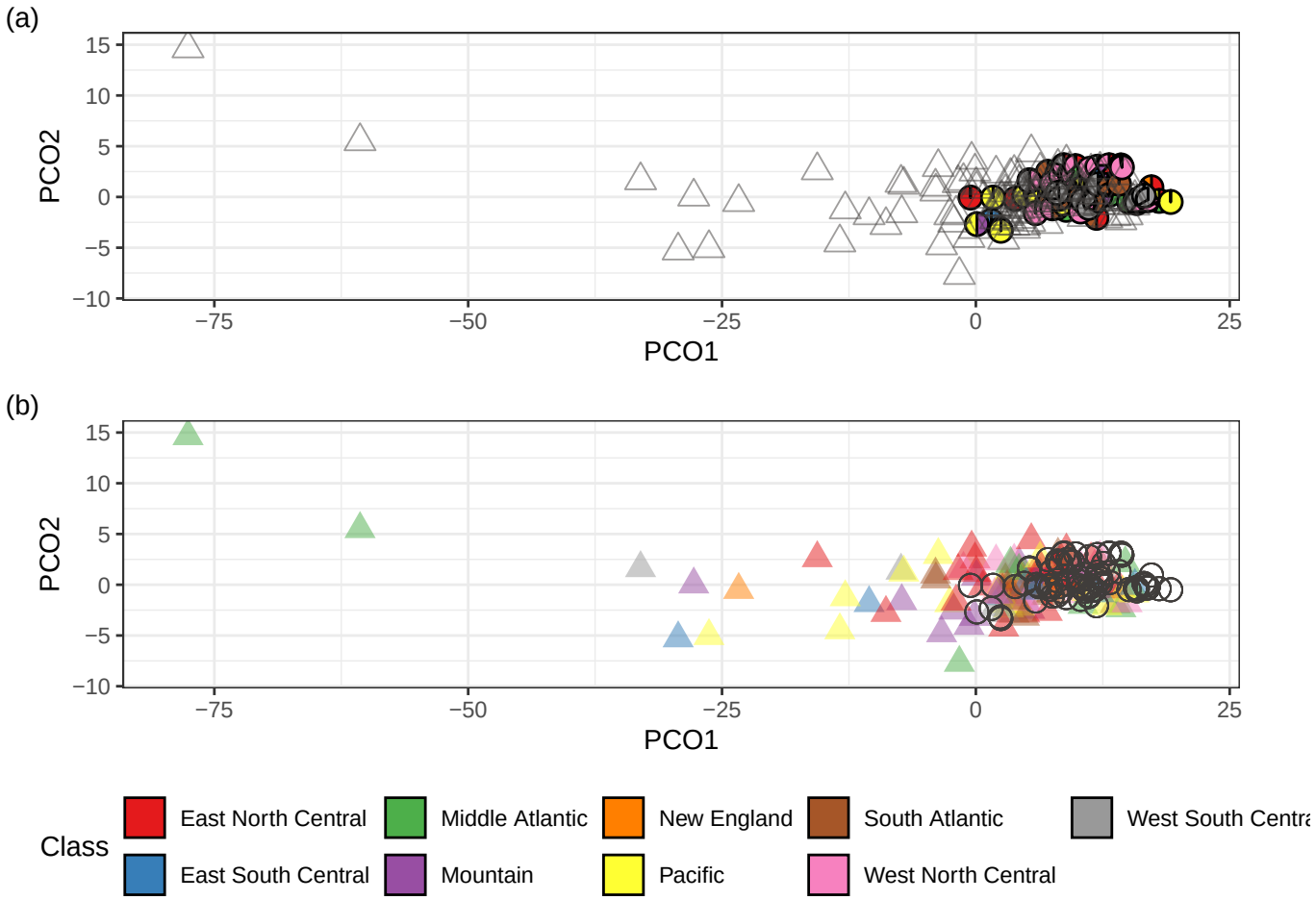

Figure B.3: The first two dimensions of encoded scores (jittered) for the categorical variable 'open response' for the PCO-encoding method. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The colour represents the proportion of observations with each level in each class for known levels (a), and absent levels (b).

B.4 The effect of different amounts of variation retained in the PCO subset on misclassification rates

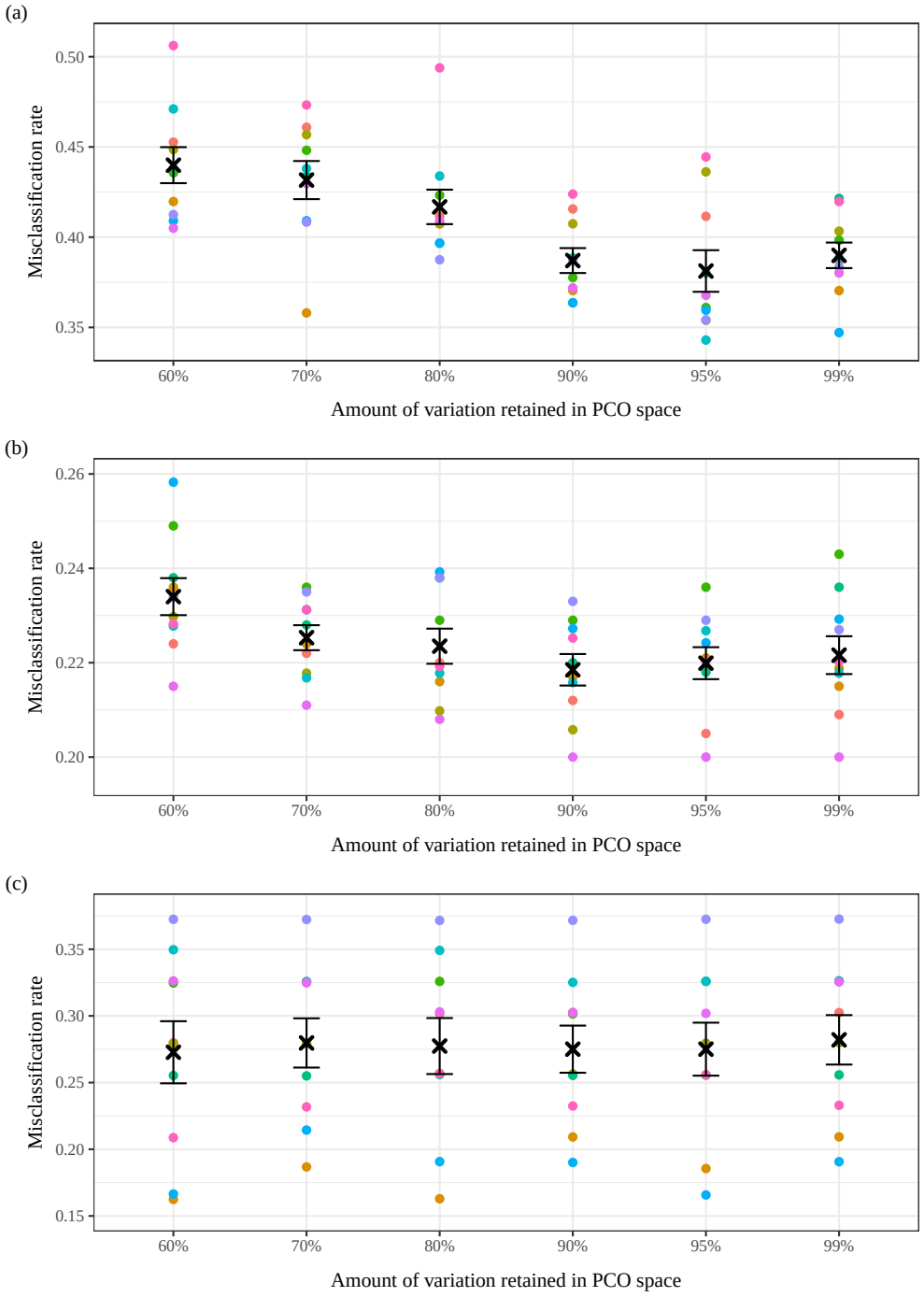

Figure B.4: Misclassification rates for the (a) Midwest survey, (b) Traffic violation, and (c) SACNZ datasets for the CAP-encoding method for differing amounts of variation retained in the PCO space. Each coloured point represents the misclassification rate of a fold, and the weighted average and standard error are depicted with a black cross and error bars.

B.5 The effect of increasing number of dimensions on random forest predictive performance

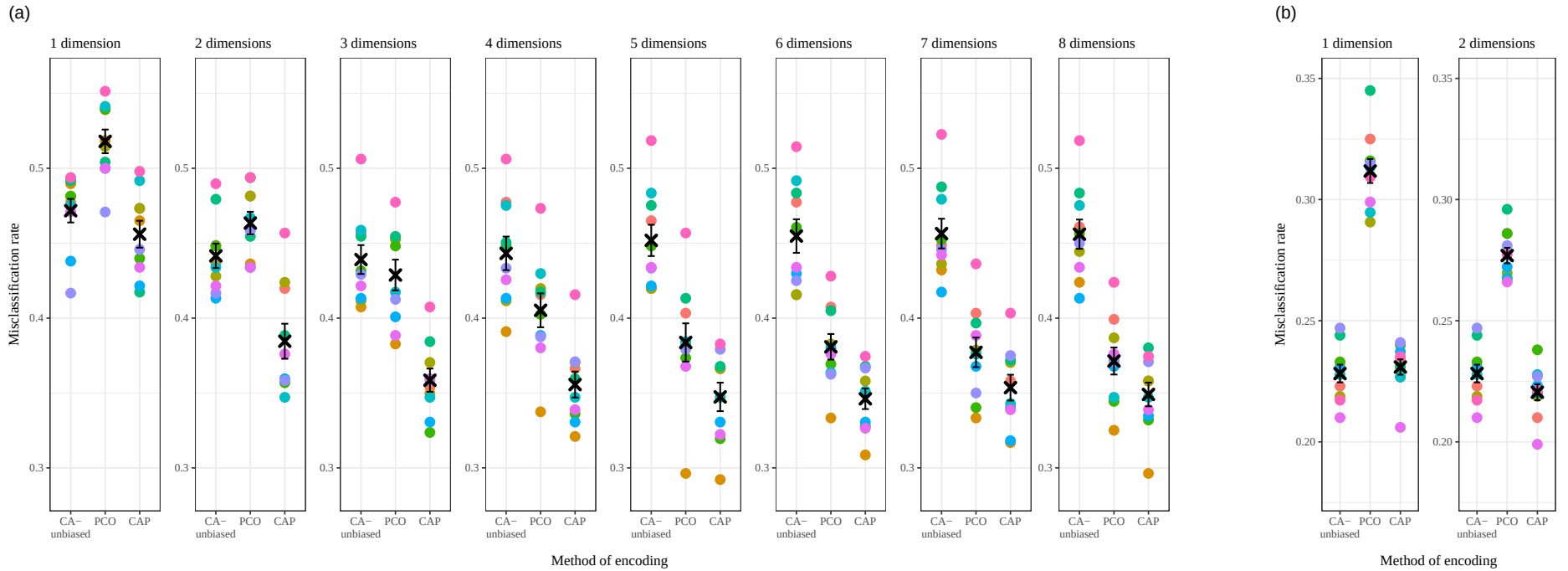

Figure B.5: Misclassification rates for the (a) Midwest survey dataset, and (b) traffic violation dataset for the three methods of encoding when different numbers of dimensions are used in the random forest analysis. Each coloured point represents the misclassification rate of a fold, and the weighted average and standard error are depicted with a black cross and error bars.

B.6 The effect of missing values on the first principal component of the variable ‘CAMP1225’ from the SACNZ dataset

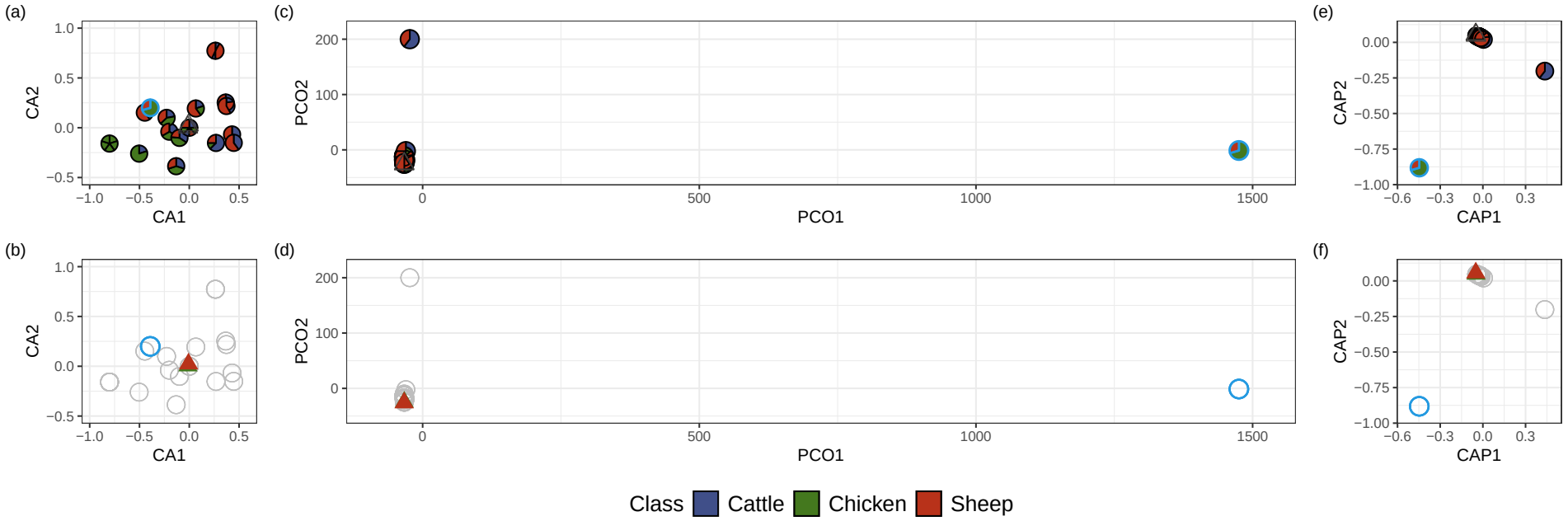

Figure B.6: The first two dimensions of encoded scores (jittered) for the categorical variable ‘CAMP1225’ for the CA-unbiased-encoding (a, b), PCO-encoding (c, d), and CAP-encoding (e, f) methods. Each level present in the training data is represented with a circle, and absent levels are represented with a triangle. The fill colour represents the proportion of observations with each level in each class (known levels, circles) or class membership (absent levels, triangles). The top row (a, c, e) shows the class membership of known levels, and the bottom row (b, d, f) shows the class membership of absent levels. The blue border indicates values that were missing in observations in the set of training data.

Appendix C :

Supplementary Files for Chapter 6 - Out of (the) Bag

C.1 The effect of increasing number of variables on the out-of-bag misclassification rate

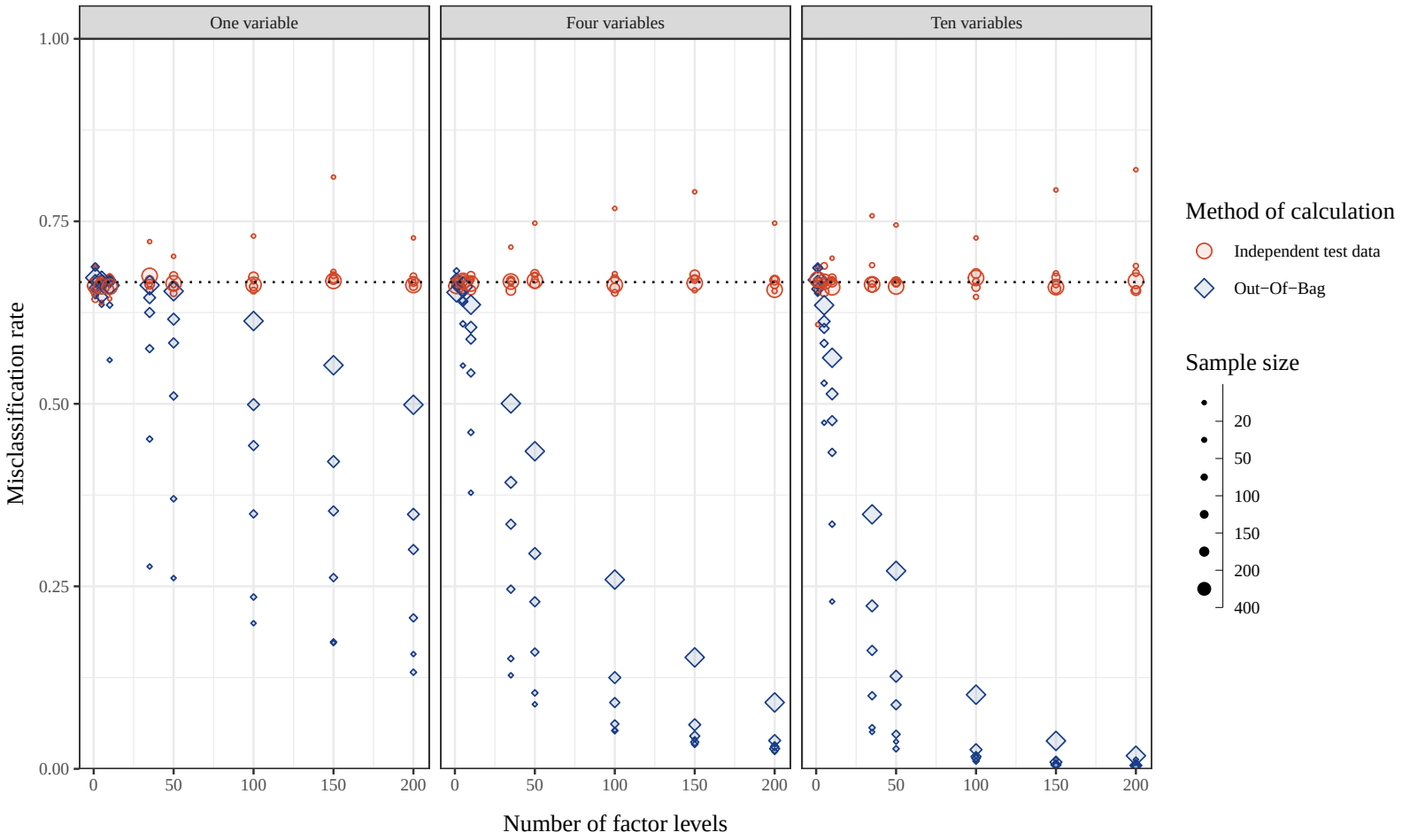

Figure C.1: The effect of increasing number of variables on the out-of-bag misclassification rate. Circles represent misclassification rates calculated using independent test data and diamonds represent misclassification rates calculated using out-of-bag samples.

C.2 The effect of increasing number of variables on measures of variable importance

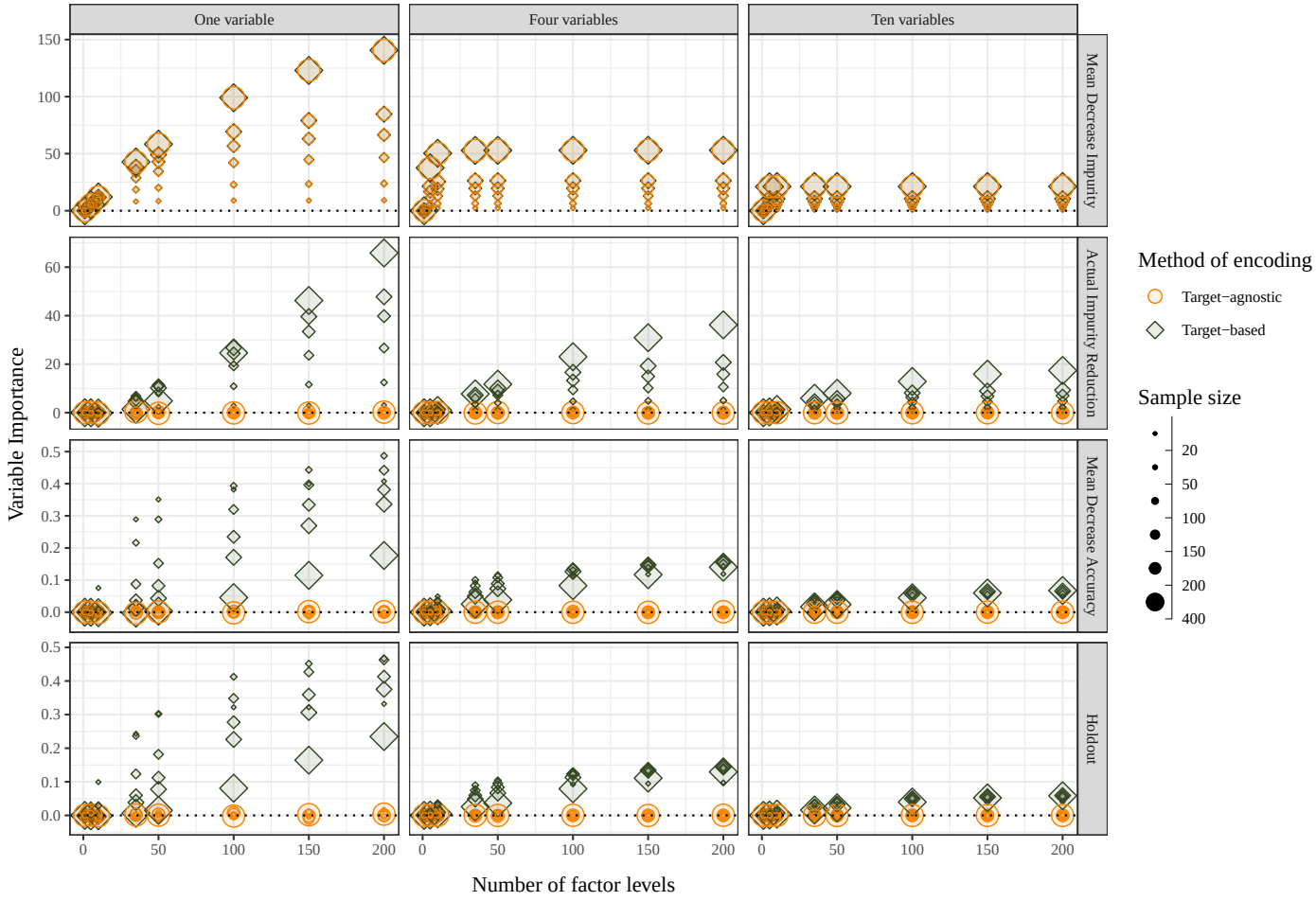

Figure C.2: The effect of method of encoding and increasing number of variables on measures of variable importance. Circles represent variable importance calculated when variables were encoded using a target-agnostic method and diamonds represent variable importance calculated when variables were encoded using a target-based method.

Appendix D :

Supplementary Files for Chapter 7 - Source Attribution of *Campylobacter*

D.1 The effect of nominal encoding of categorical variables

The impact of using nominal encoding (i.e., rather than ordinal encoding) of categorical variables was investigated for random forest analysis of the SACNZ data. Each of the 1343 core genes were nominally encoded such that each value equated to the corresponding allele number, and new alleles were encoded with unique values greater than the maximal allele number. The random forest was trained using 70% of the isolates and the remaining 30% of the isolates were assigned to one of three sources (cattle, chicken, or sheep). The average misclassification rate was calculated for each source individually, and overall. The allele labels were then shuffled and the random forest was re-trained and new misclassification rates were calculated. This was repeated 20 times and the average misclassification rates were calculated.

Following random shuffling of the labels, the average misclassification rates increased for all sources (figure D.1), but especially for sheep and for cattle which increased by 4.3% and 3.4% respectively. This highlights the potential bias of interpreting allele numbers as quantitative rather than as categorical (non-informative) labels.

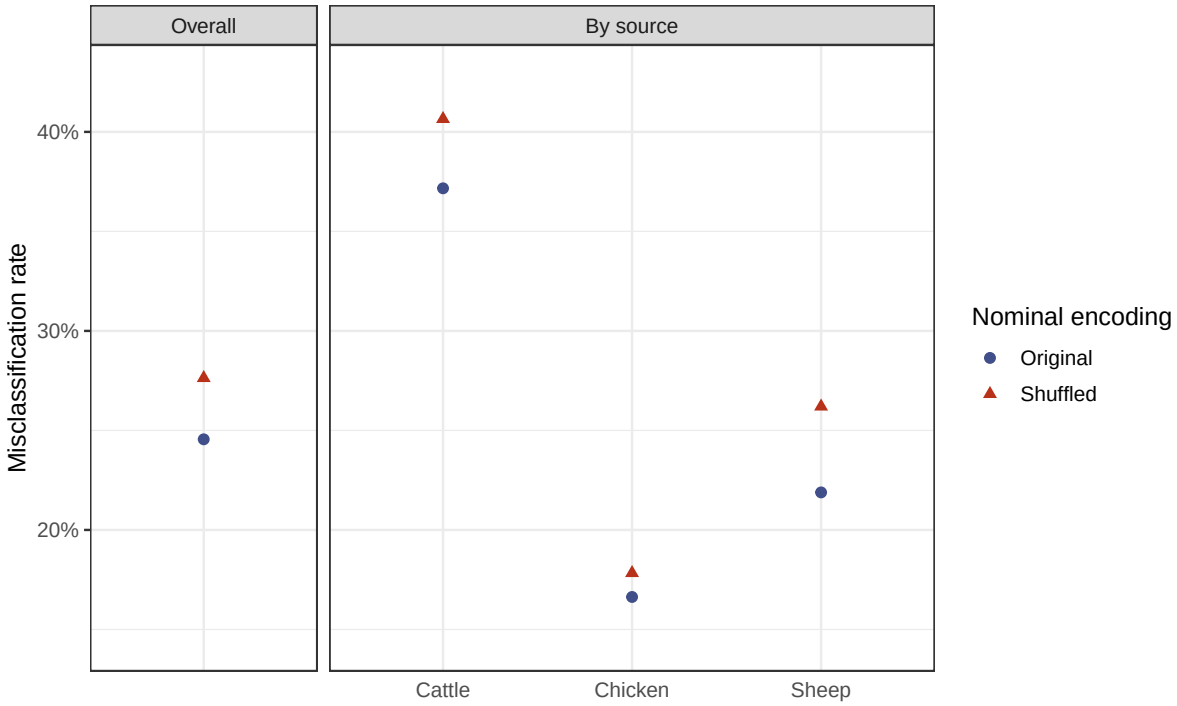

Figure D.1: The effect of nominal encoding of categorical variables. Circles represent misclassification rates calculated using original numeric allele labels and diamonds represent misclassification rates calculated following random shuffling of allele labels.

D.2 Another ten most important variables according to the independent holdout variable importance measure

Table D.2: Another ten most important variables according to the independent holdout variable importance measure.

Locus	**Variable Importance**	**Function**
CAMP1075 (rho)	0.003304	Transcription termination factor (Jolley et al., 2018)
CAMP1225*	0.003304	Putative amino acid activating enzyme (Jolley et al., 2018)
CAMP0545	0.003303	Putative integral membrane protein (Flint et al., 2014; Novik et al., 2010)
CAMP1251 (maf3)	0.003303	Motility accessory factor (Jolley et al., 2018)
CAMP0751*	0.003303	No known function (Cody et al., 2017; Kovanen et al., 2014)
CAMP0332 (cmeA)	0.003302	Periplasmic fusion protein CmeA (multidrug efflux system CmeABC) (Jolley et al., 2018)
CAMP0035	0.002479	No known function (Jolley et al., 2018)
CAMP0461	0.002479	Putative histidine triad (HIT) family protein (Jolley et al., 2018)
CAMP0890	0.002479	Putative periplasmic protein (Jolley et al., 2018)
CAMP0973 (npdA)	0.002479	NAD-dependent deacetylase (Jolley et al., 2018)

* accessory gene (i.e., not part of the cgMLST scheme).

Appendix E :

Statement of Contributions

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Helen Smith
Name/title of Primary Supervisor:	Associate Professor Jonathan Marshall
In which chapter is the manuscript /published work:	Four (4)

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

H.L. Smith, P.J. Biggs, N.P. French, A.N.H. Smith, and J.C. Marshall. Lost in the Forest: Encoding Categorical Variables and the Absent Levels Problem (2024). Data Mining and Knowledge Discovery, 38, 1889 - 1908.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate: 80

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate conceived of the modelling with supports from supervisors, carried out the statistical analyses and drafted the manuscript with inputs from supervisors.

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Helen Smith 2025.02.15 12:27:10 +13'00'
Date:	15 Feb 2025
Primary Supervisor's Signature:	
Date:	14 Feb 2025

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Helen Smith
Name/title of Primary Supervisor:	Associate Professor Jonathan Marshall
In which chapter is the manuscript /published work:	Five (5)

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N.H., and Marshall, J.C., CAP-encoding: Encoding Categorical Variables using Canonical Analysis of Principal Coordinates.

◉ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

Machine Learning

- The percentage of the manuscript/published work that was contributed by the candidate: 80

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate conceived of the modelling with supports from supervisors, carried out the statistical analyses, prepared the manuscript, and created the figures.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	H. Smith Helen Smith 2025.02.15 12:26:14 +13'00'
Date:	15 Feb 2025
Primary Supervisor's Signature:	J Marshall
Date:	14 Feb 2025

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Helen Smith
Name/title of Primary Supervisor:	Associate Professor Jonathan Marshall
In which chapter is the manuscript /published work:	Six (6)

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Smith, H.L., Biggs, P.J., French, N.P., Smith, A.N.H., and Marshall, J.C., Out of (the) bag - encoding categorical predictors impacts out-of-bag samples. PeerJ Computer Science (2024), 10:e2445, DOI 10.7717/peerj-cs.2445.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate: 80

- Describe the contribution that the candidate has made to the manuscript/published work:

The candidate conceived of the modelling with supports from supervisors, carried out the statistical analyses, prepared the manuscript, and created the figures.

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	H. Smith Helen Smith 2025.02.15 12:27:38 +13'00'
Date:	15 Feb 2025
Primary Supervisor's Signature:	J Marshall
Date:	14 Feb 2025

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

References

S. Adelabu, O. Mutanga, and E. Adam. Testing the reliability and stability of the internal accuracy assessment of random forest for classifying tree defoliation levels using different validation methods. *Geocarto International*, 30(7):810–821 – 821, 2015.

B. M. Allos. *Campylobacter jejuni* infections: Update on emerging issues and trends. *Clinical Infectious Diseases*, 32(8):1201 – 1206, 2001. ISSN 10584838.

S. F. Altschul, W. Gish, D. J. Lipman, W. Miller, and E. W. Myers. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46 – 46, 2001. ISSN 14429985.

M. J. Anderson. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62(1):245 – 253, 2006. ISSN 0006341X.

M. J. Anderson. Permutational multivariate analysis of variance (PERMANOVA). In *Wiley StatsRef: Statistics Reference Online*, pages 1–15. John Wiley and Sons, Ltd, 2017. ISBN 9781118445112.

M. J. Anderson and J. Robinson. Generalized discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics*, 45(3):301–318, 2003.

M. J. Anderson and T. J. Willis. Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology*, 84(2):511 – 525, 2003. ISSN 00129658.

A. Anjum. *Mechanistic and functional analysis of Cj0031: a phase variable methyltransferase in Campylobacter jejuni*. PhD thesis, University of Leicester, 2013.

N. Arning, S. K. Sheppard, S. Bayliss, D. A. Clifton, and D. J. Wilson. Machine learning to predict the source of campylobacteriosis using whole genome data. *PLoS Genetics*, 17(10), 2021.

T. C. Au. Random forests, decision trees, and categorical predictors: The "absent levels" problem. *Journal of Machine Learning Research*, 19:1–30, 2018.

R. Auti, A. Bhatt, and S. Tidake. Comparative analysis of machine learning algorithms for genomic data. *2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), 2023 1st DMIHER International Conference on*, 1:1 – 6, 2023. ISSN 979-8-3503-3842-3.

F. O. Bagger, L. Borgwardt, A. S. Jespersen, A. R. Hansen, B. Bertelsen, M. Kodama, and F. C. Nielsen. Whole genome sequencing in clinical practice. *BMC Medical Genomics*, 17(1), 2024. ISSN 1755-8794.

A. Bateman, M. J. Martin, S. Orchard, M. Magrane, A. Adesina, S. Ahmad, E. H. Bowler-Barnett, H. Bye-A-Jee, D. Carpentier, P. Denny, J. Fan, P. Garmiri, L. J. da Costa Gonzales, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasaamy, A. Lock, A. Luciani, J. Luo, Y. Lussi, J. S. M. Marin, P. Raposo, D. L. Rice, R. Santos, E. Speretta, J. Stephenson, P. Totoo, N. Tyagi, N. Urakova, P. Vasudev, K. Warner, S. Wijerathne, C. W. H. Yu, R. Zaru, A. J. Bridge, L. Aimo, G. Argoud-Puy, A. H. Auchincloss, K. B. Axelsen, P. Bansal, D. Baratin, T. M. Batista Neto, M. C. Blatter, J. T. Bolleman, E. Boutet, L. Breuza, B. C. Gil, C. Casals-Casas, K. C. Echioukh, E. Coudert, B. Cuche, E. de Castro, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M. Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C. J. A. Sigrist, K. Sonesson, S. Sundaram, A. Sveshnikova, C. H. Wu, C. N. Arighi, C. Chen, Y. Chen, H. Huang, K. Laiho, M. Lehvaslaiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Y. Wang, and J. Zhang. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. ISSN 13624962; 03051048. URL `https://research.ebsco.com/linkprocessor/plink?id=412c8534-c05f-3afb-8c7f-70c14f10ae2c`.

C. Benard, D. V. C. Sebastien, and R. Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881 – 900, 2022.

J. A. Benshak, N. Strachan, B. Lopes, M. Ramjee, M. Macrae, and K. Forbes. Identifying the sources of human campylobacteriosis in Nigeria. *Acta Tropica*, 237(106702), 2023. ISSN 0001-706X.

E. Berthenet, A. Ducournau, A. Buissonnière, L. Bénéjat, E. Bessède, F. Mégraud, P. Lehours, A. Thépault, M. Chemaly, K. Rivoal, and S. K. Sheppard. Source attribution of *Campylobacter jejuni* shows variable importance of chicken and ruminants reservoirs in non-invasive and invasive French clinical isolates. *Scientific Reports*, 9(1), 2019. ISSN 20452322.

X. Bian, J. M. Garber, C. M. Szymanski, M. K. Mills, D. Rafala, K. K. Cooper, S. Huynh, C. T.

Parker, W. G. Miller, J. Jones, D. Nasrin, K. L. Kotloff, and S. M. Tennant. *Campylobacter* abundance in breastfed infants and identification of a new species in the global enterics multicenter study. *mSphere*, 5(1), 2020. ISSN 23795042.

A. Bigdeli, A. Maghsoudi, and R. Ghezelbash. Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran. *Journal of Geochemical Exploration*, 233, 2022. ISSN 03756742.

M. J. Blaser. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *The Journal of Infectious Diseases*, 176:S103 – S105, 1997. ISSN 00221899.

A. L. Boulesteix, S. Janitza, L. Kruppa, and I. R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1(58):3–42, 2002.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth International Group, 1984.

M. L. Brinch, T. Hald, C. Henri, L. Wainaina, A. Merlotti, D. Remondini, and P. M. K. Njage. Comparison of source attribution methodologies for human campylobacteriosis. *Pathogens*, 12(6), 2023.

T. Bylander. Estimating generalization error on two-class datasets using out-of-bag estimates. *Machine Learning*, 48(1-3):287–297, 2002.

M. L. Calle, V. Urrea, A. L. Boulesteix, and N. Malats. Auc-rf: a new strategy for genomic profiling with random forest. *Human heredity*, 72(2):121–132, 2011.

P. Cerda, G. Varoquaux, and B. Kegl. Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107:1477—-1494, 2018.

D. Charif and J. R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.

J. Chen, R. Tan, and Y. Yang. Research on an innovative feature importance recognition algorithm based on GINI-OOB index. *2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA), Image Processing and Computer Applications*

(ICIPCA), 2023 IEEE International Conference on, pages 862 – 866, 2023. ISSN 979-8-3503-1467-0.

X. Chen and H. Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323 – 329, 2012. ISSN 08887543.

Y. Chen, S. Mukherjee, M. Hoffmann, S. Young, J. Abbott, M. K. Davidson, P. McDermott, S. Zhao, M. L. Kotewicz, Y. Luo, and M. Allard. Whole-genome sequencing of gentamicin-resistant *Campylobacter coli* isolated from U.S. retail meats reveals novel plasmid-mediated aminoglycoside resistance genes. *Antimicrobial Agents and Chemotherapy*, 57(11):5398–5405 – 5405, 2013. ISSN 00664804.

A. J. Cody, N. M. McCarthy, H. L. Wimalarathna, F. M. Colles, L. Clark, I. C. J. W. Bowler, M. C. J. Maiden, and K. E. Dingle. A longitudinal 6-year study of the molecular epidemiology of clinical *Campylobacter* isolates in Oxfordshire, United Kingdom. *Journal of Clinical Microbiology*, 50(10):3193 – 3201, 2012. ISSN 00951137.

A. J. Cody, N. D. McCarthy, M. Jansen van Rensburg, T. Isinkaye, S. D. Bentley, J. Parkhill, K. E. Dingle, I. C. J. W. Bowler, K. A. Jolley, and M. C. J. Maiden. Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. *Journal of Clinical Microbiology*, 51(8):2526 – 2534, 2013. ISSN 00951137.

A. J. Cody, J. E. Bray, K. A. Jolley, N. D. McCarthy, and M. C. J. Maidena. Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. *Journal of Clinical Microbiology*, 55(7):2086–2097 – 2097, 2017. ISSN 1098660X.

A. J. Cody, M. C. J. Maiden, N. D. McCarthy, and N. J. C. Strachan. A systematic review of source attribution of human campylobacteriosis using multilocus sequence typing. *Eurosurveillance*, 24(43), 2019. ISSN 15607917.

C. Collins and X. Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology*, 14(2):1 – 21, 2018. ISSN 1553734X.

D. Coppersmith, S. E. J. Hong, and J. R. M. Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, 1999.

P. Cressey and R. Lake. Estimated incidence of foodborne illness in New Zealand: Application of overseas models and multipliers. Technical Report 2012/11, Ministry for Primary Industries, June 2012.

N. J. Croucher, A. J. Page, T. R. Connor, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, and A. J. Delaney. Rapid phylogenetic analysis of large samples of recombinant bacterial

whole genome sequences using Gubbins. *Nucleic Acids Research*, 43(3):e15, 2015. ISSN 13624962.

L. Dai, O. Sahin, Y. Tang, and Q. Zhang. A mutator phenotype promoting the emergence of spontaneous oxidative stress-resistant mutants in *Campylobacter jejuni*. *Applied and Environmental Microbiology*, 83(24):e01685–17, 2017. ISSN 00992240; 10985336.

K. A. Dauda. Optimal tuning of random survival forest hyperparameter with an application to liver disease. *Malaysian Journal of Medical Sciences*, 29(6):67 – 76, 2022.

M. Denis, V. Rose, B. Nagard, A. Thépault, P. Lucas, M. Meunier, F. Benoit, E. Cauvin, A. Wilhem, B. Gassilloud, A. Rincé, and M. Gourmelon. Comparative analysis of *Campylobacter jejuni* and *C. coli* isolated from livestock animals to *C. jejuni* and *C. coli* isolated from surface water using DNA sequencing and MALDI-TOF. *Pathogens*, 12(9), 2023. ISSN 20760817.

M. Deviaene, D. Testelmans, P. Borzée, B. Buyse, S. V. Huffel, and C. Varon. Feature selection algorithm based on random forest applied to sleep apnea detection. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2580–2583, 2019. doi: 10.1109/EMBC.2019.8856582.

A. I. Dfuf, J. F. Perez-Minayo, J. M. M. McWilliams, and C. G. Fernandez. Variable importance analysis in imbalanced datasets: A new approach. *IEEE Access, Access, IEEE*, 8:127404 – 127430, 2020. ISSN 2169-3536.

K. E. Dingle, F. M. Colles, D. R. A. Wareing, R. Ure, A. J. Fox, F. E. Bolton, H. J. Bootsma, R. J. L. Willems, R. Urwin, and M. C. J. Maiden. Multilocus sequence typing system for *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 39(1):14–23 – 23, 2001. ISSN 00951137.

K. E. Dingle, F. M. Colles, R. Ure, J. A. Wagenaar, B. Duim, F. J. Bolton, A. J. Fox, D. R. A. Wareing, and M. C. J. Maiden. Molecular characterization of *Campylobacter jejuni* clones: A basis for epidemiologic investigation. *Emerging Infectious Diseases*, 8(9):949 – 955, 2002. ISSN 1080-6040.

K. E. Dingle, F. M. Colles, D. Falush, and M. C. J. Maiden. Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *Journal of Clinical Microbiology*, 43(1):340–347 – 347, 2005. ISSN 00951137.

A. R. Domingues, S. M. Pires, T. Halasa, and T. Hald. Source attribution of human campylobacteriosis using a meta-analysis of case-control studies of sporadic infections. *Epidemiology and Infection*, 140(6):970 – 981, 2012. ISSN 0950-2688.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:29p., 2006. ISSN 14712105.

I. Epifanio. Intervention in prediction measure: a new approach to assessing variable importance for random forests. *BMC Bioinformatics*, 18(1):1 – 16, 2017. ISSN 1471-2105.

S. V. R. Epps, R. B. Harvey, M. E. Hume, T. D. Phillips, R. C. Anderson, and D. J. Nisbet. Foodborne *Campylobacter*: Infections, metabolism, pathogenesis and reservoirs. *International Journal of Environmental Research and Public Health*, 10(12):6292 – 6304, 2013. ISSN 16604601.

ESR. Annual report concerning foodborne diseases in New Zealand 2023. New Zealand food safety technical paper no: 2024/13. Technical report, The Institute of Environmental Science and Research Limited, Porirua, New Zealand, December 2023. URL https://www.mpi.govt.nz/dmsdocument/65046-Annual-report-concerning-Foodborne-Diseases-in-New-Zealand-2023(2024/11/14).

C. R. Falcao, M. R. Edwards, M. Hurst, E. Fraser, and M. Otterstatter. A review on microbiological source attribution methods of human salmonellosis: From subtyping to whole-genome sequencing. *Foodborne pathogens and disease*, 21(3):137 – 146, 2024. ISSN 1556-7125.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164(4):1567 – 1587, 2003. ISSN 00166731.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, 7(4):574–578, 2007. ISSN 14718278.

S. Fan, C. Parsons, S. Kathariou, D. Foster, S. Zhao, S. Mukherjee, and Y. Shrestha. Genomic analysis reveals that isolation temperature on selective media introduces genetic variation in *Campylobacter jejuni* from bovine feces. *Pathogens*, 11(6):678, 2022. ISSN 20760817.

P. Fearnhead, P. J. Biggs, and N. French. Learning about recombination in *Campylobacter*. In Sheppard S. K., editor, *Campylobacter Ecology and Evolution.* Caister Academic Press, Swansea, UK, 2014. ISBN 9781908230362.

A. M. Fernandes, C. Balasegaram, S.and Willis, H. M. L. Wimalarathna, M. C. Maiden, and N. D. McCarthy. Partial failure of milk pasteurization as a risk for the transmission of *Campylobacter* from cattle to humans. *Clinical Infectious Diseases*, 61(6):903 – 909, 2015. ISSN 10584838.

I. Ferrés and G. Iraola. An object-oriented framework for evolutionary pangenome analysis. *Cell Reports Methods*, 1(5), 2021. ISSN 26672375.

W. D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53(284):789–798, 1958.

A. Flint, Y. Q. Sun, J. Butcher, M. Stahl, H. Huang, and A. Stintzi. Phenotypic screening of a targeted mutant library reveals *Campylobacter jejuni* defenses against oxidative stress. *Infection and immunity*, 82(6):2266–2275, 2014.

N. French and J. Marshall. Source attribution studies for campylobacteriosis in New Zealand. Technical report, Ministry for Primary Industries, Wellington, New Zealand, March 2014. URL `http://www.mpi.govt.nz/news-resources/publications.aspx(2020/07/12)`.

B. J. Gilpin, T. Walker, S. Paine, J. Sherwood, G. Mackereth, T. Wood, T. Hambling, C. Hewison, A. Brounts, M. Wilson, P. Scholes, B. Robson, S. Lin, A. Cornelius, L. Rivas, D. T. S. Hayman, N. P. French, J. Zhang, D. A. Wilkinson, A. C. Midwinter, P. J. Biggs, M. G. Baker, R. Eyre, N. Jones, and A. Jagroop. A large scale waterborne campylobacteriosis outbreak, Havelock North, New Zealand. *Journal of Infection*, 81(3):390–395 – 395, 2020. ISSN 15322742.

C. N. Gomes, A. G. Felice, G. do N. Pereira, V. A. S. Ceballos, S. de C. Soares, L. Tonani, P. H. G. Barião, M. R. von Z. Kress, S. da S. Duque, M. Balkey, M. W. Allard, and J. P. Falcão. Comparative genomics and virulence potential of *Campylobacter coli* strains isolated from different sources over 25 years in Brazil. *BMC Microbiology*, 24(1):512, 2024. ISSN 14712180.

J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4):325–338, 1966.

J. C. Gower. Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55(3): 582–585, 1968.

B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678, 2017.

L. Guillier, S. Cadel-Six, M. L. Vignaud, F. Palma, M. Gourmelon, S. Lozach, N. Munck, and T. Hald. Ab_sa: Accessory genes-based source attribution – tracing the source of Salmonella enterica typhimurium environmental strains. *Microbial Genomics*, 6(7):1–10 – 10, 2020. ISSN 20575858.

J. Guzinski, Y. Tang, L. Petrovska, M. A. Chattaway, and T. J. Dallman. Development and validation of a random forest algorithm for source attribution of animal and human *Salmonella Typhimurium* and monophasic variants of *S. Typhimurium* isolates in England and Wales utilising whole genome sequencing data. *Frontiers in Microbiology*, 14, 2024. ISSN 1664-302X.

J. L. Guzmán-Martín, P. González-Bustos, and J. Gutiérrez-Fernández. *Campylobacter* spp. and typing tools (review). *Applied Biochemistry and Microbiology*, 55(5):470 – 473, 2019. ISSN 00036838.

J. Hadfield, N. J. Croucher, R. J. Goater, K. Abudahab, D. M. Aanensen, and S. R. Harris. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 34(2): 292–293, 09 2017. ISSN 1367-4803.

T. Hald, D. Vose, H. C. Wegener, and T. Koupeev. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Analysis: An International Journal*, 24(1):255 – 269, 2004. ISSN 02724332.

T. Hald, W. Aspinall, B. Devleesschauwer, R. Cooke, T. Corrigan, A. H. Havelaar, H. J. Gibb, P. R. Torgerson, M. D. Kirk, F. J. Angulo, R. J. Lake, N. Speybroeck, and S. Hoffmann. World Health Organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: A structured expert elicitation. *PLoS ONE*, 11(1):1 – 35, 2016. ISSN 19326203.

E. Harrison, A. J. Wood, C. Dytham, J. W. Pitchford, J. Truman, A. Spiers, S. Paterson, and M. A. Brockhurst. Bacteriophages limit the existence conditions for conjugative plasmids. *mBio*, 6(3), 2015. ISSN 2150-7511.

L. Harrison, S. Mukherjee, C.-H. Hsu, S. Young, E. Strain, S. Zhao, Q. Zhang, G. E. Tillman, C. Morales, and J. Haro. Core genome MLST for source attribution of *Campylobacter coli*. *Frontiers in Microbiology*, 12, 2021. ISSN 1664302X.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848570.

A. H. Havelaar, A. V. Galindo, D. Kurowicka, and R. M. Cooke. Attribution of foodborne pathogens using structured expert elicitation. *Foodborne pathogens and disease*, 5(5):649–659, 2008.

W. Hickey. We've published our data on the South and Midwest, 2014. URL `https://fivethirtyeight.com/features/weve-published-our-data-on-the-south-and-midwest`. Accessed: 2024-07-01.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

T. Hothorn and A. Zeileis. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16(118):3905–3909, 2015.

M. J. Hubisz, D. Falush, M. Stephens, and J. K. Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, 9(5):1322–1332 – 1332, 2009. ISSN 1755098X.

H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

H. Ishwaran and U. B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2023. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 3.2.0.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

M. Jaillard, L. Lima, M. Tournoud, P. Mahé, A. van Belkum, V. Lacroix, and L. Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genetics*, 14(11):1 – 28, 2018. ISSN 15537390.

S. Janitza and R. Hornung. On the overestimation of random forest's out-of-bag error. *PLoS ONE*, 13(8):e0201904, 2018.

S. Janitza, E. Celik, and A. L. Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 12: 885–915, 2018.

Q. Jehanne, L. Bénéjat, A. Ducournau, A. Buissonnière, F. Mégraud, E. Bessède, P. Lehours, B. Pascoe, E. Mourkas, and S. K. Sheppard. Genome-wide identification of host-segregating single-nucleotide polymorphisms for source attribution of clinical *Campylobacter coli* isolates. *Applied and Environmental Microbiology*, 86(24):1–14, 2020. ISSN 10985336; 00992240.

K. A. Jolley and M. C. J. Maiden. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC bioinformatics*, 11:595, 2010. ISSN 1471-2105.

K. A. Jolley, J. E. Bray, and M. C. J. Maiden. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research*, 3, 2018. ISSN 2398502X.

A. B. Karki, B. Khatri, and M. K. Fakhr. Transcriptome analysis of *Campylobacter jejuni* and *Campylobacter coli* during cold stress. *Pathogens*, 7(14):960, 2023.

K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology & Evolution*, 30(4):772–780, 2013.

K. Katoh, K. Misawa, K. i. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

M. Kokot, M. Dlugosz, and S. Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761 – 2761, 2017. ISSN 13674811.

J. Kovac, B. Stessl, N. Čadež, I. Gruntar, M. Cimerman, K. Stingl, M. Lušicky, M. Ocepek, M. Wagner, and S. Smole Možina. Population structure and attribution of human clinical *Campylobacter jejuni* isolates from central Europe to livestock and environmental sources. *Zoonoses and Public Health*, 65(1):51 – 58, 2018. ISSN 18631959.

S. M. Kovanen, R. I. Kivistö, M. Rossi, M. L. Hänninen, T. Schott, U. M. Kärkkäinen, T. Tuuminen, J. Uksila, and H. Rautelin. Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. *Journal of Clinical Microbiology*, 52(12):4147–4154, 2014. ISSN 1098660X; 00951137.

J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1 – 27, 1964. ISSN 0033-3123.

M. Kuhn and H. Wickham. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*, 2020. URL `https://www.tidymodels.org`.

S. Kuwabara. Guillain-barre syndrome. *Current Neurology and Neuroscience Reports*, 7(1):57 – 62, 2007. ISSN 15284042.

R. Lake, E. Ashmore, J. Benschop, D. Campbell, P. J. Cressey, N. French, S. C. Hathaway, B. Horn, J. Marshall, A. Midwinter, S. Pirikahu, J. Sherwood, and D. Wilkinson. Source assigned campylobacteriosis in New Zealand study (SACNZS), 2020.

R. J. Lake, D. M. Campbell, S. C. Hathaway, E. Ashmore, P. J. Cressey, B. J. Horn, S. Pirikahu, J. M. Sherwood, M. G. Baker, P. Shoemack, J. Benschop, J. C. Marshall, A. C. Midwinter, D. A. Wilkinson, and N. P. French. Source attributed case-control study of campylobacteriosis in New Zealand. *International Journal of Infectious Diseases*, 103:268–277, 2021.

R. L. Lawrence, S. D. Wood, and R. L. Sheley. Mapping invasive plants using hyperspectral imagery and breiman cutler classifications (randomforest). *Remote Sensing of Environment*, 100(3):356–362, 2006.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

S. Z. Li and A. Jain, editors. *Hamming Distance*, pages 668–668. Springer US, Boston, MA, 2009. ISBN 978-0-387-73003-5.

S. J. Liao, J. Marshall, M. L. Hazelton, and N. P. French. Extending statistical models for source attribution of zoonotic diseases: A study of campylobacteriosis. *Journal of the Royal Society Interface*, 16(150), 2019. ISSN 17425662.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Y. W. Liu, K. Denkmann, K. Kosciow, C. Dahl, and D. J. Kelly. Tetrathionate stimulated growth of campylobacter jejuni identifies a new type of bi-functional tetrathionate reductase (tsda) that is widely distributed in bacteria. *Molecular microbiology*, 88(1):173–188, 2013. doi: https://doi.org/10.1111/mmi.12176.

M. Loecher. Unbiased variable importance for random forests. *Communications in Statistics: Theory & Methods*, 51(5):1413 – 1425, 2022. ISSN 03610926.

LPSN. Genus *Campylobacter*, 2024. URL `https://lpsn.dsmz.de/genus/campylobacter`.

N. Lupolova, S. J. Lycett, and D. L. Gally. A guide to machine learning for bacterial host attribution using genome sequence data. *Microbial Genomics*, 5(12), 2019. ISSN 20575858.

M. C. J. Maiden, R. Urwin, J. Zhou, B. G. Spratt, J. A. Bygraves, J. E. Russell, I. M. Feavers, E. Feil, Q. Zhang, G. Morelli, K. Zurth, M. Achtman, and D. A. Caugant. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6):3140–3145 – 3145, 1998. ISSN 00278424.

M. C. J. Maiden, M. J. J. van Rensburg, J. E. Bray, S. G. Earle, S. A. Ford, K. A. Jolley, and N. D. McCarthy. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews Microbiology*, 11(10):728 – 736, 2013. ISSN 17401526.

G. Manning, C. G. Dowson, M. C. Bagnall, I. H. Ahmed, M. West, and D. G. Newell. Multilocus sequence typing for comparison of veterinary and human isolates of *Campylobacter jejuni*. *Applied and environmental microbiology*, 69(11):6370 – 6379, 2003. ISSN 0099-2240.

N. D. McCarthy, F. M. Colles, K. E. Dingle, M. C. J. Maiden, D. Falush, M. C. Bagnall, and G. Manning. Host-associated genetic import in *Campylobacter jejuni*. *Emerging Infectious Diseases*, 13(2):267–272 – 272, 2007. ISSN 10806059.

N. D. McDonald and E. F. Boyd. Structural and biosynthetic diversity of Nonulosonic Acids (NulOs) that decorate surface structures in bacteria. *Trends in Microbiology*, 29(2):142–157, 2021. ISSN 0966-842X.

R. J. Meinersmann, C. M. Patton, G. M. Evins, I. K. Wachsmuth, and P. I. Fields. Genetic diversity and relationships of *Campylobacter* species and subspecies. *International Journal of Systematic and Evolutionary Microbiology*, 52(5):1789–1797 – 1797, 2002. ISSN 14665026.

L. Mentch and S. Zhou. Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *Journal of Machine Learning Research*, 23, 2022.

A. Merlotti, D. Remondini, G. Manfreda, F. Pasquali, N. Munck, T. Hald, E. Litrup, and E. M. Nielsen. Network approach to source attribution of *Salmonella enterica* serovar typhimurium and its monophasic variant. *Frontiers in Microbiology*, 11, 2020. ISSN 1664302X.

P. Miller, J. Marshall, N. French, and C. Jewell. sourcer: Classification and source attribution of infectious agents among heterogeneous populations. *PLOS Computational Biology*, 13(5), 2017.

M. W. Mitchell. Bias of the random forest Out-of-Bag (OOB) error for certain input parameters. *Open Journal of Statistics*, 01(03):205–211, 2011.

D. P. Mohandoss, Y. Shi, and K. Suo. Outlier prediction using random forest classifier. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), Computing and Communication Workshop and Conference (CCWC), 2021 IEEE 11th Annual*, pages 0027 – 0033, 2021. ISSN 978-1-6654-1490-6.

O. A. Montesinos López, A. Montesinos López, and J. Crossa. *Random Forest for Genomic Prediction.*, chapter 15, pages 633–681. Springer International Publishing, 2022. ISBN 978-3-030-89009-4.

Data Montgomery. Traffic violations, 2024. URL `https://data.montgomerycountymd.gov/Public-Safety/Traffic-Violations/4mse-ku6q/about_data`. Accessed: 2024-07-01.

P. Muellner, J. C. Marshall, S. E. F. Spencer, A. D. Noble, T. Shadbolt, J. M. Collins-Emerson, A. C. Midwinter, P. E. Carter, R. Pirie, D. J. Wilson, D. M. Campbell, M. A. Stevenson, and N. P. French. Utilizing a combination of molecular and spatial tools to assess the effect of a public health intervention. *Preventive Veterinary Medicine*, 102(3):242 – 253, 2011. ISSN 0167-5877.

L. Mughini-Gras, J. H. Smid, J. A. Wagenaar, A. G. de Boer, A. H. Havelaar, I. H. M. Friesema, N. P. French, L. Busani, and W. van Pelt. Risk factors for campylobacteriosis of chicken, ruminant, and environmental origin: A combined case-control and source attribution analysis. *PLoS ONE*, 7(8):1 – 13, 2012. ISSN 19326203.

L. Mughini-Gras, R. Pijnacker, C. Coipan, A. C. Mulder, S. de Rijk, A. H. A. M. van Hoek, S. Kuiling, A. Verbruggen, J. van der Giessen, M. Opsteegh, F. M. Schets, H. Blaak, E. Franz, A. Fernandes Veludo, R. Buij, G. Muskens, M. Koene, K. Veldman, B. Duim, L. van der Graaf-van Bloois, J. A. Wagenaar, A. L. Zomer, C. van der Weijden, M. van der Voort, and G. A. A. Castelijn. Sources and transmission routes of campylobacteriosis: A combined analysis of genome and exposure data. *Journal of Infection*, 82(2):216–226 – 226, 2021. ISSN 15322742.

L. Mughini-Gras, E. Benincà, S. A. McDonald, A. de Jong, J. Chardon, E. Evers, and A. A. Bonačić Marinović. A statistical modelling approach for source attribution meta-analysis of sporadic infection with foodborne pathogens. *Zoonoses and Public Health*, 69(5):475 – 486, 2022. ISSN 18631959.

L. Mughini-Gras, J. A. Paganini, R. Guo, C. E. Coipan, I. H. M. Friesema, A. H. A. M. van Hoek, M. van den Beld, S. Kuiling, I. Bergval, B. Wullings, M. van der Voort, E. Franz, and T. J. Dallman. Source attribution of *Listeria monocytogenes* in the Netherlands. *International journal of food microbiology*, 427:110953, 2025. ISSN 1879-3460.

P. Mullner, G. Jones, A. Noble, S. E. F. Spencer, S. Hathaway, and N. P. French. Source attribution of food-borne zoonoses in New Zealand: A modified Hald model. *Risk Analysis: An International Journal*, 29(7):970 – 984, 2009a. ISSN 02724332.

P. Mullner, S. E. F. Spencer, A. C. Midwinter, J. M. Collins-Emerson, N. P. French, S. Hathaway, D. J. Wilson, G. Jones, A. D. Noble, and P. Carter. Assigning the source of human campylobacteriosis in New Zealand: A comparative genetic and epidemiological approach. *Infection, Genetics and Evolution*, 9(6):1311–1319 – 1319, 2009b. ISSN 15671348.

P. Mullner, T. Shadbolt, J. M. Collins-Emerson, S. E. F. Spencer, J. Marshall, P. E. Carter, D. M. Campbell, D. J. Wilson, S. Hathaway, R. Pirie, and N. P. French. Molecular and spatial epidemiology of human campylobacteriosis: source association and genotype-related risk factors. *Epidemiology and Infection*, 138(10):1372 – 1383, 2010. ISSN 09502688.

N. Munck, P. M. K. Njage, P. Leekitcharoenphon, E. Litrup, and T. Hald. Application of whole-genome sequences and machine learning in source attribution of *Salmonella Typhimurium*. *Risk Analysis: An International Journal*, 40(9):1693 – 1705, 2020. ISSN 02724332.

O. Mutanga and E. Adam. High density biomass estimation: Testing the utility of vegetation indices and the random forest regression algorithm. *34th International Symposium on Remote Sensing of Environment - The GEOSS Era: Towards Operational Environmental Monitoring*, 2011.

I. Nachamkin, C. M. Szymanski, and M. J Blaser, editors. *Campylobacter*, volume 1267 (1). American Society for Microbiology Press, Washington, DC, USA, 2008. ISBN 978-1-55581-437-3.

S. Nembrini, I. R. König, and M. N. Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

D. G. Newell, L. Mughini-Gras, R. S. Kalupahana, and J. A. Wagenaar. *Campylobacter* epidemiology—sources and routes of transmission for human infection. In Günter Klein, editor, *Campylobacter*, pages 85–110. Academic Press, 2017. ISBN 978-0-12-803623-5. doi: https://doi.org/10.1016/B978-0-12-803623-5.00005-8.

K. K. Nicodemus. Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, 12(4):369–373, 2011.

K. K. Nicodemus and J. D. Malley. Predictor correlation impacts machine learning algorithms: Implications for genomic studies. *Bioinformatics*, 25(15):1884–1890, 2009.

K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11, 2010.

A. Nohra, A. Grinberg, N. French, A. C. Midwinter, and J. Collins-Emerson. Molecular epidemiology of *Campylobacter coli* isolated from different sources in New Zealand between 2005 and 2014. *Applied and Environmental Microbiology*, 82(14):4363–4370, 2016. ISSN 1098-5336.

A. Nohra, A. C. Midwinter, J. M. Collins-Emerson, N. P. Frencha, A. Grinberg, and J. C. Marshall. Shifts in the molecular epidemiology of *Campylobacter jejuni* infections in a sentinel region of New Zealand following implementation of food safety interventions by the poultry industry. *Applied and Environmental Microbiology*, 86(5), 2020. ISSN 10985336.

V. Novik, D. Hofreuter, and J. E. Galán. Identification of *Campylobacter jejuni* genes involved in its interaction with epithelial cells. *Infection and immunity*, 8(78):3540 – 3553, 2010.

I. D. Ogden, J. F. Dallas, M. MacRae, O. Rotariu, KW. Reay, M. Leitch, A. P. Thomson, S. K. Sheppard, M. Maiden, K. J. Forbes, and N. J Strachan. *Campylobacter* excreted into the environment by animal sources: prevalence, concentration shed, and host association. *Foodborne pathogens and disease*, 6(10):1161–1170, 2009.

S. L. W. On, B. Brett, S. Horan, H. Erskine, S. Lin, and A. J. Cornelius. Isolation and genotyping of *Campylobacter* species from kiwi (*Apteryx* spp.) in captivity: implications for transmission to and from humans. *New Zealand Veterinary Journal*, 67(3):134 – 137, 2019. ISSN 00480169.

J. Parkhill, B. W. Wren, K. Mungall, J. M. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. M. Davies, T. Feltwell, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Moule, M. J. Pallen, C. W. Penn, M. A. Quail, M. A. Rajandream, K. M. Rutherford, A. H. van Vliet, S. Whitehead, and B. G. Barrell. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403(6770):665–668, 2000.

B. Pascoe, G. Futcher, J. Pensar, S. C. Bayliss, E. Mourkas, J. K. Calland, M. D. Hitchings, L. A. Joseph, C. G. Lane, T. Greenlee, N. Arning, D. J. Wilson, K. A. Jolley, J. Corander, M. C. J. Maiden, C. T. Parker, K. K. Cooper, E. B. Rose, K. Hiett, Bruce B. B., and S. K. Sheppard. Machine learning to attribute the source of *Campylobacter* infections in the United States: a retrospective analysis of national surveillance data. *Journal of Infection*, 89(5), 2024.

B. M. Pearson, A. H. M. van Vliet, A. Rokney, L. C. Crossman, W. G. Miller, and J. Wain. Complete genome sequence of the *Campylobacter coli* clinical isolate 15-537360. *Genome Announcements*, 1(6), 2013. ISSN 21698287.

T. L. Pedersen, B. Nicolae, and R. François. farver: High performance colour space manipulation, 2024. URL `https://cran.r-project.org/package=farver`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

H. P. Pinheiro, A. de Souza Pinheiro, and P. K. Sen. Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference*, 130(1):325–339, 2005.

K. D. M. Pintar, K. M. Thomas, T. Christidis, A. Otten, A. Nesbitt, B. Marshall, F. Pollari, M. Hurst, and A. Ravel. A comparative exposure assessment of *Campylobacter* in Ontario, Canada. *Risk Analysis*, 37(4):677 – 715, 2017. ISSN 02724332.

S. M. Pires, H. Vigre, P. Makela, and T. Hald. Using outbreak data for source attribution of human salmonellosis and campylobacteriosis in Europe. *Foodborne pathogens and disease*, 7(11):1351–1361, 2010.

L. A. Pray. Discovery of DNA structure and function: Watson and crick. *Nature Education*, 1(1):100, 2008.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959 – 959, 2000. ISSN 00166731.

F. J. Pérez-Reche, O. Rotariu, B. S. Lopes, K. J. Forbes, and N. J. C. Strachan. Mining whole genome sequence data to efficiently attribute individuals to source populations. *Scientific Reports*, 10(1), 2020. ISSN 2045-2322.

J. R. Quinlan. *C4.5 : programs for machine learning.* The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, 1993. ISBN 1558602380.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

D. Reilly, M. Taylor, P. Fergus, C. Chalmers, and S. Thompson. The categorical data conundrum: Heuristics for classification problems - a case study on domestic fire injuries. *IEEE Access*, 10:70113–70125, 2022.

J. Revez, J. Zhang, T. Schott, R. Kivistö, M. Rossi, and M. Hänninen. Genomic variation between *Campylobacter jejuni* isolates associated with milk-borne-disease outbreaks. *Journal of Clinical Microbiology*, 52(8):2782–2786, 2014.

B. D. Ripley. *Pattern recognition and neural networks.* Cambridge University Press, 1996. ISBN 0521460867.

L. Rivas, P. Y. Dupont, B. Gilpin, and H. Withers. Prevalence and genotyping of *Campylobacter jejuni* and *Campylobacter coli* from ovine carcasses in New Zealand. *Journal of Food Protection*, 84(1):14–22, 2021. ISSN 0362-028X.

N. A. Rosenberg, L. M. Li, R. Ward, and J. K. Pritchard. Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, 73(6):1402–1422 – 1422, 2003. ISSN 00029297.

M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657, 2007.

N. A. Saif, J. F. Cobo-Díaz, M. Elserafy, I. El-Shiekh, A. Álvarez-Ordóñez, S. F. Mouftah, and M. Elhadidy. A pilot study revealing host-associated genetic signatures for source attribution of sporadic *Campylobacter jejuni* infection in Egypt. *Transboundary and Emerging Diseases*, 69(4):1847 – 1861, 2022. ISSN 18651674.

R. G. Same and P. D. Tamma. *Campylobacter* infections in children. *Pediatrics in review*, 39 (11):533–541, 2018.

M. Sandri and P. Zuccolotto. A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17:611–628, 2008.

E. J. Scallan Walter, S. M. Crim, B. B. Bruce, and P. M. Griffin. Incidence of *Campylobacter*-associated Guillain-Barré Syndrome estimated from health insurance data. *Foodborne Pathogens and Disease*, 17(1):23–28 – 28, 2020. ISSN 15567125.

W. Schumacher, S. Stöckel, P. Rösch, and J. Popp. Self-defining tree-like classifiers for interpretation of raman spectroscopic experiments. *Journal of Chemometrics*, 30(5):268 – 283, 2016. ISSN 08869383.

W. G. Scott, H. M. Scott, R. J. Lake, and M. G. Baker. Economic cost to New Zealand of foodborne infectious disease. *New Zealand Medical Journal*, 113(1113):281–4, 2000. ISSN 0028-8446.

A. Sears, M. G. Baker, N. Wilson, J. Marshall, P. Muellner, D. M. Campbell, R. J. Lake, and N. P. French. Marked campylobacteriosis decline after interventions aimed at poultry, New Zealand. *Emerging Infectious Diseases*, 17(6):1007 – 1015, 2011. ISSN 10806040.

R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. parts I and II. *Psychometrika*, 27(3):125–140, 219–246, 1962. ISSN 0033-3123.

S. K. Sheppard and M. C. J. Maiden. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. *Cold Spring Harbor Perspectives in Biology*, 7(8):13p., 2015. ISSN 19430264.

S. K. Sheppard, J. F. Dallas, N. J. C. Strachan, M. MacRae, N. D. McCarthy, D. J. Wilson, F. J. Gormley, D. Falush, I. D. Ogden, M. C. J. Maiden, and K. J. Forbes. *Campylobacter* genotyping to determine the source of human infection. *Clinical Infectious Diseases*, 48(8): 1072 – 1078, 2009. ISSN 10584838.

S. K. Sheppard, F. Colles, A. J. Cody, M. C. J. Maiden, N. D. McCarthy, J. Richardson, R. Elson, A. Lawson, G. Brick, C. L. Little, R. J. Owen, and R. Meldrum. Host association of *Campylobacter* genotypes transcends geographic variations. *Applied and Environmental Microbiology*, 76(15):5269–5277 – 5277, 2010a. ISSN 00992240.

S. K. Sheppard, J. F. Dallas, D. J. Wilson, N. J. C. Strachan, N. D. McCarthy, K. A. Jolley, F. M. Colles, O. Rotariu, I. D. Ogden, K. J. Forbes, and M. C. J. Maiden. Evolution of an agriculture-associated disease causing *Campylobacter coli* clade: Evidence from national surveillance data in scotland. *PLoS ONE*, 5(12):1 – 9, 2010b. ISSN 19326203.

S. K. Sheppard, K. A. Jolley, and M. C. J. Maiden. A gene-by-gene approach to bacterial population genomics: Whole genome MLST of *Campylobacter*. *Genes*, 3(2):261 – 277, 2012. ISSN 2073-4425.

S. K. Sheppard, K. A. Jolley, A. Cody, F. M. Colles, N. D. McCarthy, M. C. J. Maiden, B. Pascoe, G. Meric, X. Didelot, A. E. Darling, D. J. Kelly, N. J. C. Strachan, I. D. Ogden, K. Forbes, N. P. French, P. Carter, W. G. Miller, R. Owen, E. Litrup, M. Egholm, J. P. Affourtit, S. D. Bentley, J. Parkhill, and D. Falush. Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Molecular Ecology*, 22(4):1051–1064 – 1064, 2013a. ISSN 1365294X.

S. K. Sheppard, D. Xavier, G. Meric, T. Alicia, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. J. Maiden, J. Parkhi, and D. Falush. Genome-wide association study identifies vitamin b5 biosynthesis as a host specificity factor in *Campylobacter*. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11923 – 11927, 2013b. ISSN 00278424.

A. Signorell, K. Aho, A. Alfons, N. Anderegg, T. Aragon, C. Arachchige, A. Arppe, A. Baddeley, K. Barton, B. Bolker, H. W. Borchers, F. Caeiro, S. Champely, D. Chessel, L. Chhay, N. Cooper, C. Cummins, M. Dewey, H. C. Doran, S. Dray, C. Dupont, D. Eddelbuettel, C. Ekstrom, M. Elff, J. Enos, R. W. Farebrother, J. Fox, R. Francois, M. Friendly, T. Galili, M. Gamer, J. L. Gastwirth, V. Gegzna, Y. R. Gel, S. Graber, J. Gross, G. Grothendieck, F. E. Harrell Jr., R. Heiberger, M. Hoehle, C. W. Hoffmann, S. Hojsgaard, T. Hothorn, M. Huerzeler, W. W. Hui, P. Hurd, R. J. Hyndman, S. Jackson, M. Kohl, M. Korpela, M. Kuhn, D. Labes, F. Leisch, J. Lemon, D. Li, M. Maechler, A. Magnusson, B. Mainwaring, D. Malter, G. Marsaglia, J. Marsaglia, A. Matei, D. Meyer, W. Miao, G. Millo, Y. Min, D. Mitchell, C. F. Moser, F. Mueller, M. Naepflin, D. Navarro, H. Nilsson, K. Nordhausen, D. Ogle, H. Ooi, N. Parsons, S. Pavoine, T. Plate, L. Prendergast, R. Rapold, W. Revelle, T. Rinker, B. D. Ripley, C. Rodriguez, N. Russell, N. Sabbe, R. Scherer, V. E. Seshan, M. Smithson, G. Snow, K. Soetaert, W. A. Stahel, A. Stephenson, M. Stevenson, R. Stubner, M. Templ, D. T. Lang, T. Therneau, Y. Tille, L. Torgo, A. Trapletti, J. Ulrich, K. Ushey, J. VanDerWal, B. Venables, J. Verzani, P. J. V. Iglesias, G. R. Warnes, S. Wellek, H. Wickham, R. R. Wilcox, P. Wolf, D. Wollschlaeger, J. Wood, Y. Wu, T. Yee, and A. Zeileis. Desctools: Tools for descriptive statistics, 2024. URL `https://cran.r-project.org/package=DescTools`.

M. F. Silva, G. Pereira, C. Carneiro, A. Hemphill, L. Mateus, L. Lopes-da Costa, and E. Silva. *Campylobacter portucalensis* sp. nov., a new species of *Campylobacter* isolated from the preputial mucosa of bulls. *PLoS ONE*, 15(1):1 – 16, 2020. ISSN 19326203.

N. Singh, C. Thystrup, B. M. Hassen, M. Bhandari, H. Rajashekara, T. Hald, M. J. Manary, S. L. McKune, J. Y. Hassen, H. L. Smith, J. M. Marshall, N. French, and A. H. Havelaar. Transmission pathways of *Campylobacter jejuni* between humans and livestock in rural ethiopia are highly complex and interdependent. *Gut Pathogens*, 17(26):1–16, 2025. doi: 10.1186/s13099-025-00691-7.

H. L. Smith, P. J. Biggs, N. P. French, A. N. H. Smith, and J. C. Marshall. CAP-encoding: Encoding categorical variables using Canonical Analysis of Principal Coordinates. *Machine Learning in review*, 10:e2445, 2024a. doi: 10.7717/peerj-cs.2445.

H. L. Smith, P. J. Biggs, N. P. French, A. N. H. Smith, and J. C. Marshall. Lost in the forest: Encoding categorical variables and the absent levels problem. *Data Mining and Knowledge Discovery*, 38:1889–1908, 2024b. ISSN 1573-756X. doi: 10.1007/s10618-024-01019-w.

H. L. Smith, P. J. Biggs, N. P. French, A. N. H. Smith, and J. C. Marshall. Out of (the) bag - encoding categorical predictors impacts out-of-bag samples. *PeerJ Computer Science*, 10:e2445, 2024c. doi: 10.7717/peerj-cs.2445.

N. J. C. Strachan, F. J. Gormley, O. Rotariu, I. D. Ogden, G. Miller, G. M. Dunn, S. K. Sheppard, J. F. Dallas, T. M. S. Reid, H. Howie, M. C. J. Maiden, and K. J. Forbes. Attribution of *Campylobacter* infections in Northeast Scotland to specific sources by use of multilocus sequence typing. *The Journal of Infectious Diseases*, 199(8):1205 – 1208, 2009. ISSN 00221899.

C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

C. Strobl, A. L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 2008.

S. Suerbaum, M. Lohrengel, A. Sonnevend, F. Ruberg, and M. Kist. Allelic diversity and recombination in *Campylobacter jejuni*. *Journal of Bacteriology*, 183(8):2553–2559 – 2559, 2001. ISSN 00219193.

S. Szymczak, E. Holzinger, A. Dasgupta, J. D. Malley, A. M. Molloy, J. L. Mills, L. C. Brody, D. Stambolian, and J. E. Bailey-Wilson. r2VIM: A new variable selection method for random forests in genome-wide association studies. *BioData Mining*, 9(1), 2016.

C. K. Tanui, E. O. Benefo, S. Karanth, and A. K. Pradhan. A machine learning model for food source attribution of *Listeria monocytogenes*. *Pathogens*, 11(6), 2022. ISSN 20760817.

D. E. Taylor, M. Eaton, W. Yan, and N. Chang. Genome maps of *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Bacteriology*, 174(7):2332–2337 – 2337, 1992. ISSN 00219193.

T. Therneau, B. Atkinson, and B. Ripley. rpart: Recursive partitioning and regression trees, 2022. URL `http://CRAN.R-project.org/package=rpart`.

C. Thystrup, T. Hald, D. Belina, and T. Gobena. Outbreak detection in Harar town and Kersa district, ethiopia using phylogenetic analysis and source attribution. *BMC Infectious Diseases*, 24(1), 2024. ISSN 1471-2334.

A. Thépault, K. Rivoal, V. Rose, M. Chemaly, G. Méric, B. Pascoe, S. K. Sheppard, L. Mageiros, F. Touzain, and V. Béven. Genome-wide identification of host-segregating epidemiological markers for source attribution in *Campylobacter jejuni*. *Applied and Environmental Microbiology*, 83(7), 2017. ISSN 10985336.

A. Thépault, V. Rose, S. Quesne, T. Poezevara, V. Béven, E. Hirchaud, F. Touzain, P. Lucas, G. Méric, L. Mageiros, S. K. Sheppard, M. Chemaly, and K. Rivoal. Ruminant and chicken: important sources of campylobacteriosis in France despite a variation of source attribution in 2009 and 2015. *Scientific Reports*, 8(1), 2018. ISSN 2045-2322.

A. Thépault, V. Rose, M. Queguiner, M. Chemaly, and K. Rivoal. Dogs and cats: Reservoirs for highly diverse *Campylobacter jejuni* and a potential source of human exposure. *Animals*, 10 (5), 2020. ISSN 20762615.

L. Toloşi and T. Lengauer. Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

J. J. Truglio, D. L. Croteau, B. Van Houten, and C. Kisker. Prokaryotic nucleotide excision repair: The UvrABC system. *Chemical Reviews*, 106(2):233–252, 2006. ISSN 00092665.

W. van Pelt, A.W. Giessen, W. J. Leeuwen, W. Wannet, H. André, E. G. Evers, M. De Wit, and Y. T. H. P. Duynhoven. Origin, extent and costs of human salmonellosis. part 1 origin of human salmonellosis with respect to pig, cattle, chicken, eggs and other sources (in Dutch). *Infectieziekten Bulletin*, 10:240–243, 01 1999.

L. Wainaina, A. Merlotti, D. Remondini, C. Henri, P. M. K. Njage, and T. Hald. Source attribution of human campylobacteriosis using whole-genome sequencing data and network analysis. *Pathogens*, 11(6), 2022. ISSN 20760817.

M. L. Wallace, L. Mentch, B. J. Wheeler, A. L. Tapia, M. Richards, S. Zhou, L. Yi, S. Redline, and D. J. Buysse. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, 23 (1), 2023. ISSN 1471-2288.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645 – 1658, 2023.

D. J. Wilson, E. Gabriel, A. J. H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, P. Fearnhead, C. A. Hart, and P. J. Diggle. Tracing the source of campylobacteriosis. *PLoS Genetics*, 4(9):1 – 9, 2008. ISSN 15537390.

D. J. Wilson, E. Gabriel, A. J. H. Leatherbarrow, J. Cheesbrough, S. Gee, E. Bolton, A. Fox, C. A. Hart, P. J. Diggle, and P. Fearnhead. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, 26(2):385 – 397, 2009. ISSN 0737-4038.

T. L. Wong, L. Hollis, A. Cornelius, C. Nicol, R. Cook, and J. A. Hudson. Prevalence, numbers, and subtypes of *Campylobacter jejuni* and *Campylobacter coli* in uncooked retail meat samples. *Journal of Food Protection*, 70(3):566 – 573, 2007. ISSN 0362028X.

M. N. Wright and I. R. König. Splitting on categorical predictors in random forests. *PeerJ*, 2019 (2):e6339, 2019.

M. N. Wright and A. Ziegler. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

S. Yu, P. Fearnhead, B. R. Holland, P. Biggs, M. Maiden, and N. French. Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli*. *Journal of Molecular Evolution*, 74(5-6):273 – 280, 2012. ISSN 0022-2844.

S. Zhang, S. Li, H. Den Bakker, X. Deng, W. Gu, B. B. Bruce, E. Trees, P. I. Fields, D. Boxrud, A. Taylor, C. Roe, E. Driebe, D. M. Engelthaler, M. Allard, E. Brown, P. McDermott, and S. Zhao. Zoonotic source attribution of *Salmonella enterica* serotype typhimurium using genomic surveillance data, United States. *Emerging Infectious Diseases*, 25(1):82–91 – 91, 2019. ISSN 10806059.

A. Ziegler and I. R. König. Mining data with random forests: Current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(1): 55–63, 2014.

A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):1 – 17, 2017. ISSN 1474-760X.

M. Zilbauer, N. Dorrell, B. W. Wren, and M. Bajaj-Elliott. *Campylobacter jejuni*-mediated disease pathogenesis: an update. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(2):123 – 129, 2008. ISSN 00359203.