Fitting a spatial coalescent model

David Welch, Hongbin Guo, Stéphane Guindon

University of Auckland

9 July 2015

@phydyn #MMEE2015, Paris

Supported by the Marsden Fund of NZ

Phylogenetic problem:

Given sequence data D, want to infer genealogy g and any parameters θ controlling mutation or populations processes.

$$P(g, \theta | D) \propto P(D | g, \theta) P(g | \theta) P(\theta)$$

Most common model for genealogy, $P(g|\theta)$, is Kingman's coalescent.

Phylogeographic problem:

As well as D have L, location of each sequence.
Now want to infer g, θ and μ, parameters controlling spatial movement.

$$P(g, \theta, \mu | D, L) \propto P(D | g, \theta) P(g, L | \theta, \mu) P(\theta, \mu)$$

Existing models

1. Structured coalescent, fixed number of panmictic demes

$$P(g, \theta | D, L) \propto P(D | g, \theta) P(g, L | \theta, \mu) P(\theta, \mu)$$

$$= P(D | g, \theta) \int_{L_{ancestral}} P(g, L, L_{ancestral} | \theta, \mu) \, dL_{ancestral} P(\theta, \mu)$$

2. Finite demes but genealogy process does not a prior depend on location process

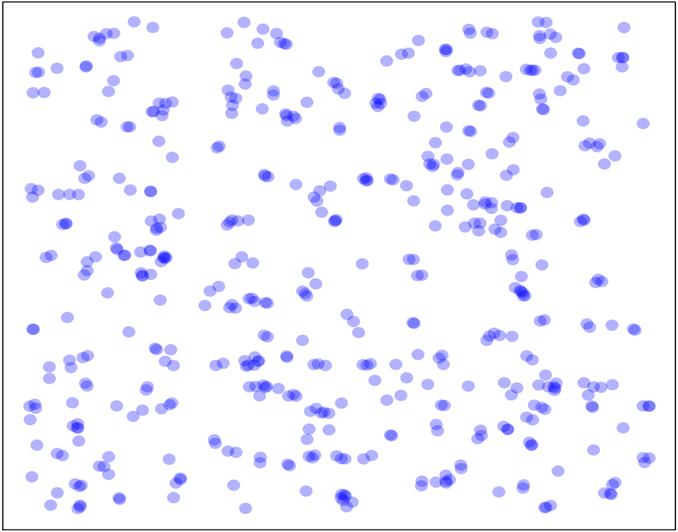
$$P(g, \theta | D, L) \propto P(D, L | g, \theta, \mu) P(g | \theta) P(\theta, \mu)$$

3. Continuous space with Brownian motion down lineages, separate from genealogy process. Based on Wright-Malecot forward model where position of off-spring is normally distributed with centre at parent.

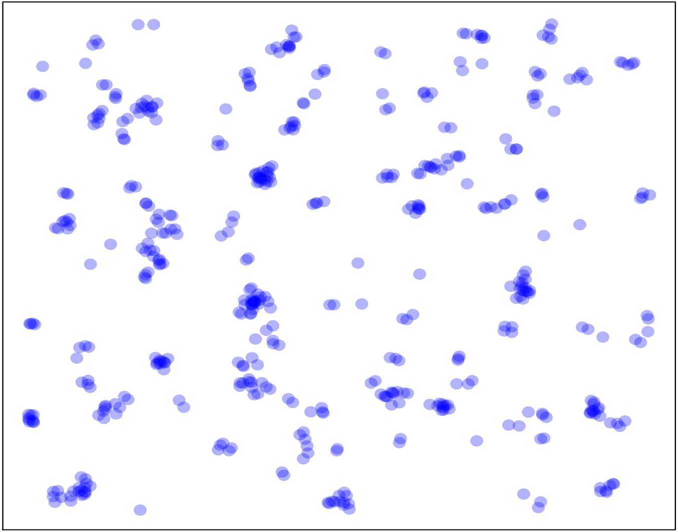
Problems with existing models

- Deme structure often not natural or known
- Even when known, number of demes must be small for structured coalescent (3-4 max?)
- *A priori* assumption of neutrality of location process unsatisfactory
- Wright-Malecot model does not produce uniform distribution across space

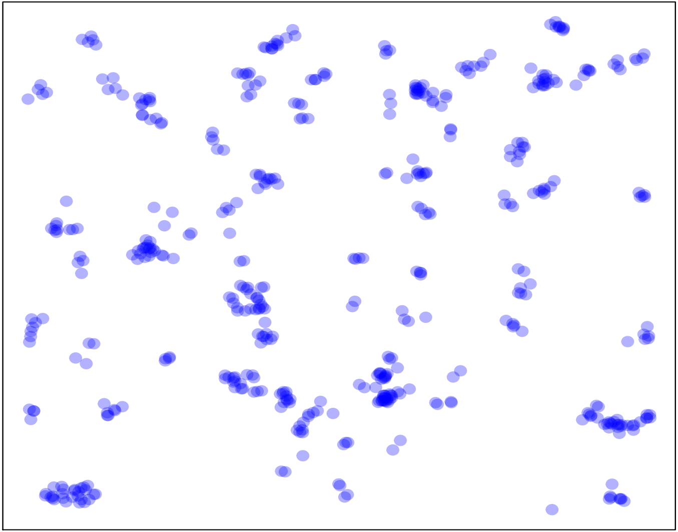
Clumping in Wright-Malecot model



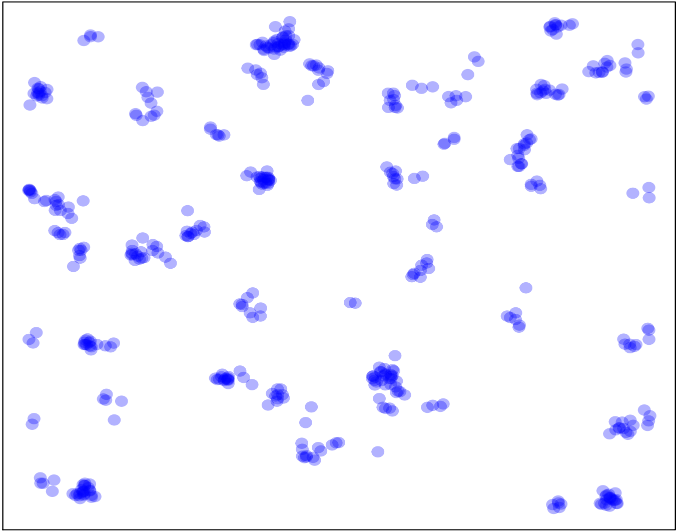
Clumping in Wright-Malecot model



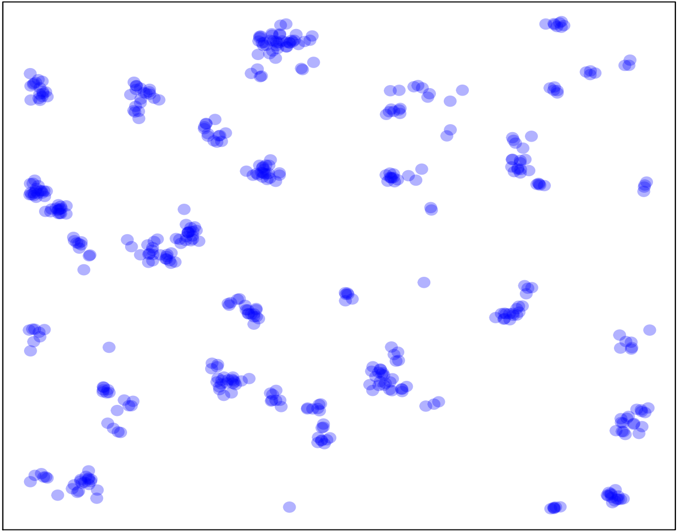
Clumping in Wright-Malecot model



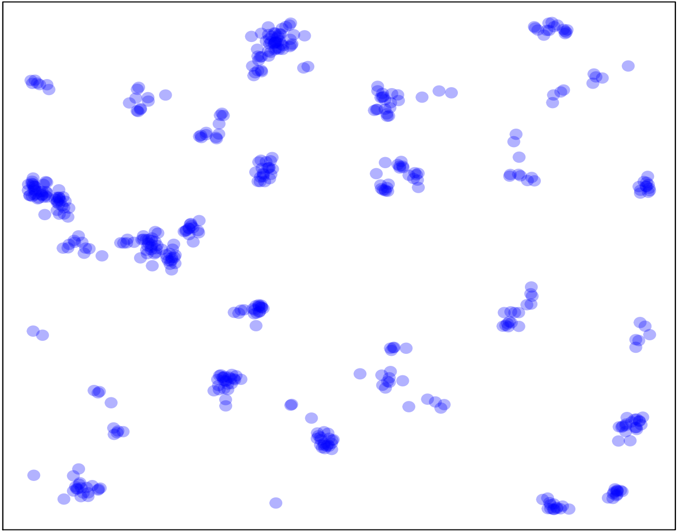
Clumping in Wright-Malecot model



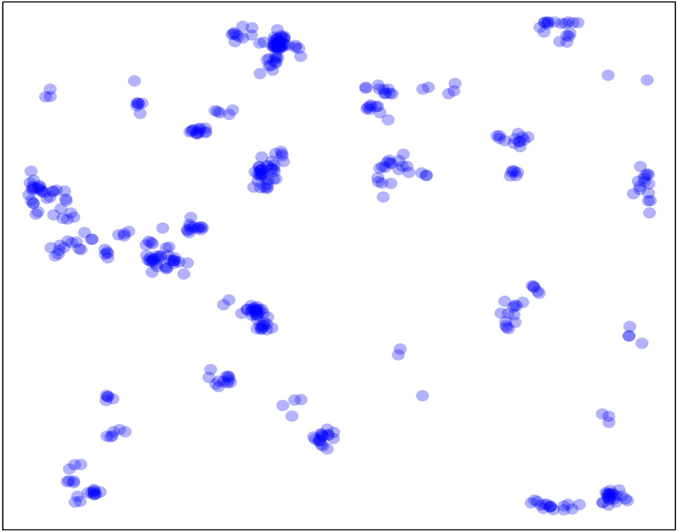
Clumping in Wright-Malecot model



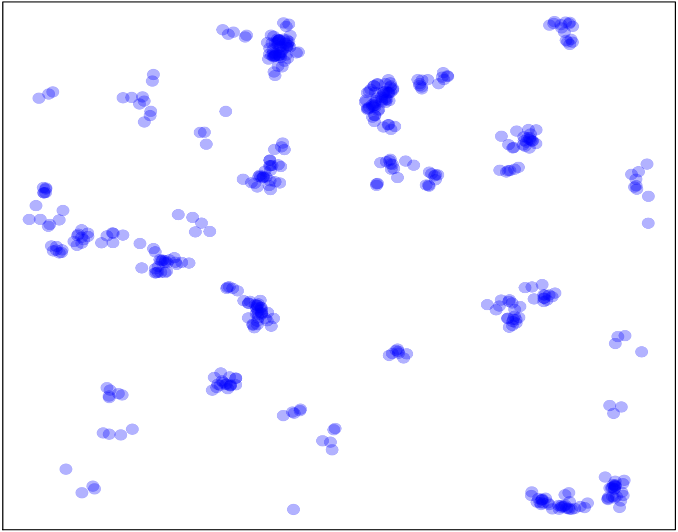
Clumping in Wright-Malecot model



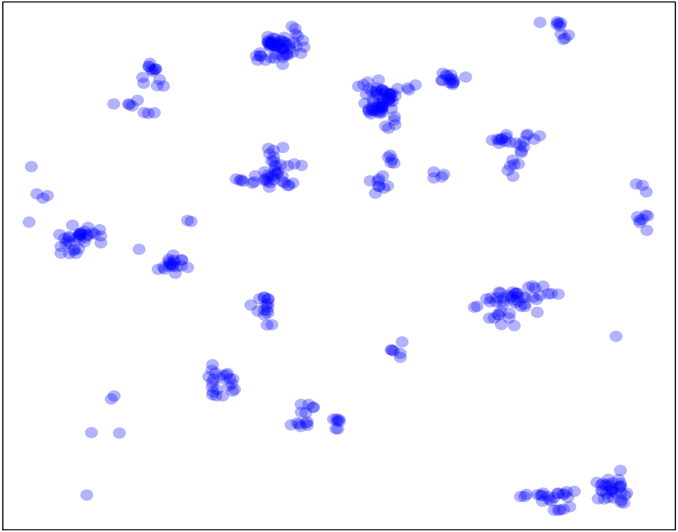
Clumping in Wright-Malecot model



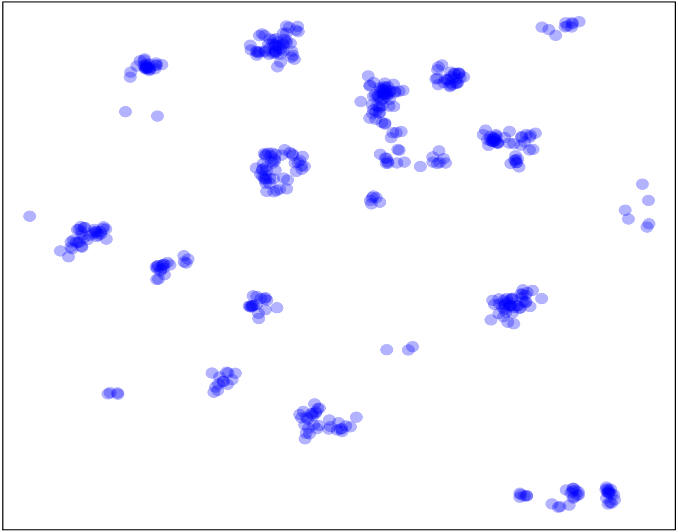
Clumping in Wright-Malecot model



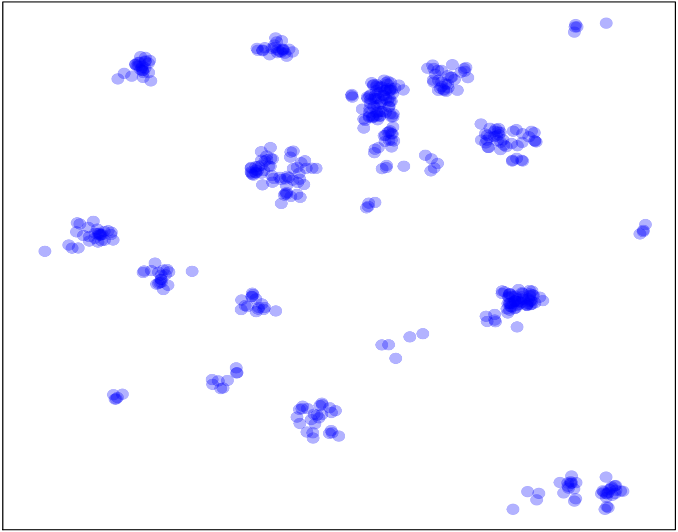
Clumping in Wright-Malecot model



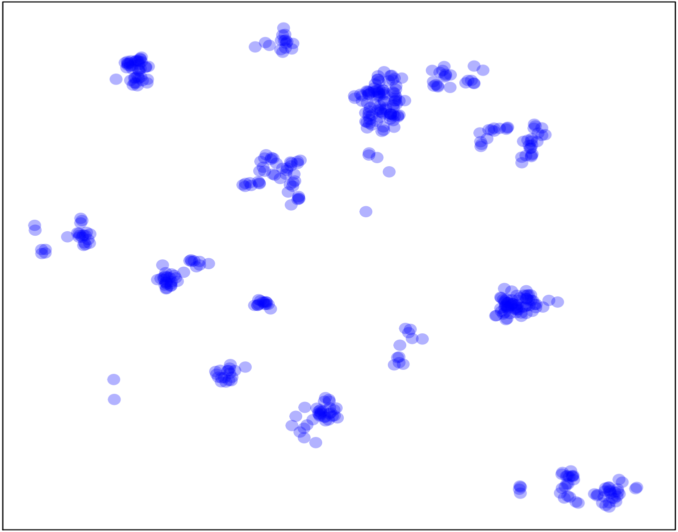
Clumping in Wright-Malecot model



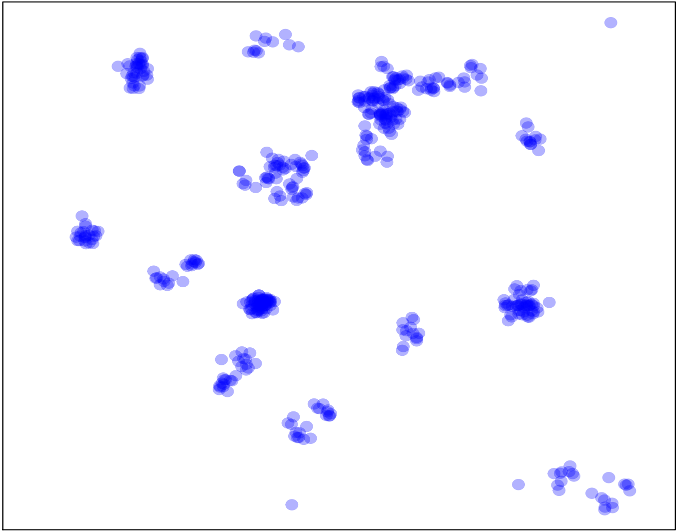
Clumping in Wright-Malecot model



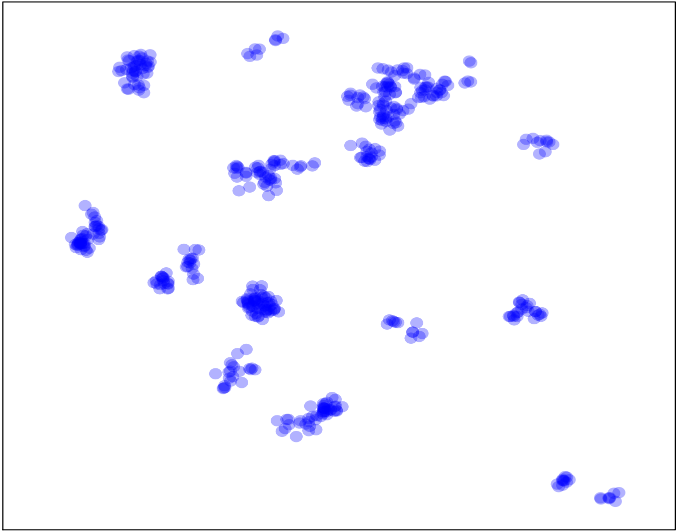
Clumping in Wright-Malecot model



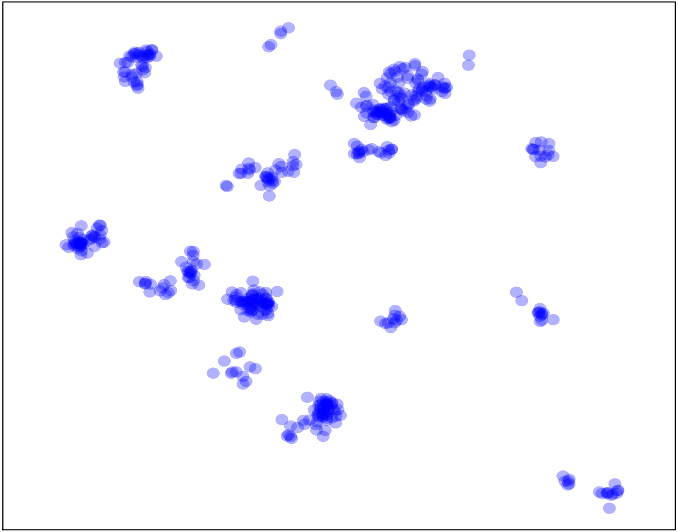
Clumping in Wright-Malecot model



Clumping in Wright-Malecot model



Clumping in Wright-Malecot model



Continuous landscape coalescent — forward process

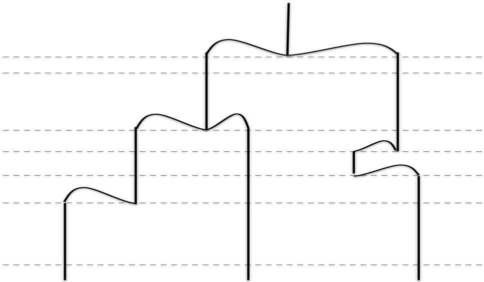
Variation of the spatial Λ-Fleming-Viot process of Etheridge, Barton, Véber et al.

- reproduction/death/migration events no longer centred on individuals
- Start with individuals spread uniformly across landscape
- Reproduction and extinction events (REXs) occur at exponential intervals with rate λ
- at a REX, a centre c is chosen uniformly across landscape
- each individual at l dies with some probability according to its distance from the centre, $u(l,c) = \mu K(l,c,\theta)$
- new individuals are born at location l' are rate according to distance from centre, so at rate $\propto u(l',c)dl'$
- All newly born individuals are the off-spring of a single individual at k who was alive before event and is chosen according to distance from centre $v(k,c) \propto K(l,c,\theta)$

Continuous landscape coalescent

The reverse process follows the ancestry of a sample of lineages. Suppose a single lineage is at location l.

- REX events still occur at rate λ
- Lineage at location l hit by REX with centre c with probability $u(l, c)$, jumps to new location l' according to pdf $v(l', c)$
- Lineages coalesce when both hit by same REX event, move to same new location.



Inference

Want the posterior $P(\lambda, \mu, \theta, g|D, L)$
To calculate, need to augment the space to include full history:

$$P(\lambda, \mu, \theta, g|D, L) = \int_{L_{anc}, M} P(\lambda, \mu, \theta, L_{anc}, M|D, L)\, dL_{anc}\, dM.$$

Approximate this integral using Bayes theorem and Markov chain Monte Carlo sampling.

Choose a more interpretable parametrisation

Hard to interpret λ, μ, θ except in terms of model.
Instead, use parameters common from Wright-Malecot model:
neighbourhood size

$$\mathcal{N} = \frac{2}{\mu}$$

diffusion rate

$$\sigma^2 = 4\theta^4 \lambda \pi \mu.$$

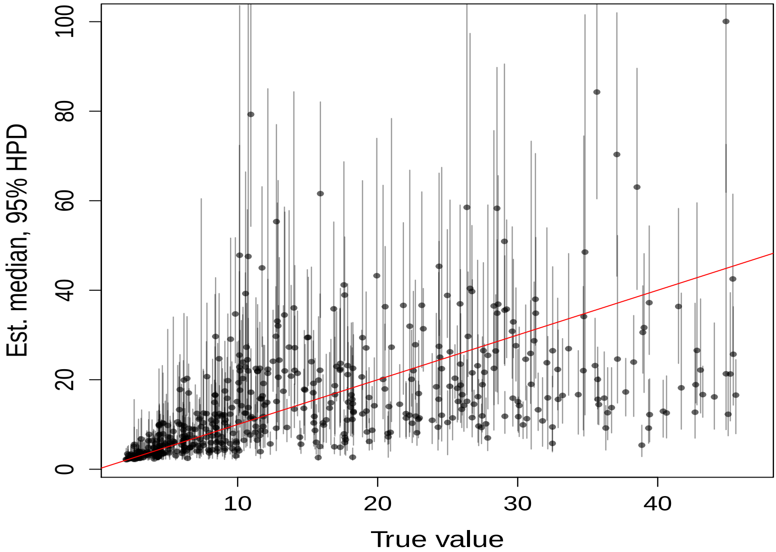
and θ.

Derivation is based on relationship between coalescent rate and effective population size N_e.

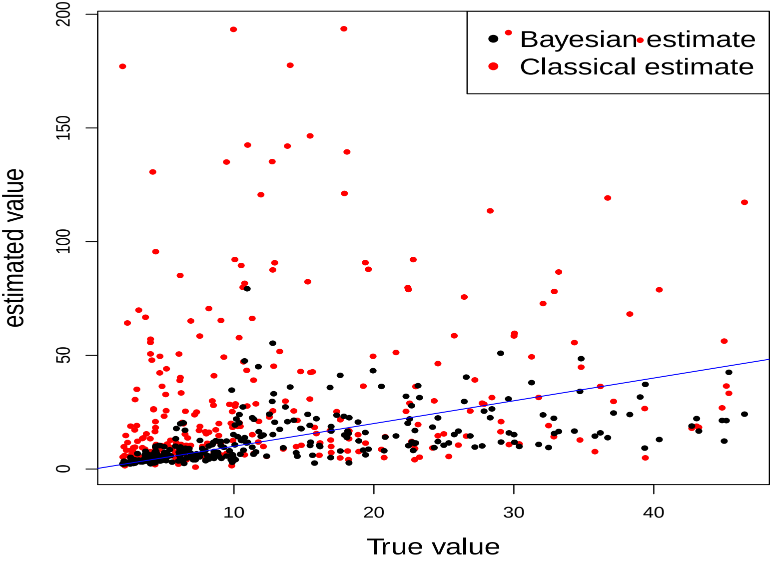
Simulations

- Landscape is 10×10
- $N_e \sim U([100, 5000])$
- $\mathcal{N} | N_e \sim U([N_e \times 10^{-3}, N_e \times 10^{-2}])$.
- $\theta \sim U([1.5, 4])$.
- 50 samples taken uniformly from 10 triangular regions comprising an average 17% of landscape
- Sequences of length 500bp simulated under Kimura model over tree.
- 500 repetitions

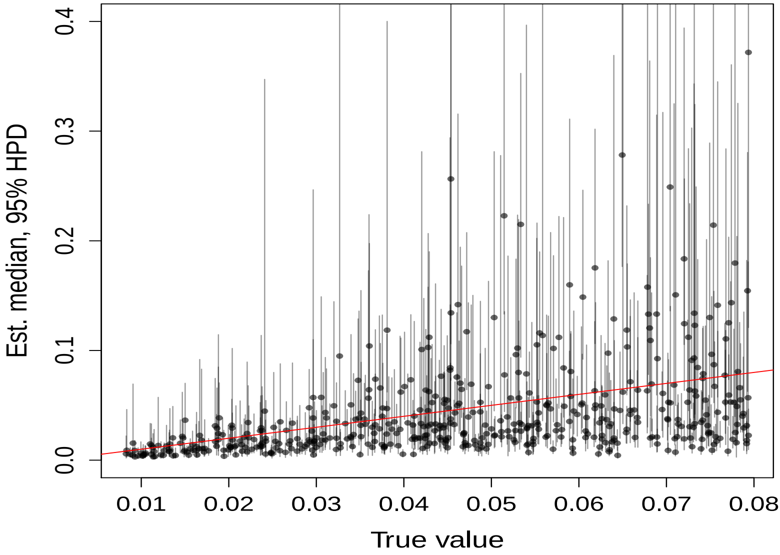
Median and 95% credible interval estimates for $\mathcal{N}$



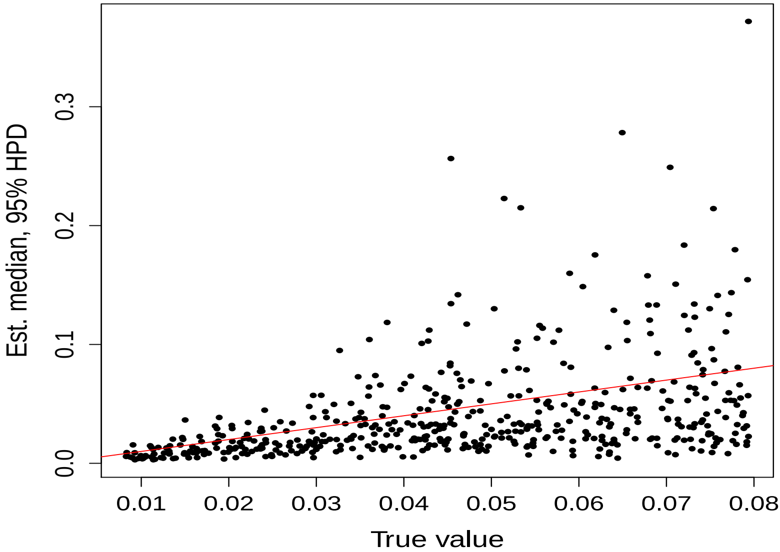
Comparison of Bayesian estimation of $\mathcal{N}$ with fixation index based estimation method



Median and 95% credible interval estimates for σ



Median estimates for σ



Summary

- It may be a feasible alternative to structured coalescent or other approximate models when doing inference
- But will need to generalise: to allow changing landscapes and non-constant populations
- Paper and software will be available soon