Source attribution modelling

Jonathan Marshall

11 June 2013

Collaborators

Petra Müllner
Geoff Jones
Alasdair Noble
Martin Hazelton
Simon Spencer
Michael Baker
Donald Campbell
Phil Carter
Tui Shadbolt
Nigel French
Hopkirk lab team
NZFSA for funding

Source Attribution

There are many potential sources of infection and pathways through which that infection may occur.

Source attribution is the process of determining which proportion of a particular disease is acquired from a given source and through a given pathway.

Knowledge of which sources are contributing the most to the disease burden allows targeting of intervention strategies.

How can we determine the origin of a human case?

- Often cases are sporadic rather than being part of outbreaks (e.g. *Campylobacter*).
- Epidemiological information associated with a case may be minimal.
- We often have no information on the exposure for sporadic cases.
- Often the only thing we have is genotype information for each case.

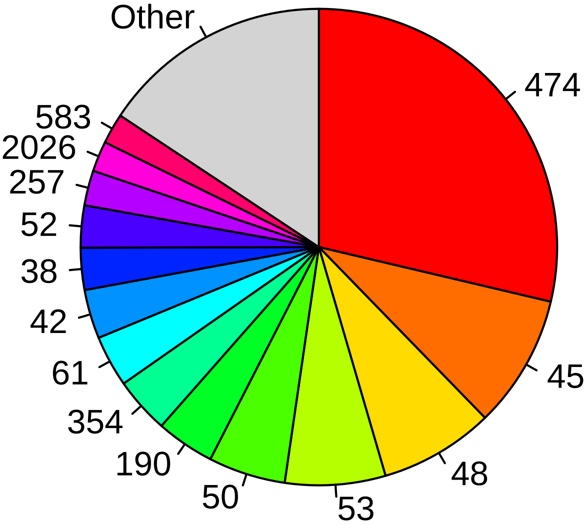
MLST for *Campylobacter*

Seven housekeeping genes around the genome (loci).

The combination of alleles at the seven loci gives the **multilocus sequence type**.

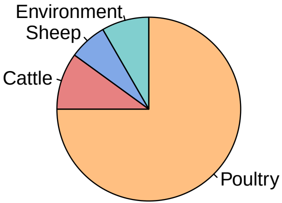
ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
474	2	4	1	2	2	1	5
61	1	4	2	2	6	3	17
190	2	1	5	3	2	3	5
2381	175	251	216	282	359	293	102
48	2	4	1	2	7	1	5

Human MLST types

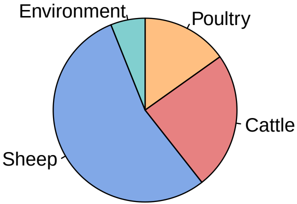


Source specific types

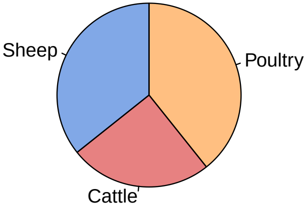
ST-474: $n = 60$



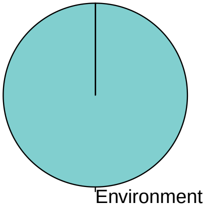
ST-61: $n = 33$



ST-190: $n = 28$



ST-2381: $n = 22$



Attribution using MLST

For each human ST, assign to the most likely source, given the distribution of STs on the source.

Use **Bayes Theorem**

$$P(\text{source} = k | \text{ST} = i) = \frac{P(\text{ST} = i | \text{source} = k) P(\text{source} = k)}{\sum_k P(\text{ST} = i | \text{source} = k) P(\text{source} = k)}$$

where

- $P(\text{ST} = i | \text{source} = k)$ is the distribution of STs on each source.
- $P(\text{source} = k)$ is the prior probability that an isolate picked at random is from source k.

The Dutch model

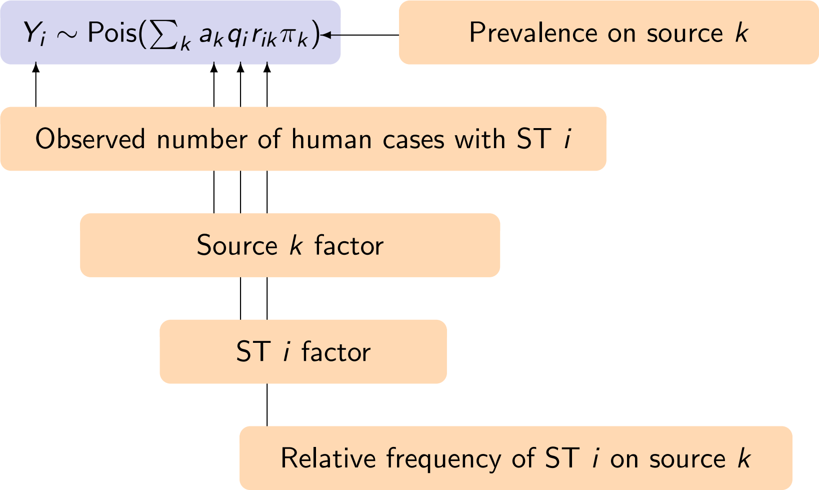
- The simplest way to estimate $P(\mathrm{ST} = i | \mathrm{source} = k)$ is to use the relative frequency r_{ik} of ST i on source k.

- The simplest way to estimate $P(\mathrm{source} = k)$ is to assume *apriori* that all sources are equally likely.

- This yields the Dutch model:

$$P(\mathrm{source} = k | \mathrm{ST} = i) = \frac{r_{ik}}{\sum_k r_{ik}}$$

Hald model (Hald et. al. 2004, Müllner et. al. 2009)



$Y_i \sim \text{Pois}(\sum_k a_k q_i r_{ik} \pi_k)$

Prevalence on source k

Observed number of human cases with ST i

Source k factor

ST i factor

Relative frequency of ST i on source k

Hald model

- Estimates the source and isolate-specific parameters a_k and q_i by matching up the expected number of human cases of type i with the observed number of cases.

- We have an identifiability problem, in that we have $I + K$ parameters to estimate, and only I data points (number of human isolates of each type).

- Add in some prior information to reduce this problem and fit in a Bayesian framework.

Island model (D. Wilson, 2008)

Model the sampling distribution of the allelic profiles on the sources by assuming that the observed sequences arise due to:

- **Mutation**, where an allele at a locus is novel.
- **Migration** between sources, where the allelic profile has been observed before in one of the sources, including the current one.
- **Recombination**, where the allele at a given loci has been observed before but not in this allelic profile.

Island model

Let M_{jk} is the probability of sampling an allele from source k that has already been observed in source j.

Let f^l_{aj} be the frequency with which allele a has been observed at locus l in those genotypes sampled from source j.

Then, in the absense of mutation and recombination, the probability of sampling allele a at locus l in source k is

$$B^l_{ak} = \sum_{j=1}^{K} M_{jk} f^l_{aj}$$

Island model: Unlinked loci

Let μ_k be the probability of sampling an allele from source k that is a novel mutant (among all sources).

Then if we assume the loci are independent, the sampling formula for observing sequence $y = (y^1, y^2, \dots, y^7)$ on source k is

$$\phi(y|k) = \prod_{l=1}^{7} \begin{cases} \mu_k & \text{if } y^l \text{ is novel,} \\ (1-\mu_k) B^l_{y^l k} & \text{otherwise.} \end{cases}$$

Island model: Linked loci

In the absense of mutation, recombination and migration, we assume that a genotype y sampled from source k will be identical to one already observed in the sample of source k.

Migration allows y to be a copy of a genotype c from a source C other than k.

Mutation and recombination mean that y may contain novel alleles, or comprise of a novel combination of existing alleles.

Island model: Linked loci

Let μ_k be the probability, per locus, that a genotype sampled from source k contains a novel mutant allele.

Let R_k be the probability, per locus, that a genotype sampled from source k has undergone recombination. The allele is independently sampled from other alleles observed at that locus.

The sampling formula for observing sequence $y = (y^1, y^2, \dots, y^7)$ in source k is

$$\phi(y|k) = \sum_c \frac{M_{Ck}}{N_C} \prod_{l=1}^{7} \begin{cases} \mu_k & \text{if } y^l \text{ is novel,} \\ (1-\mu_k) R_k B^l_{y^l k} & \text{if } y^l \neq c^l \\ (1-\mu_k)\left[1 - R_k(1 - B^l_{y^l k})\right] & \text{if } y^l = c^l \end{cases}$$

Island model: Linked loci

ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
474	2	4	1	2	2	1	5
?	2	4	1	2	29	1	5

We have a novel allele at the pgm locus. We assume this genotype has arisen through **mutation**.

Island model: Linked Loci

ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
474	2	4	1	2	2	1	5
?	2	4	1	2	1	1	5

The pgm allele looks familiar...

ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
45	4	7	10	4	1	7	1
3718	2	4	1	4	1	1	5

But we haven't seen this genotype before. We assume it arose through **recombination**.

Island model: Linked loci

ST	aspA	glnA	gltA	glyA	pgm	tkt	uncA
474	2	4	1	2	2	1	5
?	2	4	1	2	2	1	5

This is just 474 - we've seen this before, but possibly not in this source. We assume it arose through **migration**.

Island model: Likelihood of human sequences

Let F_k be the proportion of human sequences h_i from source k.

The posterior distribution of F, given the probabilities μ, M and R is

$$p(F|h,\mu,M,R) \propto \prod_i \left[\sum_k F_k \phi(h_i|k)\right] p(F)$$

where $p(F)$ is the prior distribution, where we assume each source is equally likely.

Island model: Parameter estimation

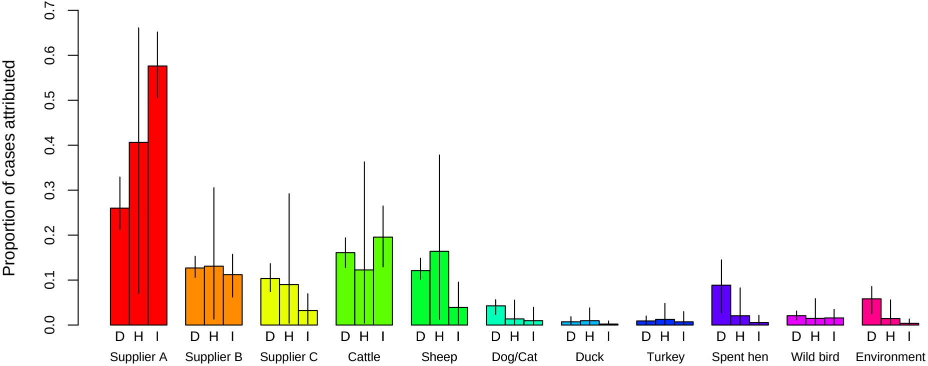
The parameters are estimated in two passes:

- The probabilities μ, M, and R are estimated from the known source cases with likelihood approximated by a leave-one-out approach.

- For each posterior sample of these, an MCMC side-chain is run to estimate the probabilities F_k.

- Within the MCMC, various Metropolis-Hastings updates are used.

Dutch, Hald, Island attribution



Temporal associations: Attribution through time

- Extend existing source attribution models to be dynamic in time.
- Assume mutation, recombination, migration rates are static, as are the distribution of STs on sources.
  - Due to practicality (lack of data) - may be extended later.
- Assume the probability of sources F_j may change through time.

Temporal model for F_j

Assume F_k arises from a linear model on the logit scale.

$$F_{kt} = \begin{cases} \dfrac{e^{-f_{kt}}}{1 + \sum_{k=1}^{K-1} e^{-f_{kt}}} & k = 1 \dots K-1, \\ \dfrac{1}{1 + \sum_{k=1}^{K-1} e^{-f_{kt}}} & k = K. \end{cases}$$

where

$$\begin{aligned} f_{kt} &= X_t \beta_k + e_{kt}, \\ e_{kt} &\sim \text{Normal}(\rho_k e_{k(t-1)}, \sigma_k^2). \end{aligned}$$

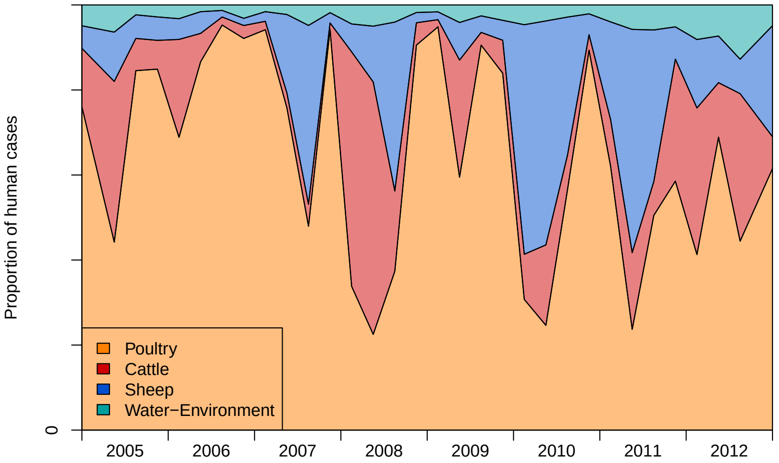
- X is the design matrix, specifying covariates in time.
- β_k measures the effect of those covariates on F_k.
- ρ_k is an auto-correlation AR(1) parameter.
- σ_k^2 is the variance.

Temporal model for F_j

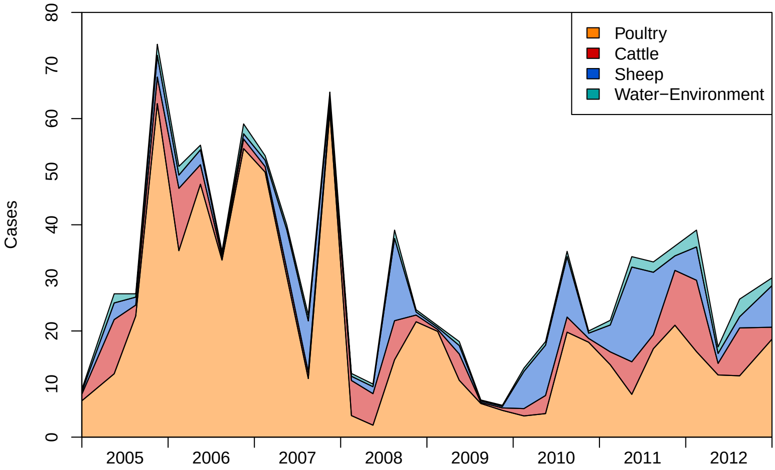
Start simple:

- Assume $X_t = 1$ for all t.
- Thus, the model represents potentially different means within each group.
- Assume also that $\sigma_k^2 = \sigma^2$ for all k.
- Let t be quarters from 2005-2012.

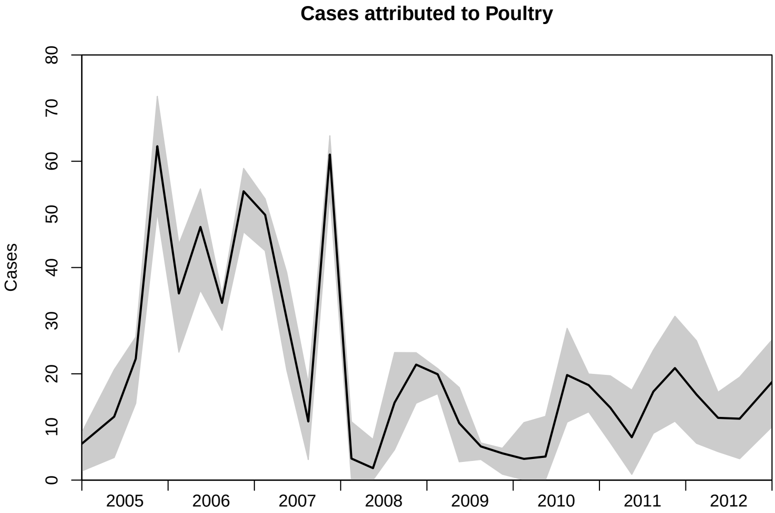
Campylobacter in the Manawatu



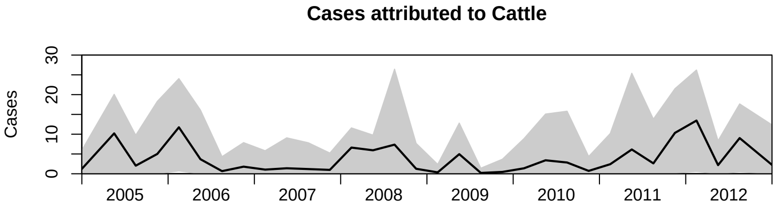
Campylobacter in the Manawatu

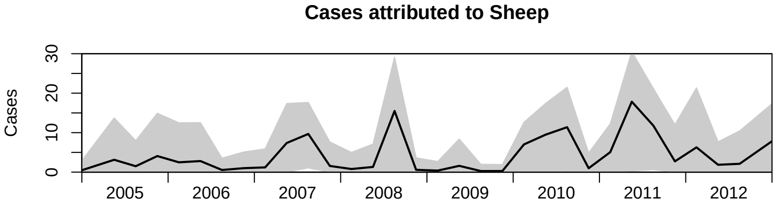


Campylobacter in the Manawatu

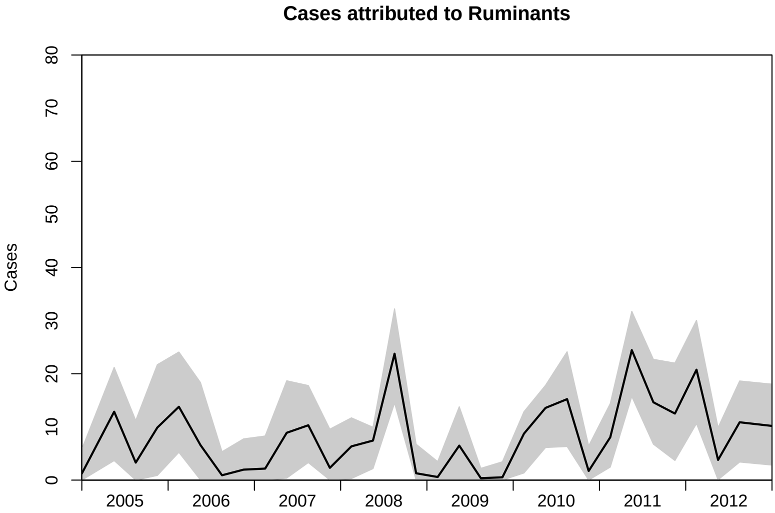


Campylobacter in the Manawatu

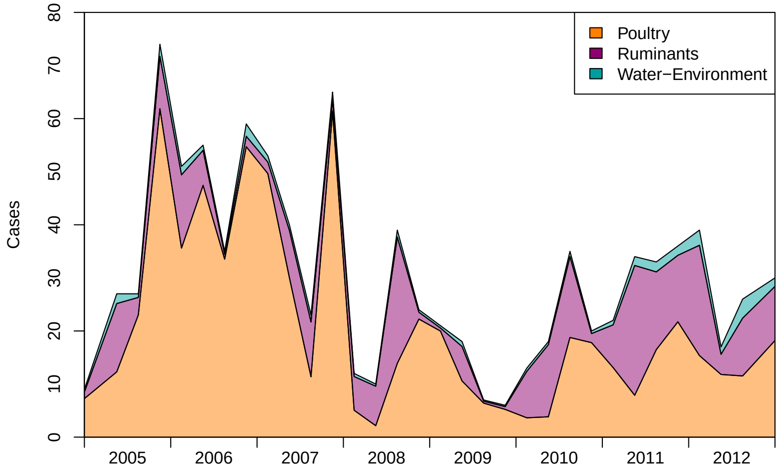




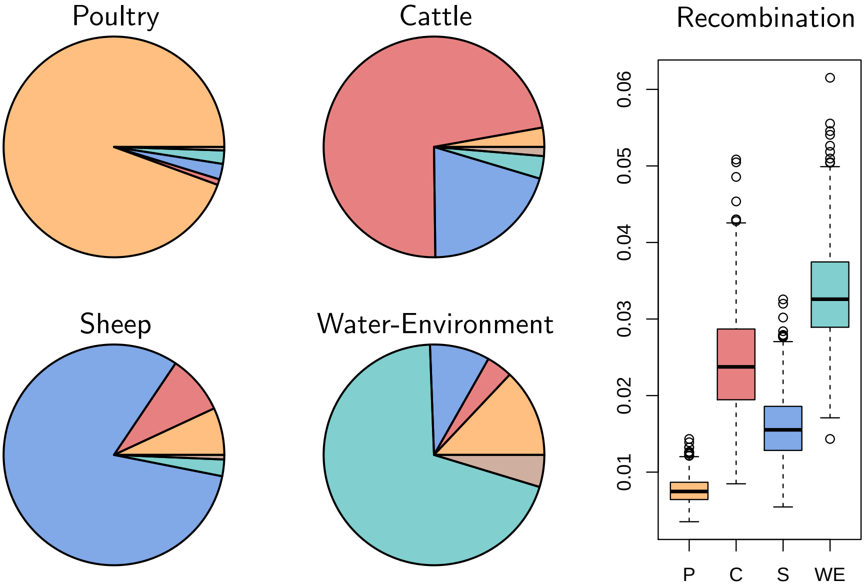
Campylobacter in the Manawatu



Campylobacter in the Manawatu



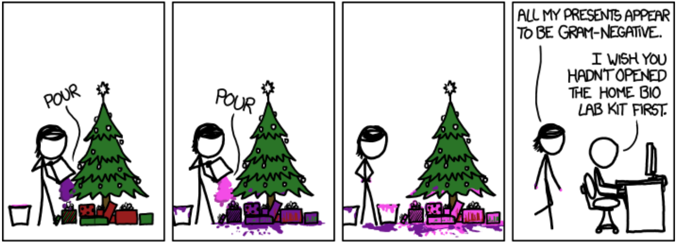
Migration, mutation, and recombination



Future work

- Improve time series model. e.g. investigate seasonality.
- Investigate the effect of intervention in poultry.
- Investigate temporal changes in ST distribution on sources.

Thanks for listening


POUR
POUR
ALL MY PRESENTS APPEAR TO BE GRAM-NEGATIVE.
I WISH YOU HADN'T OPENED THE HOME BIO LAB KIT FIRST.