Source Attribution: The R Package sourceR

Poppy Miller Massey University
Jonathan Marshall Massey University
Nigel French Massey University
Chris Jewell Massey University

Abstract

Numerous zoonotic diseases cause morbidity, mortality and productivity losses in both humans and animal populations. For many zoonotic diseases that are important to human health (such as campylobacteriosis), it is difficult to attribute human cases to sources because there is little epidemiological information on the cases. Genotyping systems allow the zoonotic pathogens to be categorised, and the relative distribution of the genotypes among the sources (food sources or reservoirs of bacteria) and in human cases allows inference on the likely source of each genotype. Current source attribution models, specifically the Island model (Wilson, Gabriel, Leatherbarrow, Cheesebrough, Hart, and Diggle 2008), Hald (Hald, Vose, Wegener, and Koupeev 2004) and modified Hald models (Mullner, Jones, Noble, Spencer, Hathaway, and French 2009) are not fully joint and have many (often unverifiable) assumptions. Identifiability of the parameters in this model is an issue because a large number of parameters need to be estimated, the data is imbalanced, and many of the combinations of source and type and have very low counts. We present techniques to overcome these issues within a Bayesian framework by developing a fully joint model which non-parametrically clusters the type effects (using a Dirichlet Process) allowing identification of groups of bacterial subtypes with similar pathogenicity, survival and/ or virulence mechanisms. This model is applied to *Campylobacter* data from the Manawatu area of New Zealand (previously analysed by Mullner *et al.* (2009)) using the sourceR package, and compared to current source attribution models.

Keywords: bayesian, source attribution, food-bourne zoonoses, non-parametric, R.

library(tikzDevice)

1. Introduction

1.1. Background

Food-borne diseases are a major source of human morbidity and mortality world wide. In 2010, an estimated 600 million cases occurred globally, of which approximately 90% were cause by food borne diarrhoeal disease pathogens (Havelaar, Kirk, Torgerson, Gibb, Hald, Lake, Praet, Bellinger, de Silva, Gargouri, Speybroeck, Cawthorne, Mathers, Stein, Angulo, Devleesschauwer, and on behalf of World Health Organization Foodborne Disease Burden Epidemiology Reference Group 2015). Identifying the source from which a food-borne disease is acquired, and the pathway by which it enters the food chain, is crucial for the identification and prioritization of food safety interventions. Traditional approaches to source attribution include full risk assessments, analysis and extrapolation of surveillance or outbreak data, and analytical epidemiological studies (Crump, Griffin, and Angulo 2002). However, their results

may be highly uncertain due to long and variable incubation times of food-borne diseases in the face of many and various exposures of an individual to potential sources. Given this difficulty, quantitative methods using pathogen genotype frequency have become popular for identifying important sources of food-borne illness (Mullner *et al.* 2009).

For a given disease, quantitative source attribution relies on molecular typing data from pathogen genotypes isolated from human cases as well as from a number of putative sources of infection. For bacterial diseases, source samples are usually collected from food (such as raw chicken, beef etc) and environmental (such as water or faeces samples) sources, and tested for the presence of the zoonotic bacterium usually using polymerase chain reaction (PCR) methods. The bacterial samples are then categorised into subtypes using a genetic typing methodology. Multilocus Sequence Typing (MLST) is commonly used because it is a relatively inexpensive, rapid, and unambiguous procedure for coarse characterisation of isolates of bacterial species. Here, genetic variations in small fragments of several house-keeping genes are assigned distinct allelic identifiers. A sequence type is therefore defined as a unique combination of alleles at each gene locus (Dingle, Colles, Wareing, Ure, Fox, Bolton, Bootsma, Willems, Urwin, and Maiden 2001). Being defined on conserved regions of the bacterial genome, the evolution of new MLST types is slow, enabling data collected over a period of months to be classed as cross-sectional, making it suitable for use in source attribution models. Recent statistical approaches designed specifically to use MLST data are reviewed in Section 3.

Routine surveillance for food-borne pathogens is now commonplace in many countries and is performed by national authorities, for example FoodNet in the US (Allos, Moore, Griffin, and Tauxe 2004), the Danish Zoonosis Centre (`food.dtu.dk`), and the Ministry for Primary Industries in New Zealand (`foodsafety.govt.nz`). However, despite this availability of data there are no implementations in standard statistical software for source attribution modelling, with analyses being performed using a variety of *ad hoc* methodologies. Moreover, as the example of human *Campylobacter jejuni* cases in New Zealand between 2005 and 2007 shows, current statistical source attribution models are subject to computational approximations and inherent identifiability problems.

This paper presents an `R` package **sourceR** implementing a flexible Bayesian non-parametric model, designed for use by epidemiologists and other scientists to attribute cases of zoonotic infection to putative sources of infection. The paper is structured as follows. We first describe a motivating example in Section 2, before briefly reviewing a set of related models that have been previous applied to this dataset in Section 3, and for which our model represents a significant advance. We describe our source attribution model in Section 4.1, and demonstrate its utility through worked examples on simulated and real-world data in Sections 5.1 and 5.2 respectively.

2. Motivating example

Campylobacter is the most commonly identified cause of food-borne bacterial gastro-enteritis in the developed world (Miller, On, Wang, Fontanoz, Lastovica, and Mandrell 2005) is estimated to be responsible for over 26% of bacterial foodbourne illnesses world-wide (Havelaar *et al.* 2015). In 2006, New Zealand had one of the highest incidences of campylobacteriosis in the developed world, with an annual incidence in excess of 400 cases per 100,000 people

(Baker, Wilson, Ikram, Chambers, Shoemack, and Cook 2006). A campaign to change poultry processing procedures, supported in part by results from quantitative source attribution methods, was successful in leading to a sharp decline in campylobacteriosis incidence after 2007 (Mullner *et al.* 2009). This example provides the dataset that motivates the construction of the **sourceR** package. The dataset was first published in Mullner, Collins-Emerson, Midwinter, Carter, Spencer, van der Logt, Hathaway, and French (2010), with a detailed description of the data (and data collection methods) available in French and Marshall (2009) and French and Marshall (2013).

Briefly, our data consist of MLST-genotyped *Campylobacter jejuni* isolates from both human cases of campylobacteriosis and potential food and environmental sources between 2005 and 2008 in the Manawatu region of New Zealand. The human isolates were obtained from the local medical microbiology service (MedLab Central, Palmerston North), with isolates from food and environmental sources collected during a sample-based surveillance study. Samples of beef and lamb were collected from local retail stores, water from popular local riverine swimming locations, and sheep and cattle faeces from farms within local river catchments. These samples were then grouped into one of six sources: poultry supplier A, poultry supplier B, poultry supplier C, bovine (beef mince and liver, and cattle faecal samples), ovine (lamb mince and liver, and sheep faecal samples) and environmental (water samples).

These data are included within **sourceR**, named `campy`, comprising a data frame of the number of positive isolates of each MLST type identified from humans and each potential source of infection. We use this dataset as a source attribution case study in Section 5.2, comparing our results with previously published *ad hoc* statistical approaches.

3. Review of models and notation

This section briefly reviews the current source attribution models. Throughout, we adopt a convention where $i = 1, \ldots, n$ denotes a bacterial subtype, and $j = 1, \ldots, m$ denotes a putative source of infection.

3.1. Dutch model

The Dutch method (van Pelt, van de Giessen, van Leeuwen, Wannet, Henken, and Evers 1999) is one of the simplest models for source attribution. It compares the number of reported human cases caused by a particular bacterial subtype with the relative occurrence of that subtype in each source. The number of reported cases per subtype and reservoir is estimated by:

$$\lambda_{ij} = \frac{r_{ij}}{\sum_j r_{ij}} y_i \tag{1}$$

where r_{ij} is the relative occurrence of bacterial subtype i in source j, y_i is the estimated number of human cases of type i per year, λ_{ij} is the expected number of cases per year of type i from source j. A summation across subtypes gives the total number of cases attributed to source j, denoted by λ_j:

$$\lambda_j = \sum_i \lambda_{ij} \tag{2}$$

As the Dutch model has no inherent statistical noise model, confidence intervals for the estimated total attributed cases $\hat{\lambda_j}$ by bootstrap sampling over the dataset.

3.2. Hald model

One of the first stochastic source attribution models was the Hald model (Hald *et al.* 2004) which was developed to model the source attribution of salmonellosis in Denmark. It extends the Dutch method by incorporating source and type effect parameters into the model, and assuming that the number of human cases are Poisson distributed conditional on the source typing data. Source and type effect parameters are used to account for source- and type-specific influences on the rates. Type effects summarise the characteristics that determine a type's capacity to cause an infection (survivability, pathogenicity and virulence). Source effects account for the ability of a particular food source to act as a vehicle of infection. This is a significant advantage over the Dutch model as it is not plausible that type and source effects are equal for most zoonoses. Inference is performed in a Bayesian framework allowing the model to explicitly include and quantify the uncertainty surrounding each of the parameters.

The number of human cases y_i of isolate type $i = 1, \ldots, n$ is Poisson distributed such that

$$y_i \sim \text{Poisson}(\lambda_i) \tag{3}$$

$$\lambda_i = q_i \sum_{j=1}^{m} a_j c_j p_{ij} \tag{4}$$

where $p_{ij} = r_{ij} \times \pi_j$ is the absolute prevalence of each type in source j, π_j is the prevalence of positive samples in source j, c_j is the offset for the annual consumption of each food source j, n_j is the total number of samples from each source, $r_{ij} = \frac{x_{ij}}{\sum_{i=1}^{I} x_{ij}}$ is the relative prevalence of each type in source j, x_{ij} is the number of MLST positive samples for type i in source j, a_j is the j^{th} source effect, and q_i is the i^{th} type effect. The rate of cases attributed to each source is given by $\lambda_j = \sum_{i=1}^{n} a_j c_j p_{ij}$.

Note that the prevalence π_j is calculated by dividing the number of positive samples (using PCR to detect the presence of *Campylobacter*) by the total number of samples for each source. Samples testing positive for *Campylobacter* using PCR are MLST typed. It is possible for MLST typing to fail, hence, the number of positive samples for a given source (used in the prevalence calculation) can exceed the number of source samples used in the source data matrix (x).

This model is overparameterised because there are $m + n$ parameters (the source and type effects) but only n independent observations (the observed human case totals y_i). Identifiability was obtained by assuming some source and type effects were equal. This was done by pooling the bacterial subtypes into groups (where types within the same group have the same type effect) and assuming the source effects were the same for Danish and imported pork. Although in some cases there may be some physical justification to set some parameters equal, it is not possible for all zoonoses. Furthermore, the intensity of the source surveillance system in Denmark justified the use of point estimates of p_{ij}, rather than explicitly modelling the source sampling process.

3.3. Modified Hald model

The Modified Hald model (Mullner *et al.* 2009) was developed because the Hald model had some assumptions that were not suitable for modelling campylobacteriosis. There was no evidence to justify *a priori* fixing some source and type effects to be equal, and the source data came from a less intensive surveillance system with fewer source samples taken (suggesting

it would be beneficial to introduce uncertainty into the source prevalence matrix). For their application, they wished to include the environment as a potential source of infection. Since it is not possible to quantify annual exposure to the environment, the annual consumption offset was removed from the model.

The number of human cases y_i of isolate type $i = 1, \ldots, n$ is again Poisson distributed with rate λ_i for each type i as in Equation 4, omitting the annual consumption term c_j. In contrast to the Hald model, identifiability of the model is ensured by treating $\boldsymbol{q}$ as a a log Normal$(0, \tau)$ distributed random effect. However, a strong prior is needed on τ to shrink $\boldsymbol{q}$ towards 0 sufficiently to avoid overfitting the model, the choice of which is arbitrary.

In a further development, the modified Hald model introduces uncertainty into the relative prevalence matrix by modelling the source sampling process. The p_{ij}'s were first modelled in a separate Bayesian scheme, where independent symmetric Dirichlet priors were used to model columns of the $\boldsymbol{r}$ matrix, and a non-informative Beta distribution was used for the source prevalences:

$$r_{\cdot j} \sim \mathsf{Dirichlet}(\mathbf{1}) \; \forall \; j \tag{5}$$

$$\pi_j \sim \mathsf{Beta}(1, 1) \; \forall \; j \tag{6}$$

This model was fitted in WinBUGS using an approximate two stage process (Mullner *et al.* 2009). First, a posterior distribution was estimated for the absolute prevalence of source subtypes $\boldsymbol{p}$, using the model specified in Equations 5 and 6. The marginal posterior for each element of $\boldsymbol{p}$ was then approximated by a Beta distribution

$$p_{ij} \sim \mathsf{Beta}(\alpha_{ij}, \beta_{ij})$$

using the method of moments for α_{ij} and β_{ij}. which were included as independent priors in the Poisson model. Due to convergence issues for very small α_{ij} values, α_{ij} was limited to be at least 1 and the β_{ij} parameter was adjusted accordingly (French and Marshall 2009).

Using independent Beta priors on each p_{ij} removes the constraint that they sum to π_j over each type i. Thus, the absolute prevalence for source j ($\sum_{i=1}^{I} p_{ij}$) is no longer constrained to be a probability.

3.4. Asymmetric Island model

The Asymmetric Island Model (Wilson *et al.* 2008; Wilson 2016) takes a different approach to the models described above. Here, the evolutionary processes (mutation, migration and recombination) of the sequence types is modelled to probabilistically infer the source of each human infection. This means it requires genetic typing for all samples limiting the range of data that can be used with this model (for example, phenotypic typing cannot be used). The extra information in the genetic typing allows the model to attribute human cases not observed in any sources to a likely source of infection by looking at the genetic similarity of that type to other types that are observed in the sources; this is not possible with the Dutch, Hald or Modified Hald models. However, they are much simpler with fewer strong assumptions and work with a wider range of data than the Island model.

4. Methods

In this section, we address the problems inherent in both the Hald and Modified Hald models. Our approach builds on these models by introducing a fully joint model for both source and human case sampling. This allows us to integrate over uncertainty in the source sampling process, estimating both the prevalence of contaminated source samples and the relative prevalence of each identified subtype, without resorting to an approximate marginal probability distribution on $\boldsymbol{r}$. Furthermore, we introduce non-parametric clustering of pathogen types using a Dirichlet process model on the type effect vector $\boldsymbol{q}$, providing an automatic data-driven way of reducing the dimensionality of $\boldsymbol{q}$ to aid model identifiability. We are able, therefore, to circumvent the Hald model requirement for heuristically grouping pathogen types (Section 3.2), as well as avoiding an arbitrarily strong prior distribution on a random effect precision parameter as required by the Modified Hald model (Section 3.3).

4.1. Model

As with the Hald and Modified Hald models, the number of human cases y_i identified by isolation of subtype i is assumed to be Poisson distributed so that

$$y_i \sim \mathsf{Poisson}(\lambda_i) \tag{7}$$

The mean intensity λ_i is a linear combination of type and source-specific effects such that

$$\lambda_i = q_i \sum_{j=1}^{m} a_j p_{ij} \tag{8}$$

where a_j represents the source effect, p_{ij} the absolute prevalence of subtype i in samples from source j, and q_i is the type effect for subtype i.

For each source $j = 1, \ldots, m$, we model the number of positive samples x_{ij} identified as type $i = 1, \ldots, n$ as

$$\boldsymbol{x}_j \sim \text{Multinomial}(n_j, \boldsymbol{r}_j) \tag{9}$$

where $\boldsymbol{x}_j$ denotes the vector of type-counts in source j, n_j denotes the number of positive samples obtained from source j, and $\boldsymbol{r}_j$ denotes a vector of relative prevalences of isolate types in source j. The advantage of this model is that it automatically places the constraint $\sum_{i=1}^{n} r_{ij} = 1$, avoiding the approximation made in Mullner *et al.* (2009) where independent Beta-distributed priors were assigned marginally to components of $\boldsymbol{r}_j$. The source case model is then coupled to the human case model through the simple relationship

$$p_{ij} = r_{ij}\pi_j \tag{10}$$

where π_j is the prevalence of any isolate in source j.

We note that in principle, a Beta distribution could be used to model π_j, arising as the conjugate posterior distribution of a Binomial sampling model for x_j positive samples from n_j tested, and a Beta prior on π_j. However, since within a particular source the number of positive and negative samples are typically high, we choose to fix the source prevalences at their point estimates ($\pi_j = x_j/n_j$).

The type effects, $\boldsymbol{q}$ are drawn from a Dirichlet Process

$$q_i \sim \text{DP}\left(\alpha_q, Q_0\right). \tag{11}$$

The Dirichlet Process is a random probability measure defined by a base distribution Q_0 and a concentration parameter α_q (Ferguson 1973). The base distribution constitutes a prior distribution in the values of each element of the type effects $\mathbf{q}$ whilst the concentration parameter encodes prior information on the number of groups K to which each subtype i is assigned. For small values of α_q, samples from the DP are likely to have a small number of atomic measures with large weights. For large values, most samples are likely to be distinct, and hence, concentrated on Q_0. A value of 1 implies that, *a priori*, two randomly selected types have probability 0.5 of belonging to the same cluster (Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin 2013).

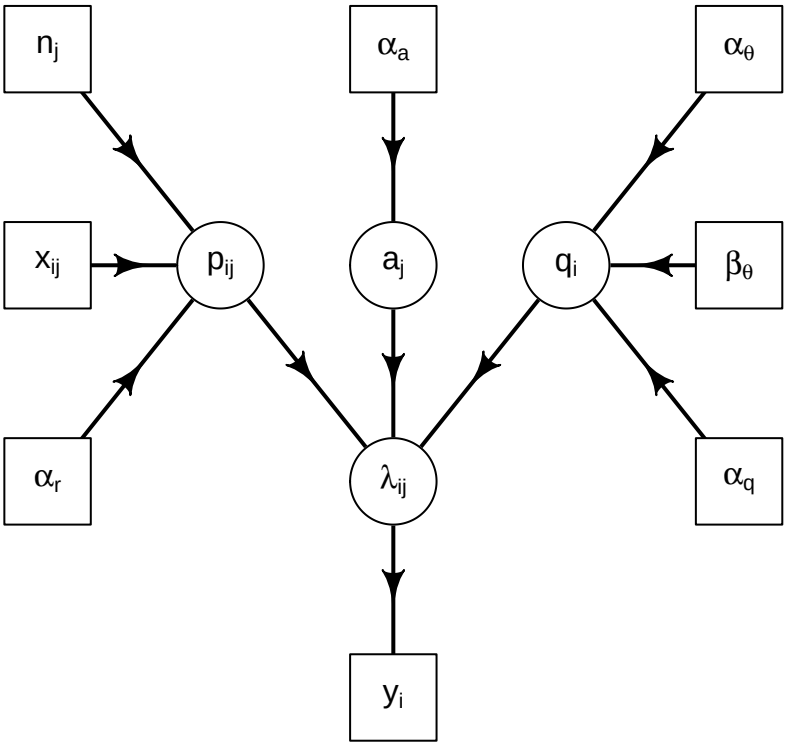

Figure 1: Directed acyclic graph of the source attribution model. See Table 1 for a concise description of the parameters.

Table 1: Description and definition of parameters used in our model.

Parameter	Description	Estimation
λ_{ij}	Number of human cases from type i, source j	$\lambda_{ij} = a_j \cdot q_{k(i)} \cdot r_{ij} \cdot \pi_j$
λ_i	Number of human cases from type i	$\lambda_i = \sum_{j=1}^{m} \lambda_{ij}$
λ_j	Number of human cases from source j	$\lambda_j = \sum_{i=1}^{n} \lambda_{ij}$
y_i	Number of human cases from type i	$y_i \sim \mathsf{Poisson}(\lambda_i)$
x_{ij}	Number of positive samples (that were successfully MLST typed) from source j, type i	Data
h_{ij}	Number of positive samples (PCR) that could not be MLST typed.	Data
n_j	Total number of samples from source j	Data
π_j	Prevalence of contamination for each source	$\sum_{i=1}^{I}(x_{ij} + h_{ij})/n_j$
r_{ij}	Relative occurrence of type i on source j	$\mathbf{r}_j \sim \mathsf{Dirichlet}(\alpha_r)$ or $x_{ij}/\sum_{i=1}^{n} x_{ij}$
p_{ij}	Absolute prevalence of type i in source j	$r_{ij} \cdot \pi_j$
a_j	Unknown source effect for source j	$\mathbf{a} \sim \mathsf{Dirichlet}(\alpha_a)$
q_i	Unknown type effect for type i in group k, where group k has an unknown value θ_k	$\mathbf{q} \sim \mathsf{DP}(\mathsf{Gamma}(\alpha_\theta, \beta_\theta), \alpha_q)$

4.2. Model extensions

The models can further be extended to incorporate a time and location dependence into the model allowing different rates over time and in different locations (such as urban vs rural cases). Let λ_{ijtl} be the expected number of infections of sequence type i attributable to source j at time t with location l. Then the observed human counts from a particular type i during that time period in a particular location y_{itl} is given by

$$y_{itl} \sim \mathsf{Poisson}(\sum_{j=1}^{m} \lambda_{ijtl})$$

where

$$\lambda_{ijtl} = q_i a_{jtl} r_{ijt} \pi_{jt}. \tag{12}$$

Note that type effects $\boldsymbol{q}$ are assumed to be constant over all times and locations, and source effects $\boldsymbol{a}$ are allowed to vary between times and locations. Importantly, the source sampling information ($\boldsymbol{r}_t$ and $\boldsymbol{\pi}_t$) are allowed to vary by time only. This is because of the the nature of food source sampling at the point of sale, where food retailers move packaged meat long distances from the farm to retail store. This is implemented in **sourceR** as shown in Section 5.1.

5. Case Studies

In this section we demonstrate the use of the **sourceR** package using two example data sets. The first uses simulated data to demonstrate the general use case, with both time and location

Table 2: Minimum parameters required for the `saBayes` function.

`formula`	A formula object of the form $y \sim x1 + x2 + ... + xJ$, where y is the name of the human cases column, and $x1, ..., xJ$ are the names of the source count columns in the data.
`time`	A formula object of the form $\sim t$, where t is the name of the column containing the times in the data.
`location`	A formula object of the form $\sim l$, where l is the name of the column containing the locations in the data.
`type`	A formula object of the form $\sim s$, where s is the name of the column containing the (sub) types in the data.
`data`	Correctly formatted data.
`priors`	List with parameters for the prior distributions for each of the model parameters.
`n_iter`	Specifies the number of iterations to run the algorithm for.
`likelihood_dist`	Specifies the likelihood distribution to be used for the human cases from each type. Must be one of `nbinom` or `pois`.

extensions for the model. The second is a specific case study using the the data described in the motivation section (Section 2). The results (proportion of cases attributed to each source) are compared to the results from the Dutch, Modified Hald and Island models.

5.1. Simulated data

In this section, we provide a worked example using simulated data with multiple times and locations for source attribution data generated from the model in Section 4.1 (available in the **sourceR** data sets). There are two times (1, 2) and two locations (A, B) over which the human cases vary. The data expected by `saBayes` is in long format, with a column for the number of human cases, the number of positive samples for each source, and columns identifying the type, time and locations. Note, the source data is the same for all locations within a time. For this data, the source prevalences (π_j) are all set to be 1. If source prevalences are not provided, `saBayes` will automatically set them all to 1 (with a warning). The data must be in long format, with columns giving the number of human cases for each type, a column for each of the sources giving the number of positive samples for each type, and columns giving the time, location and type id's for each observation.

```
require(sourceR)
set.seed(63164)
data(sim_SA)
data(sim_SA_true)

# Set priors
priors <- list(a = 1, r = 1, theta = c(0.01, 0.00001))

# Run model
res_sim <- saBayes(formula = Human ~ Source1 + Source2 + Source3 + Source4 + Source5,
                   time = ~Time, location = ~Location, type = ~Type,
```

```
                data = sim_SA$data, priors = priors,
                alpha_conc = 1, prev = sim_SA$prev,
                likelihood_dist = "pois", n_iter = 1010,
                mcmc_params = list(burn_in = 20, thin = 1))
```

The algorithm is run for 102,000 iterations using the **sourceR** command, with an initial burn in of 2000 iterations, followed by a further 100,000 iterations, of which every 100$^{\text{th}}$ sample is saved. The acceptance rates for all parameters (except those updated using a Gibbs sampler) can be found in a list called `acceptance` in the output from `saBayes`. Trace and autocorrelation plots for the parameters (Figure 2) indicate that the Markov chain is mixing well and has converged, and that thinning by 100 is adequate. The posteriors are returned as nested lists for each parameter. The following R code demonstrates how to access the posteriors for a given source a_{jtl} or type q_i effects and a relative prevalence r_{ijt}.

```
## Plot the marginal posterior for the source effect 2, at time 1, location A
plot(res_sim$posterior$a$time1$locationA[,"Source3"], type="l")
## Plot the marginal posterior for the type effect 21
plot(res_sim$posterior$q[,"type21"], type="l")
## Plot the marginal posterior for the relative prevalence of source effect 5,
## type 17, at time 2
plot(res_sim$posterior$r$time2["type17","Source5",], type="l")
```

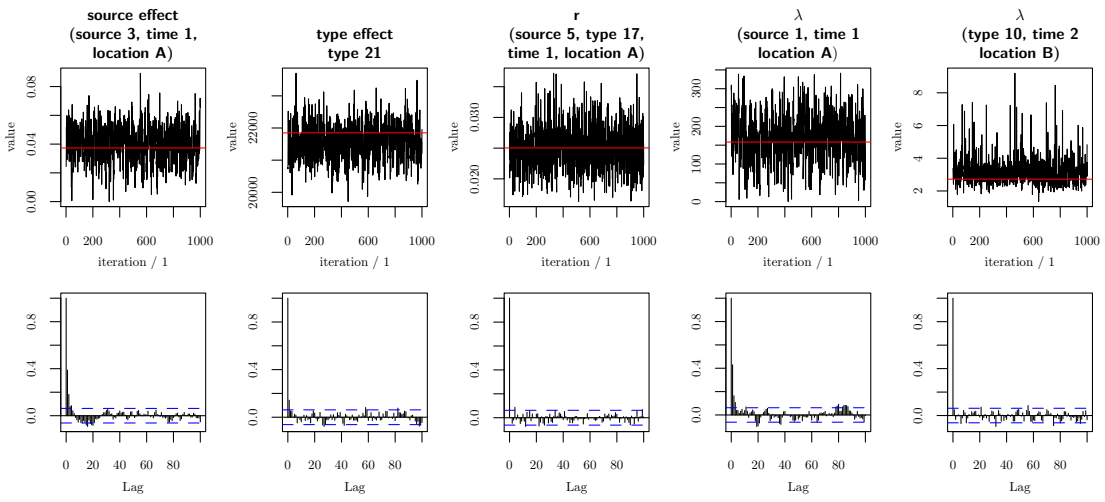

Figure 2: Trace and acf plots for a sample of the model parameters. True values of the parameters are shown in red.

Medians and Chen-Shao highest posterior density credible intervals (Chen and Shao 1991) can be obtained for each parameter using the `summary` command.

```
summary(res_sim, alpha = 0.05, thin = 1, burn_in = 0)
```

The data can be subsetted using the `subset_posterior` command.

```
subset_posterior(res_sim, params = c("a", "li", "q"),
                 t = "1", l = "B", j = c("Source2", "Source1"),
                 i = c("47", "10"), iters = c(3:10))
```

Both the full posterior and a subset of the posterior (generated using `subset_posterior`) can be flattened into a data frame.

```
flatten(res_sim)
```

The marginal density plots of the proportion of cases attributed to each source at each time and location (λ_{jtl}) show that the true values (shown by a cross on the graph) are within the credible intervals (Figure 3). The residual plots for λ_i (Figure 4) show that the model is fitting well. The heatmap shows the grouping of the type effects (Figure 5) computed using a dissimilarity matrix from the clustering output of the mcmc. The coloured bar under the dendrogram gives the correct grouping from the simulated data. This shows that all the types have been grouped correctly if the dendrogram is cut at the true number of groups (5).

5.2. Campylobacteriosis cases in the Manawatu (2005-2008)

In this section, we apply **sourceR** to the `campy` (campylobacteriosis) dataset from Manawatu, New Zealand, described in Section 2. We compare the results of our Bayesian non-parametric approach with results from the Modified Hald and Island models. Types which do not have any source cases need to be removed from the data set before running the analysis because there is no information to attribute human cases to a source if the subtype only occurs in humans.

```
data(campy)
set.seed(59623)
# remove rows with no source cases as there is no information for
# source attribution of human cases for these sources
zero_rows <- which(apply(campy[,c(2 : 7)], 1, sum) == 0)
campy <- campy[-zero_rows,]

# Set priors
priors <- list(a = 1, r = 1, theta = c(0.01, 0.00001))

# set prevalences
# Number of samples  tested  for c. jejuni, for each source.
tot_samples<-c(239, 196, 127, 595, 552, 192 + 332)
# Number of samples positive for c. jejuni, for each source.
```

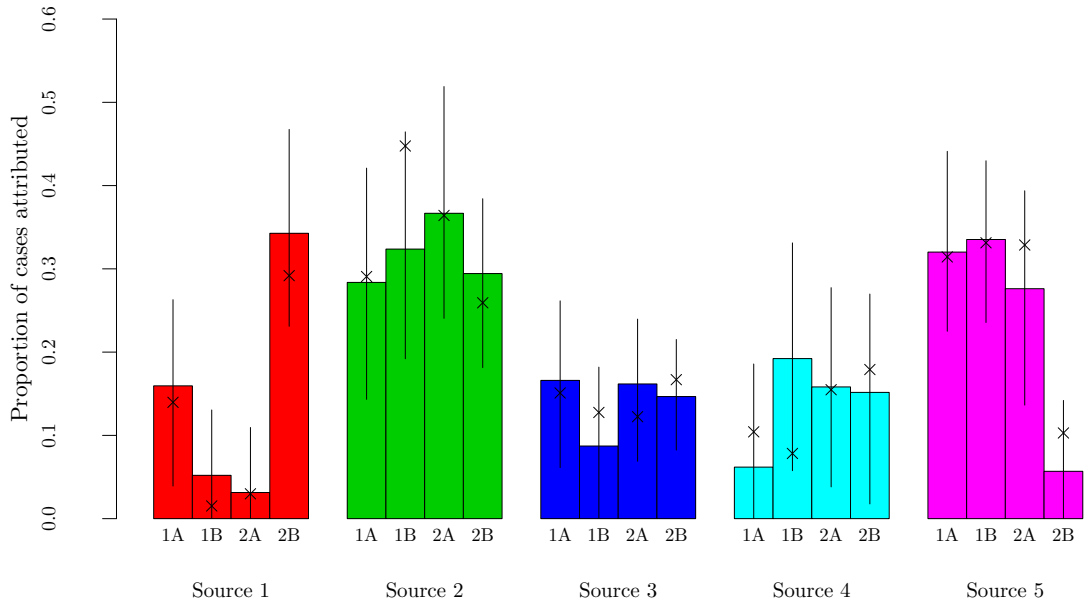

Figure 3: Proportion of cases attributable to each source for each time (1, 2) and location (A, B) for simulated data. Error bars represent 95% Chen-Shao credible intervals. True λ_j values are shown as crosses.

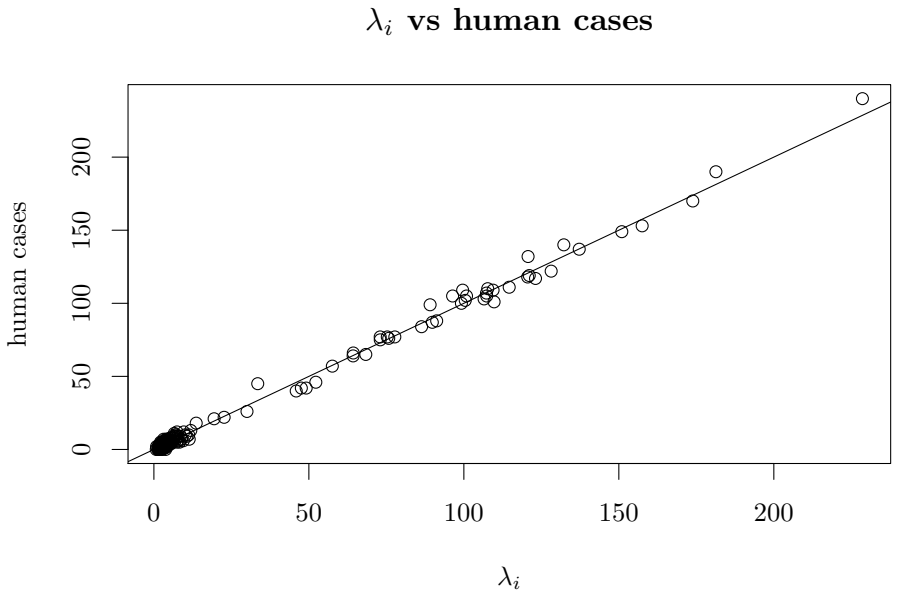

Figure 4: Residual plots (simulated data).

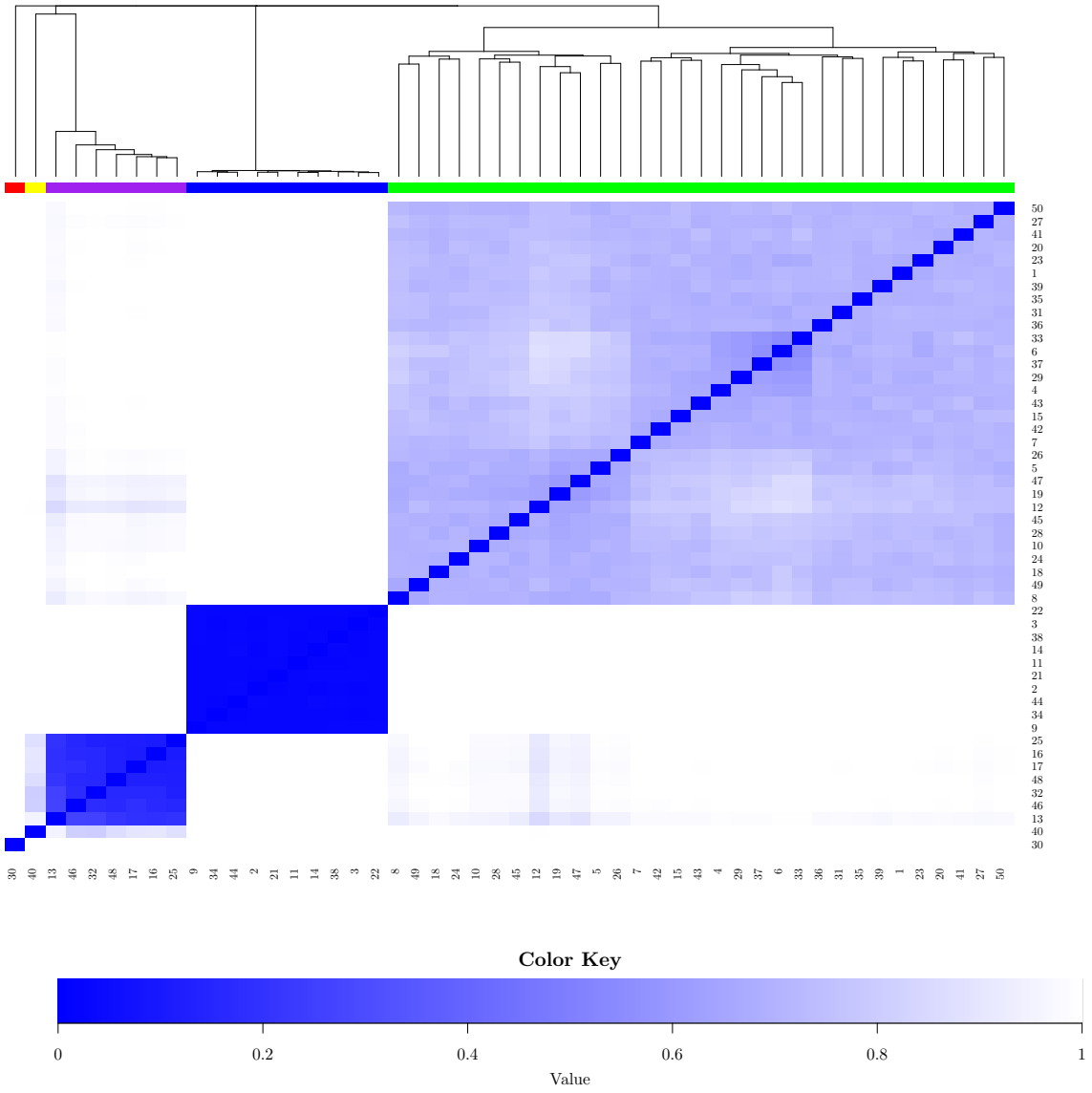

Figure 5: Heatmap showing the grouping of the type effects (q) using simulated data (true groupings given by the 5 colours in the bar under the dendrogram).

```
pos_samples<-c(181, 113, 109, 97, 165, 24 + 62)
prevs <- data.frame(value = pos_samples / tot_samples,
                    source_id = colnames(campy[, 2:7]))

# Run model
# the model assumes one time and location if none are specified in saBayes
res_real <- saBayes(formula = Human ~ ChickenA + ChickenB + ChickenC +
                  Bovine + Ovine + Environment,
                type = ~Type, data = campy, priors = priors, alpha_conc = 1,
                prev = prevs, likelihood_dist = "pois", n_iter = 1020,
                mcmc_params = list(burn_in = 20, thin = 1))
```

Trace and autocorrelation plots for the parameters indicate that the Markov chain is mixing well and has converged, and that thinning by 500 is adequate for most of the parameters (Figure 6). The residual plots for the λ_is (Figure 7) show that the model is fitting the data well. The proportion of cases attributed to each source (λ_j) using the new model can be compared to the previous models (Figure 8). The new model has very similar medians to the modified Hald and Island models. The credible intervals are much narrower than the modified Hald model, but still relatively wide compared to the Island model. The heatmap and dendrogram of the type effects (Figure 9) shows that there are 3 main groups of type effects. The violin plots (Figure 10) show that the largest group of types have very small type effects. These correspond to types that are observed in source samples, but no human cases. There is a group of 5 types which have very large type effects (including type 474 which is endemic to NZ and largely associated with poultry). Although the clustering was determined without reference to genetic relatedness of the types, three members of this group (subtypes 38, 48 and 474) are members of the same clonal complex (CC48) and therefore genetically closely related (pubmlst 2016). Subtype 52 was frequently placed in both the groups with the largest and middling type effects (as can be seen in Figure9, although overall it was attributed to the group with the largest type effects.

6. Discussion

The **sourceR** package is the first implementation of a source attribution model in standard statistical software that is easily accessible and intended for use by epidemiologists. The simulation and case studies illustrate how the **sourceR** package might be used in practice to identify important sources of infection. The new model is widely applicable, fully joint, and does not require approximations or a large number of assumptions. Mixing and aposteriori correlations are significantly decreased in comparison to the Modified Hald model. Furthermore, it can identify clusters of bacterial sub types with similar virulence, pathogenicity and survivability.

6.1. Clustering of the type effects

The clustering has significantly reduced the effective number of parameters in the model. The dendrogram and heatmap indicate that there are three main groups identified by the model.

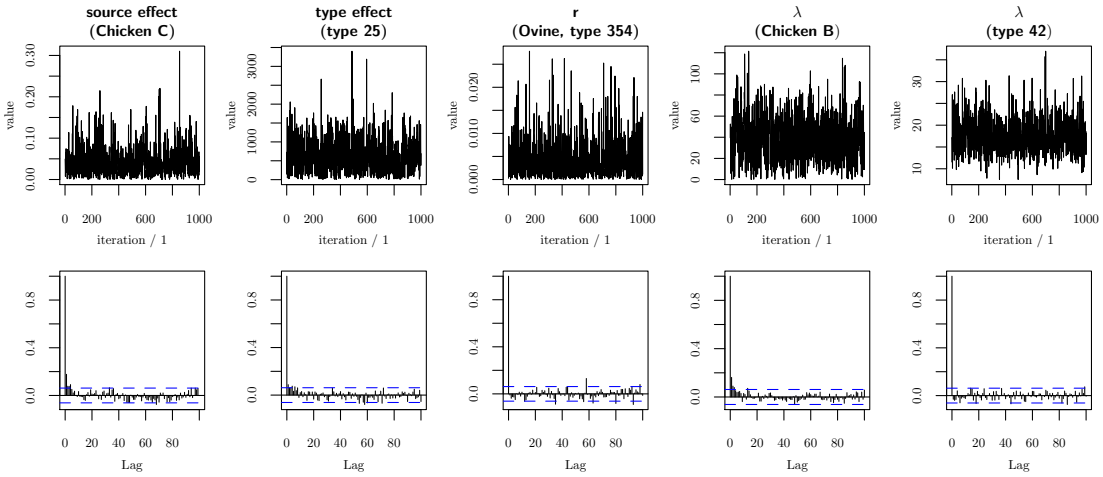

Figure 6: Trace and acf plots for a sample of the model parameters (Manawatu *Campylobacter* data).

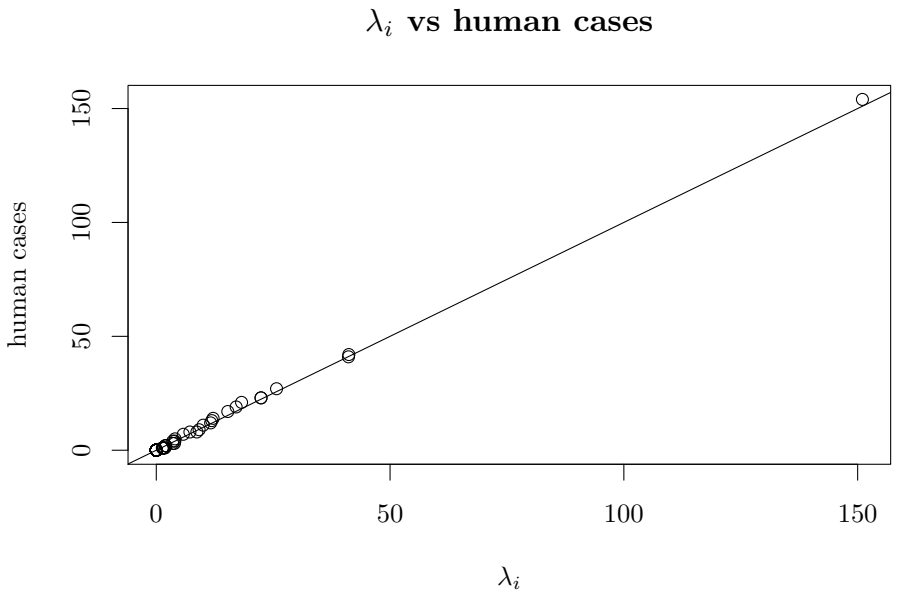

Figure 7: Residual plots (Manawatu *Campylobacter* data).

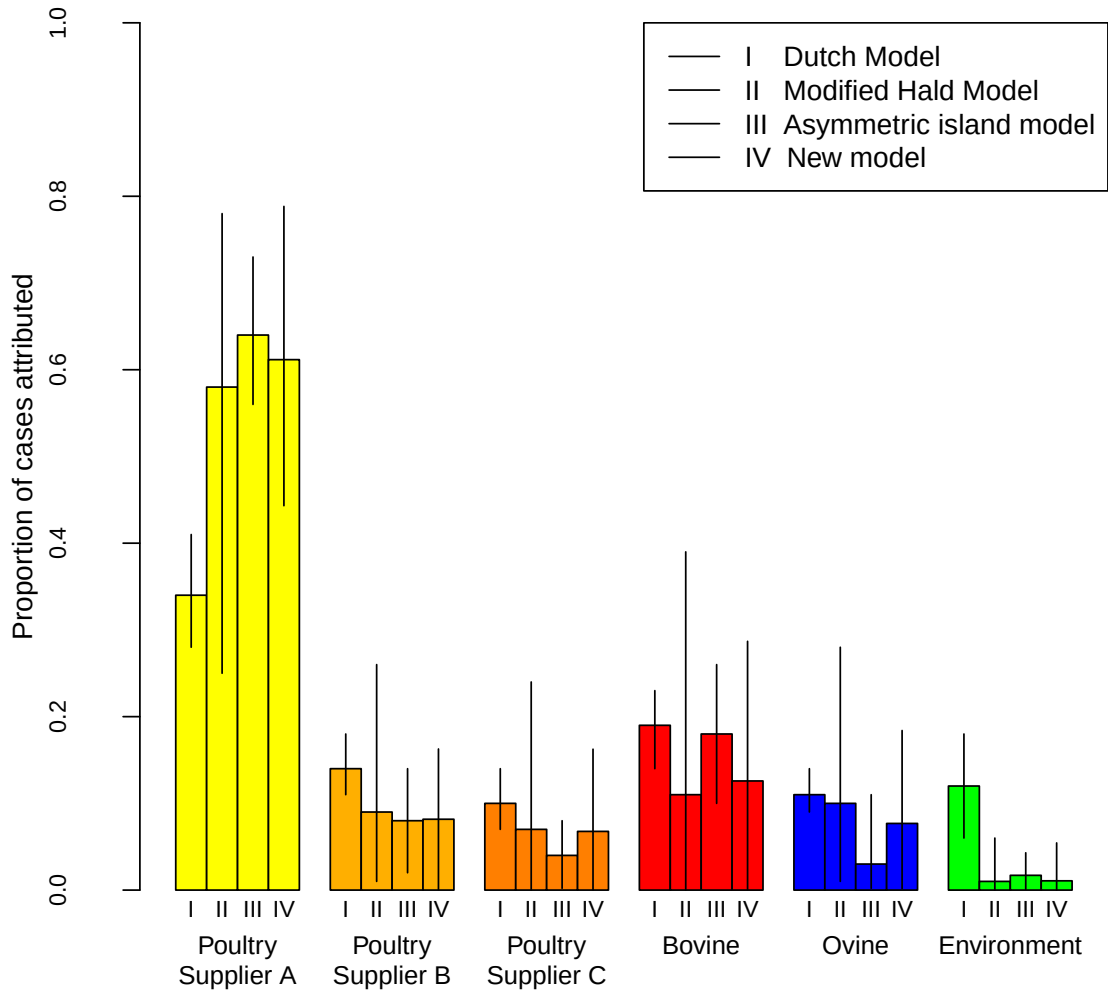

Figure 8: Proportion of human campylobacteriosis cases attributable to each source (Manawatu *Campylobacter* data). Error bars represent 95% confidence or credible intervals.

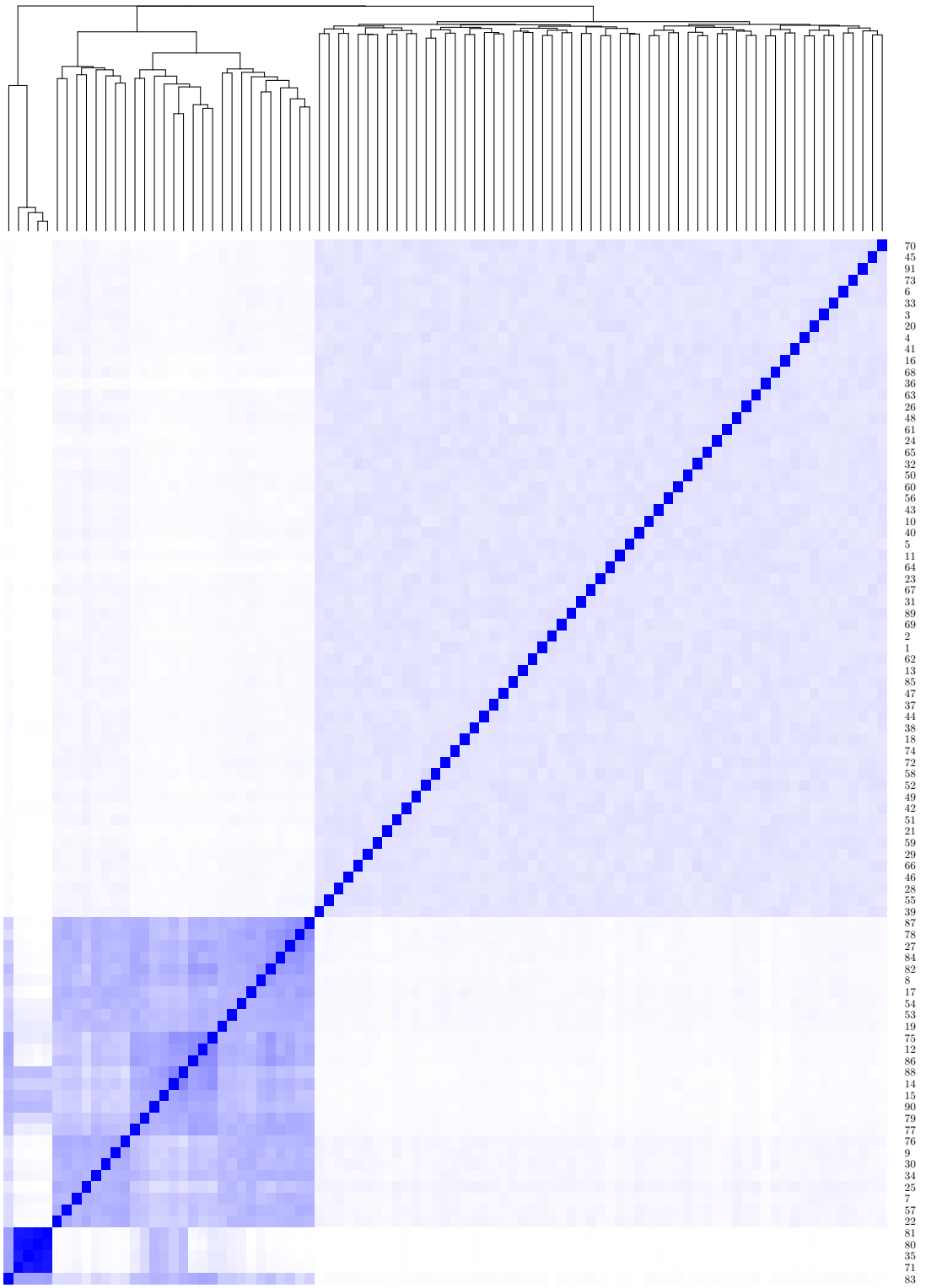

Figure 9: Heatmap showing the grouping of the type effects (q) using the Manawatu *Campylobacter* data.

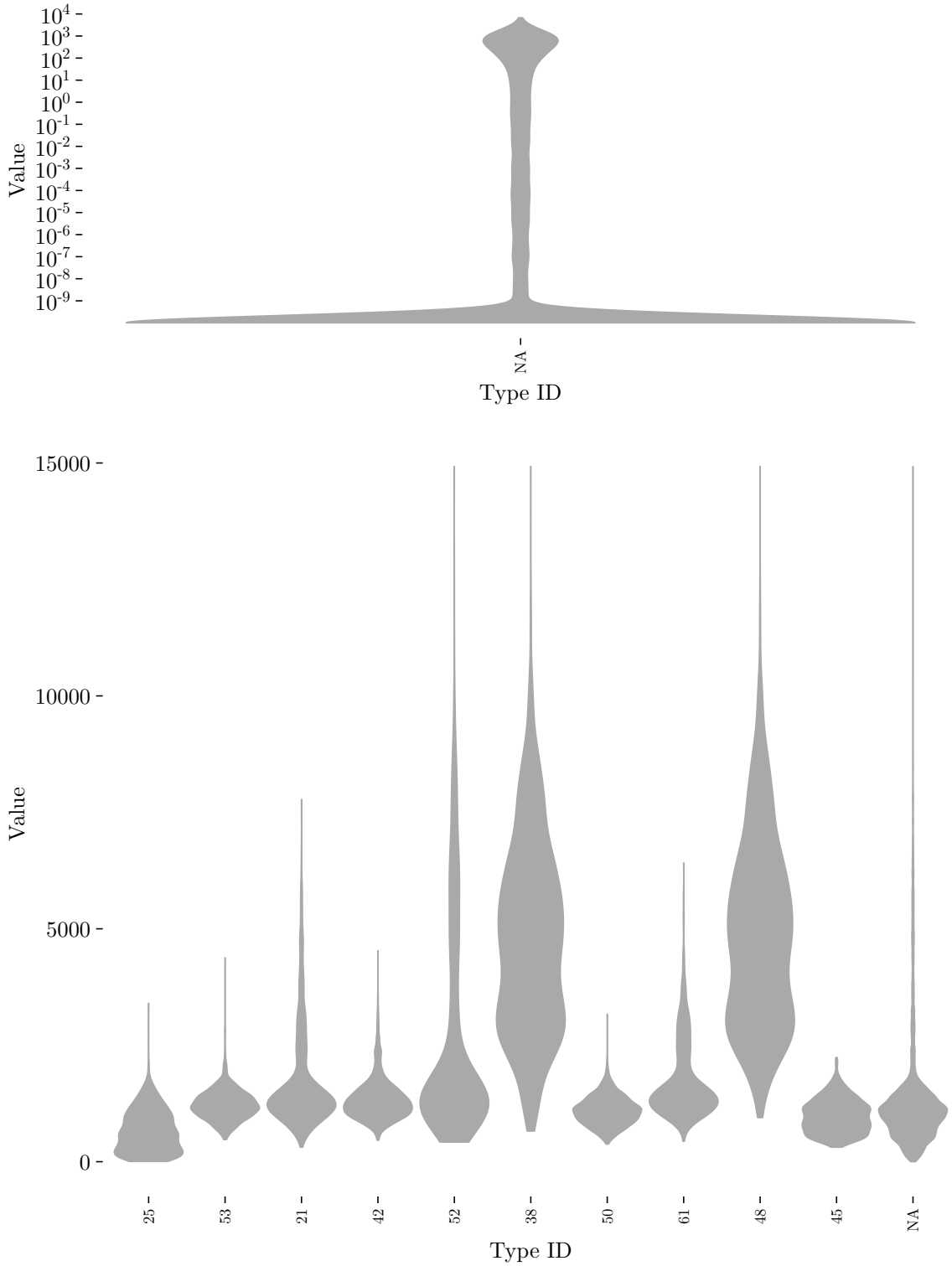

Figure 10: Violin plots of the type effects (q) using the Manawatu *Campylobacter* data.

It is more difficult to estimate the type effect if there are very few positive source samples for that type because large changes in the type effect may not result in much change to the estimated number of human cases. In the *Campylobacter* data set (Mullner *et al.* 2009), there is very little information about the source effects for the types in the largest group identified in the clustering because the source matrix for these types is very sparse, and all have zero human cases. This means that the type effect becomes very small and dominates the source effect.

Care must be taken in performing marginal interpretations of the number of parameters. It is much easier to split a group into two (with similar group means) than it is to merge two groups with clearly different means. Hence, a histogram of the number of groups per iteration is positively skewed compared to the true number of groups. When fitting the model with simulated data, visually assessing the dendrogram and heatmap to determine the number of groups usually provides a closer value to the true number of groups than looking at a histogram, particularly when the group means are well separated.

The results of the clustering of the type effects is of biological interest as it could be used to identify alleles that are correlated with high (or low) virulence, survivability and pathogenicity. The analysis could therefore provide an early warning system for the emergence of dangerous pathogen types in, for example, a particular food processing facility. Additionally, it may identify clusters of strains having particular traits that could be explored using further genotyping or phenotyping assays.

6.2. Posterior correlations between the source and type effects

In the Hald and Modified Hald models there is an inherent posterior correlation between the mean of the source and type effects because the model does not include an explicit mean or constrain the scale of the source or type parameters. This causes a decrease in mixing quality and increases the width of the credible intervals for the source and type effects. This correlation has been greatly reduced in our model by constraining the scale of the source effects using a Dirichlet prior. However, aposteriori correlations between some source and type effects (and hence some λ_j parameters and the source and type effects) may occur if the source matrix is highly unbalanced (especially if it contains many zero's as in the *Campylobacter* data set used above). Although a highly unbalanced source matrix can make fitting the model difficult, a heterogeneous distribution of types is a essential for the model to find the solution with the highest probability of occurrence, as there would be little information contained in the observations of human cases if the types were approximately equally distributed among the food sources (Hald *et al.* 2004). The introduction of uncertainty into a relative prevalence matrix prevents any source-type combination from being 0 which reduces the heterogeneity of the relative prevalence matrix because it forces the larger components to be reduced (as the vector $\mathbf{r}_j$ must sum to 1 over the types for each source). Whether or not to allow true zero's in the prevalence matrix depends on whether there are truly apathogenic types (for a particular source).

The source matrix for the simulated data analysed in Section 5.1 was drawn from a $\mathsf{Uniform}(1, 100)$ distribution which meant the matrix was not sparse, nor highly imbalanced. Simulating data with a sparse, highly unbalanced source matrix reduced mixing quality and increased posterior correlations between some source and type effects. Alternative fitting algorithms such as NUTS (Homan and Gelman 2014) converge to high-dimensional target distributions much

more quickly than simpler methods such as random walk Metropolis or Gibbs sampling (that are currently used in **sourceR**). This is because they avoid the random walk behaviour and sensitivity to correlated parameters that are causing slow mixing for the highly unbalanced *Campylobacter* data set. Currently, Dirichlet processes cannot be fitted using NUTS, hence, a hybrid algorithm, where the clustering is fitted using a standard CRP, and the other parameters are updated using NUTS would likely improve mixing significantly. This may be implemented in a future release for the **sourceR** package.

If there are multiple times and / or locations, it is much easier to identify the type effects groups because they are constant over all times and locations.

6.3. Comparison of the number of cases attributed to each source for current source attribution models

Figure 8 shows the proportion of cases attributed to each source for each of the commonly used source attribution models in addition to the new model. The median values are similar between all models except the Dutch method. The credible intervals of the Dutch model are very narrow because there are far fewer parameters in the model, however, the lack of source and type effects in the model biases the results.

The Island model has much narrower credible intervals than the other models, however it is much more complex than the other models, and hence it has many implicit assumptions (such as the assumptions about mutation, recombination and migration rates, which are likely to be gross simplifications). The narrower credible intervals produced by the Island model could be due to more bias (if the model assumptions are not correct) or more accuracy (due to the additional genetic information that is used). Our model (as with the Hald and Modified Hald models) ignores the genetic similarities between types, which loses some information and prevents attribution of novel types in human isolates to a likely source. However, the Island models reliance on detailed genetic data prevents the use of data where phenotypic typing methods were used, reducing the range of data for which the model is applicable.

The modified Hald model has very wide credible intervals compared to the other models. This may be because the prevalence matrix is less restricted (as it is modelled using independent Beta's for each ij), that uncertainty is modelled for the source prevalences (π_j) or that the model is less identifiable due to the effective number of parameters still being large. Although it would be preferable to allow uncertainty in the source prevalences, we decided to use point estimates in our model. This is because they have the same functional form as the source effects in the model, hence they cannot be identified from the source effects in a fully joint model without very strong priors.

6.4. Source and type effect interpretation

The interpretation of source and type effects for this model depends on the quality and type of data collected, the model specification, and the characteristics of the organism of interest. Type effects summarise the characteristics that determine a types capacity to cause an infection, such as survivability during food processing, pathogenicity or virulence (measured in cases per dose of bacteria population). Source effects account for the ability of a particular source to act as a vehicle of infection. This includes factors such as the amount of the food source consumed (if an offset for consumption is not used), the physical properties of the source and the environment provided for the bacteria through storage and preparation. A

high source effect may reflect a high exposure, but not necessarily a high ability of the individual food source to cause disease. Including an environmental source in the model can be thought of as grouping the (individually) unmeasured wildlife sources into one. It may also be a transmission pathway for pathogens present in livestock sources (for example, through the contamination of waterways) which complicates the interpretation meaning the source effects no longer directly summarize the ability of the source to act as a vehicle for food-borne infections (Hald *et al.* 2004).

6.5. Apathogenic subtypes

Potentially pathogenic types (that is types found in the sources but not humans) are included in the model as it is assumed that these types are rarely (rather than never) found in humans. The model cannot attribute types that have been detected in humans but not in any of the sources because there is no information relating them to the sources (as with the Hald and modified Hald models). The Island model (Wilson *et al.* 2008) can attribute types undetected in a source using inferences on genetic relatedness, however, it cannot use data where types are distinguished by phenotypic characteristics. In addition to excluding human cases for types not detected in any sources, cases with a history of travel in the incubation period are assumed to have acquired the disease overseas, and are therefore excluded from the model.

At present it is assumed that both humans and all sources can potentially be infected by all types, albeit some very rarely. If a type is truly apathogenic in humans, then this approach is likely to overestimate the type incidence λ_i. A future development may therefore be to to allow for zero inflation in the prevalence matrices and human data. However, the **sourceR** package currently allows the relative prevalence matrix to be fixed at the maximum likelihood estimates, which includes zero values where a particular type was not detected in any samples from a source. Fixing the relative prevalence matrix increases the posterior precision, but the results may be biased if the source data is not of a high quality. The relative prevalence matrix can be fixed by setting `r` to `TRUE` in the `params_fix` argument to `saBayes`.

6.6. Model extensions

There are many alternative model extensions to those implemented in the **sourceR** package. These include:

1. Adding in an independent time and/or location term
2. Adding time and/or location dependence to the type factors q_i
3. Model autocorrelation between parameters over time
4. Add interaction terms between the source and type effects

Extension 1 is useful for modelling dynamic behaviour, however, it assumes that the changes are independent of sources and types. It is more likely that the changes are specific to a few sources or types. Hence, it is preferable to add the dependence into the source and/or type terms so that is is possible to identify which parts of the epidemiology are likely to be the cause of the observed changes in the attribution. This model is a subset of the other

models where the time/ location dependent behaviour is independent of the source, type and prevalence.

Extension 2 involves adding temporal or location dependence to the type effects. Type effects do not change in each location as they depend on the genetics of each bacterial subtype. Evolution of subtypes over time could cause changes to their virulence, pathogenicity and survivability (and hence type effect). There is evidence that *Campylobacter* can evolve quickly Wilson, Gabriel, Leatherbarrow, Cheesbrough, Gee, Bolton, Fox, Hart, Diggle, and Fearnhead (2009), however, assuming the type effects are fixed over time is equivalent to assuming that the types are likely adapted to a particular source, and that any further adaptation to a new source is likely to coincide with a change in biology, and hence, the introduction of a new sequence type (French and Marshall 2009). Changes to the type effects over time (or location) are likely to have a much smaller impact on the source attribution than changes to the source effects because the source factor applies to all subtypes on a given source (and there are many more types than sources). At present, the package does not support type effects changing over time, however, this is a feature that may be implemented in a further release.

Future releases of the package will also allow the user to independently specify whether the source, type and prevalence parameters are time or location dependent. Currently, if the human cases are modelled with time and location information, the source effects must also vary over the same times and locations, whilst the relative prevalence matrix must vary over only the times. For example, it may be preferable to use a single prevalence matrix (if subsetting the matrix over time makes it too sparse, or if the source data was collected at different times to the human data), but allow the human cases and source effects to vary over time.

Extension 3: another improvement would be to allow autocorrelation between the parameters over time, rather than modelling them as separable. An AR(1) model has been used in NZ attribution studies via modifications to the asymmetric island model (French and Marshall 2015).

Extension 4 involves adding interaction terms between the source and type effects to the model to allow for the biologically plausible possibility that certain subtypes are more or less likely to survive and cause disease, dependent on the food source they appear in. However, this would significantly increase the number of parameters and reduce identifiability of the model.

7. Conclusions

In this article, we have presented a novel source attribution model which builds upon, and unites, the Hald and Modified Hald approaches. This model allows the data to inform type effect clustering using a Bayesian non-parametric model. This is a significant improvement over the previous attempts to improve model identifiability by reducing the effective number of parameters (fixing some source and type effects, or modelling the type effects as random using a 2 stage model). Like the Modified Hald model, the new model incorporates uncertainty in the prevalence matrix into the model, however, it does this by fitting a fully joint model rather

than a 2 step model. This has the advantage of allowing the human cases to influence the uncertainty in the source cases and preserves the restriction on the sum of the prevalences for each source. The **sourceR** package implements this flexible Bayesian non-parametric model to enable straightforward attribution of cases of zoonotic infection to putative sources of infection by epidemiologists and other scientists.

Acknowledgements

The research for this paper was financially supported by the Ministry for Primary Industries, the Institute of Fundamental Sciences (Massey University), the mEpiLab (Massey University), and CHICAS (Lancaster University). This project was also supported by the Livestock Improvement Corporation (LIC) and Soroptimist International Palmerston North. We acknowledge the following individuals and groups: mEpiLab (Massey University), MidCentral Public Health Services and Petra Mullner (for the Manawatu data set) and Geoff Jones (for his helpful input on automatic clustering methods).

References

Allos BM, Moore MR, Griffin PM, Tauxe RV (2004). "Surveillance for Sporadic Foodborne Disease in the 21st Century: The FoodNet Perspective." *Clinical Infectious Diseases*, **38**(Supplement 3), S115–S120. doi:10.1086/381577. http://cid.oxfordjournals.org/content/38/Supplement_3/S115.full.pdf+html, URL http://cid.oxfordjournals.org/content/38/Supplement_3/S115.short.

Baker M, Wilson R, Ikram R, Chambers S, Shoemack S, Cook G (2006). "Regulation of Chicken Contamination Urgently Needed to Control New Zealand's Serious Campylobacteriosis Epidemic." *The New Zealand Medical Journal*.

Chen M, Shao Q (1991). "Monte Carlo Estimation of Bayesian Credible and HPD Intervals." *Journal of Computational and Graphical Statistics*.

Crump JA, Griffin PM, Angulo FJ (2002). "Bacterial Contamination of Animal Feed and Its Relationship to Human Foodborne Illness." *Clinical Infectious Diseases*, **35**(7), 859–865. doi:10.1086/342885. http://cid.oxfordjournals.org/content/35/7/859.full.pdf+html, URL http://cid.oxfordjournals.org/content/35/7/859.abstract.

Dingle K, Colles F, Wareing D, Ure R, Fox A, Bolton F, Bootsma H, Willems R, Urwin R, Maiden M (2001). "Multilocus sequence typing system for Campylobacter jejuni." *Journal of Clinical Microbiology*.

Ferguson T (1973). "Bayesian Analysis of some Nonparametric Problems." *Ann. Stat.*, **1**, 209–230.

French N, Marshall J (2009). "Dynamic Modelling of Campylobacter Sources in the Manawatu." *Technical report*, Hopkirk Institute, Massey University. Prepared for Dr Donald Campbell, New Zealand Food Safety Authority.

French N, Marshall J (2013). "Completion of Sequence Typing of Human and Poultry Isolates and Source Attribution Modelling." *Technical report*, Hopkirk Institute, Massey University.

French N, Marshall J (2015). "Final Report: MPI Agreement 11777, Schedule 1A Source Attribution January to December 2014 of Human *Campylobacter jejuni* Cases from the Manawatu." *Technical report*, Molecular Epidemiology and Public Health Laboratory Infectious Disease Research Centre Institute of Veterinary, Animal and Biomedical Sciences College of Sciences Massey University New Zealand.

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science.

Hald T, Vose D, Wegener H, Koupeev T (2004). "A Bayesian Approach to Quantify the Contribution of Animal-Food Sources to Human Salmonellosis." *Risk Analysis*, **24**(1), 255–269.

Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, Praet N, Bellinger DC, de Silva NR, Gargouri N, Speybroeck N, Cawthorne A, Mathers C, Stein C, Angulo FJ, Devleesschauwer B, on behalf of World Health Organization Foodborne Disease Burden Epidemiology Reference Group (2015). "World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010." *PLoS Med*, **12**(12), 1–23. doi:10.1371/journal.pmed.1001923. URL http://dx.doi.org/10.1371%2Fjournal.pmed.1001923.

Homan MD, Gelman A (2014). "The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*

Miller W, On S, Wang G, Fontanoz S, Lastovica A, Mandrell R (2005). "Extended Multilocus Sequence Typing System for Campylobacter coli, C. lari, C. upsaliensis, and C. helveticus." *Journal of Clinical Microbiology*.

Mullner P, Collins-Emerson J, Midwinter A, Carter P, Spencer S, van der Logt P, Hathaway S, French N (2010). "Molecular Epidemiology of Campylobacter jejuni in a Geographically Isolated Country with a Uniquely Structured Poultry Industry." *Applied and Environmental Microbiology*, **76**(7), 2145–2154.

Mullner P, Jones G, Noble A, Spencer S, Hathaway S, French N (2009). "Source Attribution of Food Borne Zoonoses in New Zealand: A Modified Hald Model." *Risk Analysis*, **29**(7).

pubmlst (2016). "Campylobacter MLST." This publication made use of the Campylobacter Multi Locus Sequence Typing website (http://pubmlst.org/ campylobacter/) sited at the University of Oxford (Jolley & Maiden 2010, BMC Bioinformatics, 11:595). The development of this site has been funded by the Wellcome Trust., URL http://pubmlst.org/campylobacter/.

van Pelt W, van de Giessen A, van Leeuwen W, Wannet W, Henken A, Evers E (1999). "Oorsprong, Omvang en Kosten van Humane Salmonellose. Deel1. Oorsprong van Humane Salmonellose met Betrekking tot Varken, Rund, Kip, ei en Overige Bronnen." *Infectieziekten Bull.*

Wilson D (2016). "iSource." URL http://www.danielwilson.me.uk/iSource.html.

Wilson D, Gabriel E, Leatherbarrow A, Cheesebrough J, Hart C, Diggle P (2008). "Tracing the Source of Campylobacteriosis." *PLoS Genetics*.

Wilson DJ, Gabriel E, Leatherbarrow AJ, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P (2009). "Rapid Evolution and the Importance of Recombination to the Gastroenteric Pathogen Campylobacter jejuni." *Molecular Biology and Evolution*, **26**(2), 385–397.

A. McMC Algorithm

This section gives the full details of the algorithm used to fit the fully joint non-parametric source attribution model described in Section 4.

Affiliation:

Poppy Miller
CHICAS, Faculty of Health and Medicine,
Furness College,
Lancaster University,
Lancaster, LA1 4YG,
United Kingdom
E-mail: p.miller@lancaster.ac.uk

Algorithm 1 Chinese restaurant algorithm to update the type effects.

Initialise: Setup output matrices and initial values

for z in 1:*niter* **do**

Step 1 : Update source effects (**a**) for each time t and location l (adaptive single site Normal random walk Metropolis-Hastings).

Propose $a_j^* \sim \mathsf{Normal}\,(a_j, \sigma)$ where σ equals Σ_a w.p. 0.95 and σ_a otherwise.
if z $\mathsf{mod}(50) == 1$ **then**
Update Σ_a: $\Sigma_a^* = \Sigma_a + \mathsf{sign}\,(\mathsf{current\ acceptance\ rate}\ a_j - 0.45) \times \mathsf{min}\left(0.01, z^{-0.5}\right)$
end if
Rescale a_{-j} such that $\sum_{j=1}^{m} a_j = 1$
Accept proposed a_j^* w.p. $\dfrac{\prod_{i=1}^{n}\left(q_{k(i)} \sum_{j=1}^{m} a_j^* p_{ij}\right)^{y_i} e^{-q_{k(i)} \sum_{j=1}^{m} a_j^* p_{ij}}}{\prod_{i=1}^{n}\left(q_{k(i)} \sum_{j=1}^{m} a_j p_{ij}\right)^{y_i} e^{-q_{k(i)} \sum_{j=1}^{m} a_j p_{ij}}} \times \dfrac{\prod_{j=1}^{m} a_j^{*\alpha_a - 1}}{\prod_{j=1}^{m} a_j^{\alpha_a - 1}}$

Step 2 : Update components of the relative prevalence matrices (r_{ij}) for each time t (adaptive single site Normal random walk Metropolis-Hastings).

Propose $r_{ij}^* \sim \mathsf{Normal}\,(r_{ij}, \sigma)$ where σ equals Σ_r w.p. 0.95 and σ_r otherwise.
if z $\mathsf{mod}(50) == 1$ **then**
Update Σ_r: $\Sigma_r^* = \Sigma_r + \mathsf{sign}\,(\mathsf{current\ acceptance\ rate}\ r_{ij} - 0.45) \times \mathsf{min}\left(0.01, z^{-0.5}\right)$
end if
Rescale r_{-ij} such that $\sum_{i=1}^{n} r_{ij} = 1$
Accept proposed r_{ij}^* w.p. $\dfrac{\prod_{i=1}^{n}\left(q_{k(i)} \sum_{j=1}^{m} a_j r_{ij}^* \pi_j\right)^{y_i} e^{-q_{k(i)} \sum_{j=1}^{m} a_j r_{ij}^* \pi_j}}{\prod_{i=1}^{n}\left(q_{k(i)} \sum_{j=1}^{m} a_j r_{ij} \pi_j\right)^{y_i} e^{-q_{k(i)} \sum_{j=1}^{m} a_j r_{ij} \pi_j}} \times \dfrac{\prod_{i=1}^{n}\left(r_{ij}^{*\left(x_{ij}+\alpha_r-1\right)}\right)}{\prod_{i=1}^{n}\left(r_{ij}^{\left(x_{ij}+\alpha_r-1\right)}\right)}$

Step 3 : Update type effects (**q**) using a blocked Gibbs sampler (chinese restaurant construction).
See Algorithm **??** for details.

end for