The effect on inferences of population size of the sampling scheme for intraspecific DNA sequences

Suzy Whoriskey

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow

January 2020

Declaration

I, Suzy Whoriskey, declare that this thesis titled *The effect on inferences of population size of the sampling scheme for intraspecific DNA sequences* and the work presented in it are my own. I confirm that:

- This work was done wholly while a candidate for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

The results in Chapter 4 were presented at the Research Students' Conference in Dublin 2016. Further work from Chapter 4 and initial results in Chapter 5 were presented at the Research Students' Conference in Durham in 2017.

And he puzzled and puzzled 'till his puzzler was sore.

Dr. Seuss, *The Grinch*

Abstract

Variation in samples of DNA sequences from within one species can be informative about the demographic processes that have affected that species, revealing signals of migration patterns and population size changes in the past. The demographic models that are fitted to the data might vary, as might the way the data are used, but one almost ubiquitous assumption is that the samples sequenced in the study are randomly chosen. Yet this is rarely plausible either because random sampling is practically impossible to perform or indeed because the samples for analysis are very consciously selected in some non-random way.

This thesis explores the robustness of a particular flexible class of models used for inference of variable population size, the so-called skyline plot methods, to non-randomness of sampling by taking a simulation approach. The particular sampling scheme investigated takes sequences belonging to one subtree (or haplogroup) of the genealogy of a non-recombining locus. Pitfalls of analyses ignoring the sampling scheme are reported and a recommendation for the interpretation of such analyses is made.

This work uses the Bayesian skyline plot model to infer population sizes and in simulation settings this model proves to be accurate in estimating population size as a function of time, from random samples. When a non-random sample defined by a haplogroup is analysed, the model can infer the *shape* of the population curve well but fails to capture the *magnitude*, when compared to the population curve inferred from a random sample or to the true population curve. Functional data analysis techniques were used to explore the relationship between the population curves inferred from random and non-random samples. After establishing that there is indeed a strong relationship between the two, the goal was to develop a straightforward post hoc correction to the inferred population curve from the non-random sample that is easy to apply and permits practitioners to allow for the violations of model assumptions caused by the non-random sample, so obtaining a more reliable estimate of population size. This is illustrated by applying the approach to samples of sequences taken from human mitochondrial DNA. The correction uses information on the prevalence of the mutation defining the non-random subtree.

Acknowledgements

First, and foremost, to Dr. Vincent Macaulay and Dr. Mayetri Gupta, thank you for your patience, guidance and friendship. My time working with and learning from you has been most enjoyable. You truly went above and beyond. Oh, how I'll miss our Tuesday mornings.

I would like to thank the EPSRC for their generous funding of this project. Thank you too to Dr. Surajit Ray for his advice and help with the Functional Data Analysis work presented in Chapter 5.

Thank you to the School of Mathematics and Statistics, especially our Head of Statistics Professor Claire Miller, for the many opportunities I have had. I will always be grateful to the staff of this department.

My fellow PhD students, being your student representative was the best fun, largely due to the coffee and cake. In particular, thank you to Eilidh. If everyone had a friendship like ours, they'd all be doing PhDs.

Mum and dad, thank you for your constant belief in me and always encouraging me to be the best I can be, without you this wouldn't be possible. To Sophie, Sam and Lucy, thank you for your endless entertainment, the laughter has helped more than you know. A special thank you to my Nanny and Aunty Phyllis, your thoughts and prayers have meant so much to me. Thank you to Joe for your advice and reassurance - particularly over the phone and especially when you were going through it too.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Deoxyribonucleic acid (DNA) plays a key role in defining biological functions in organisms. It serves as a record of the past that we carry in our cells today. Analysing this genetic record sheds light on patterns of human behaviour that can be modelled by certain mathematical processes from the field of human evolutionary genetics. In particular, we can compare DNA sequences from a number of humans to each other and this allows us to draw inferences on events that happened in the past, for example, changes in population size.

Inference about population size from DNA is a statistical problem (Wakeley, 2009). The historic population size of humans is unobservable directly and any statements about it can only be made from data collected today. By studying DNA we can learn about our ancestors through the genetic material we have inherited over thousands of years. To be able to gain any insights on what this genetic material can tell us, we need some form of variation to exploit. All human beings differ from one another, whether that is through physical traits, susceptibility to certain diseases, or even how each of us respond to certain medicines. Differences between people are evident in their DNA sequence, long chains containing specific orderings of compounds called nucleotides (denoted T, C, A and G). These differences in our sequences arise from *mutations*, changes in sequence when DNA is passed from generation to generation.

Humans are diploid individuals meaning that we carry most of our DNA in pairs of chromosomes. Humans have 23 pairs of chromosomes: one pair of sex chromosome (females have XX and males have XY) and 22 pairs of autosomal chromosomes. These chromosomes contain genes (sections of DNA with some function) and any gene can have one or more *allele*. An allele is a variant form of a gene (a specific ordering of nucleotides in a section of DNA) and most genes have many different alleles. Their proportions in a population (allele frequencies) are estimated from samples. The pair of alleles present at any location (or locus) in the genome is called the genotype.

However, studying chromosomes in terms of population size poses the problem of tracing a section of the DNA back to a most recent common ancestor (MRCA). Each offspring will have two parents, and each parent will have two parents, and each of those parents will have

two parents, and so on. Different loci turn out to be inherited through different ancestors (Li and Durbin, 2011). However, there exist two types of DNA that are inherited only through one parent, and as such are haploid. This means that the Y-chromosome is haploid because the cells of men contain one copy of that chromosome (females have no copies). Mitochondrial DNA (mtDNA) is haploid because, although cells in general have a large number of copies (in both males and females), all those copies are usually identical (Jobling et al., 2014).

Each of these types of DNA is passed on through one parent: the paternal and maternal lines, respectively. By studying haploid DNA like this, we can reconstruct one ancestral tree back to the most recent common ancestor of a sample of individuals. This thesis will focus entirely on mtDNA and Figure 1.1 shows a schematic phylogeny of mtDNA. Figure 1.1 shows how the mtDNA from many individuals is connected to one most recent common ancestor. In this particular example, this is the African mitochondrial tree, and the MRCA is the so-called *Mitochondrial Eve* (Cann et al., 1987). Throughout generations, the mtDNA carried within Mitochondrial Eve would have been passed on to her offspring and then passed on to their offspring, and so on, through the direct female line of descent. The tree branches at particular instances of a mother having more than one daughter with surviving descendants in the modern sample. Along a particular branch, by chance, a mutation or mutations may occur which distinguishes the descendants of that branch. The set of these descendants is called a haplogroup and a conventional set of names has evolved to describe them, L0, L1, L2, etc. (Kivisild et al., 2006). A haplogroup can be though of as a subtree of the full genealogy.

Because of these mutations our DNA holds information from the past that we can now extract due to the increasing advances in the study of the mechanisms of the molecular evolutionary process. This is done by estimating rates of mutations, approximate times at which mutations may have occurred, the population size through time and by testing models of mutation using DNA sequence data, made possible by the rapid accumulation of genetic sequence data, improved computational abilities and the development of statistical models suitable for these evolutionary biological processes. Population genetics is a branch of evolutionary biology that studies genetic differences within and between populations with an aim to draw inferences about properties of that population going back in time. This is done by modelling changes of DNA sequences over time and space. With dramatic advances in technology, it is now possible to determine the ordering of the nucleotide bases (the DNA sequence) in a matter of days (National Human Genome Research Institute, 2019). Compare this to the Human Genome Project (Watson, 1990), which was an international collaborative project and was successful in the first sequencing of the human genome. It began in 1990 and was not declared complete until 2003 (National Human Genome Research Institute, 2019).

In this thesis I use methods for inferring historic population sizes from the genetic variation of samples of DNA sequences. This introductory chapter will discuss the forces that affect genetic variation before giving a review of the different methods used for drawing inferences about

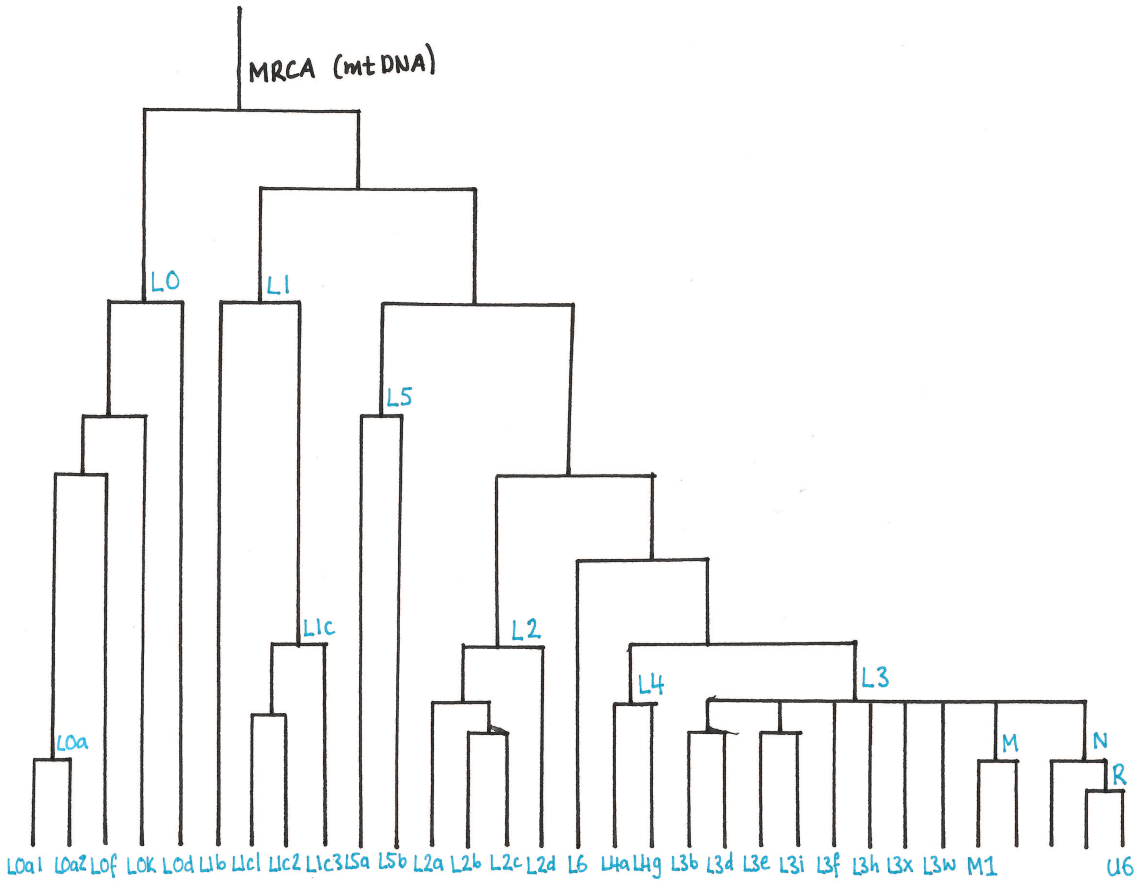


Figure 1.1: A reconstructed phylogenetic tree showing the evolution of mtDNA in African populations. Adapted from Figure 4 of the paper by Kivisild et al. (2006). The labels of each haplogroup are shown in blue, and the MRCA is labelled at the top.

a population from a given set of DNA data. I will then provide examples of the methods used in practice for sampling these DNA sequences and discuss the repercussions of the sampling scheme. The chapter will then conclude with a section describing the aims and the structure of the thesis.

1.1 Forces of genetic variation

Genetic variation in a population is shaped by different processes which can be described by mathematical models that approximate reality. These models are used to help us understand the nature of evolutionary processes and allow us to infer past processes from present day genetic diversity. These mathematical models allow us to learn information about mutation rates, population growth rates, the age of a mutation or the migration rate between two populations. Together these illustrate the structure, or the demography, of a population. This section begins with a brief overview of some of the forces, alongside mutation, that shape genetic variation in

a population, such as natural selection, recombination, migration and genetic drift.

The process of swapping one nucleotide for another from generation to generation is a type of mutation called a substitution. As such, new sequences are then introduced to the population. Another type of mutation is called an *indel* which is a term that stands for an *insertion* or *deletion* of nucleotides in the DNA sequence. The former defines the event of one or more nucleotides being added to the DNA sequence and the latter the opposite, one or more nucleotides are lost from the DNA sequence. These are copying errors and happen when the DNA is replicating itself. Indels can have disastrous consequences for the translation of the genetic code into proteins. During translation, the genetic code is mapped from a triplet of nucleotides which code for amino acids. These amino acids are then the constituents of proteins. If the triplet is read incorrectly, the relevant protein will likely break.

Recombination is a similar process to substitution in that it introduces new sequences to a population. Recombination in humans is the process by which pieces of DNA from each parent are broken up and rearranged (or recombined) to produce new combinations of nucleotides in the offspring. It occurs during the production of gametes (sex cells, i.e., sperm or egg) by meiosis and allows completely new genetic combinations for their offspring. As such, mutation and recombination are both processes that create new variation in a population.

Natural selection is another process that is at play in changing the genetic make-up of a species over time. As defined by Darwin (1859), it is the differential reproduction of individuals of different genotypes in sequential generations. Selection occurs at any point from the formation of a genotype at fertilisation all the way to the bearer of that genotype successfully creating their own offspring (Jobling et al., 2014). Fitness of an individual genotype is attributed to the following four conditions:

1. Survival into reproductive age (viability and mortality)
2. Success in attracting a mate (sexual selection)
3. Ability to fertilise (fertility and gamete selection)
4. Number of descendants (fecundity).

Over time, if a new allele is introduced into the population it will be competing against others in terms of its fitness (or against the environment in which it is living) to survive. If it is a detrimental allele, in such a way that it has a negative effect on any of the four conditions above, it will rapidly be eliminated from the population over generations and is called purifying selection. One example of natural selection is in antibiotic resistance, where bacteria have become resistant to some antibiotic treatments, leaving humans at risk of highly infectious diseases (Neu, 1992). This has happened because evolution in the bacterium is selecting alleles that can resist antibiotics and this is an opposite effective of purifying selection called positive selection.

Genetic drift is another important force affecting genetic variation where population allele frequencies change over time due to chance. Specifically, genetic drift is a type of sampling fluctuation arising from which allele copies are transmitted to the next generation from the gene pool of the current generation, given that some individuals will produce more offspring than others. Genetic drift happens in all populations but its effects are stronger in smaller populations, just as sampling variation is larger in smaller samples..

The final force of genetic variation is migration. The movement of individuals between genetically distinguishable populations will in general change allele frequencies, e.g., by introducing a new allele. This process has been modelled in many ways with stochastic models such as the island model (Wright, 1931) and the stepping-stone model (Kimura and Weiss, 1964).

1.2 Models in classic population genetics

Theoretical population genetics, now over a century old, explores mathematical models of the way allele frequencies are changed by the forces of genetic variation. Perhaps the most common model of this type is the Wright-Fisher model (Wright, 1931), which mathematically describes the change in allele frequencies in time otherwise known as genetic drift.

The Wright-Fisher model makes the following assumptions:

1. All individuals in the population die each generation and are replaced by offspring.
2. The population size N stays constant over time and is finite.
3. There is no difference in fitness between two alleles.
4. Mutations do not occur.
5. The population is randomly mating, that is, not subdivided.

Clearly, these assumptions are not realistic, but the Wright-Fisher model captures essential properties of genetic drift. For a locus with two alleles in a diploid population, we can think of the allele copies in the present generation being formed by randomly sampling with replacement from the allele copies in the previous generation. Suppose the latter consists of i copies of allele a and $2N-i$ copies of allele A, so the proportion of allele a in the previous population is $p = \frac{i}{2N}$ and the proportion of allele $A = 1-p$. Then, the probability P_{ij} that an allele with i copies in the previous generation is found in j copies in the next generation has a binomial form:

$$P_{ij} = \binom{2N}{j} p^j (1-p)^{2N-j}, \qquad j = 0, 1, \ldots, 2N. \tag{1.1}$$

The allele copies that are transmitted each generation are assumed independent, so these transition probabilities apply each generation (with an updated i and hence P), generating a

Markov chain. Then, let the current generation be generation 0 and K_t be the count of allele a in future generations. So, by Equation 1.1, K_1 is binomially distributed with parameters $2N$ and p, with $K_0 = i$. So

$$E(K_1) = 2Np = i \qquad \text{and} \qquad Var(K_1) = 2Np(1-p).$$

This says that the number of copies of allele a can be expected to remain the same on average but could become extinct ($K_1 = 0$) or become fixed ($K_1 = 2N$ copies) in the population within a single generation. Iterating Equation 1.1, the frequency of alleles will drift randomly and eventually one allele will be lost from the population. From the Wright-Fisher model we can evaluate the heterozygosity of the locus in the population (i.e., the probability that two randomly sampled allele copies are different). Under the Wright-Fisher model, heterozygosity decays geometrically with a half-life of approximately $2N \log 2$ generations (Wakeley, 2009). This gives a first indication of how genetic variation information (in this case captured by heterozygosity) might allow the estimate of population size.

In contrast to the Wright-Fisher model, the Moran model (Moran, 1962) allows generations to overlap. This model is formulated for haploid individuals (organisms that have only one set of chromosomes, compared to diploid individuals that have two sets of chromosomes) of population size N. At discrete times $t = 0, 1, 2, \ldots$, two individuals are chosen at random from the population (these individuals could be the same or different). Each individual has a chance of $\frac{1}{N}$ of being drawn. The first individual reproduces itself and the second one dies (if the same individual has been chosen then they first copy themselves and promptly die). Therefore, the population size does not change.

Let there be i copies of allele a and $N-i$ copies of allele A at some time, and j is the number of copies of allele a after one time unit. Now K_1 (the counts of allele a after one time unit) can take three possible values, $i+1$, i or $i-1$. The probability of the first is the probability that an A allele dies and an a allele is chosen to reproduce. Assuming independence, and arguing similarly for the other cases, we obtain transition probabilities

$$P_{ij} = \begin{cases} p(1-p) & \text{if } j = i+1, \\ p(1-p) & \text{if } j = i-1, \\ p^2 + (1-p)^2 & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

It follows that

$$E(K_1) = Np = i \qquad \text{and} \qquad Var(K_1) = 2p(1-p).$$

As in the Wright-Fisher model, Equation 1.2 leads to variation in the number of copies of an allele but the expected number after one time step is unchanged. To compare the rate of genetic drift under the Moran model to the Wright-Fisher model, a generation is defined to be

N birth-death events (Wakeley, 2009). Iterating (1.2) one finds that heterozygosity again decays but at double the rate of the Wright-Fisher model. The important part is that the half-life is still proportional to N.

The final model to be discussed is the Coalescent Process (Kingman, 1982), which approximates the previous two models (Wright-Fisher and Moran) for large population size. The Coalescent Process describes the genetic ancestral relationships between DNA sequences in a sample in a tree whose root is the most recent common ancestor and helps us understand how gene variants originated from a common ancestor some time in the past. This ancestral tree is called the *gene genealogy*. Under the coalescent, we trace the ancestral lineages (the set of genetic ancestors) from the sample back in time as pairs of lineages join (or coalesce). The history of a sample of size n will have $n-1$ coalescent events.

The key insight is that the rate of coalescence depends on population size. If one can estimate the rate of coalescence from modern DNA sequences, then the relationship between that rate and the population size can be exploited. Consider the transmission of allele copies through one generation of the Wright-Fisher model. Looking at that transmission backwards amounts to allele copies in the more recent generation having random ancestors in the previous generation. The chance that any pair have the *same* ancestor is $\frac{1}{2N}$ (for a diploid population of size N). It is straightforward to show that the probability that at least two of k gene copies coalesce in that previous generation is

$$\frac{1}{2N}\binom{k}{2}+O(N^{-2}), \tag{1.3}$$

where $O(N^{-2})$ contains the higher order terms of the function, which for biological and mathematical reasons, are rarely seen.

For large N, the probability of coalescence per generation is approximately $\frac{1}{2N}\binom{k}{2}$, which is inversely related to N. This rate of coalescence clearly impacts the lengths of branches in the gene genealogy. DNA sequence data permits us to learn about the lengths of those branches, hence about the rate of coalescence and hence to infer N.

Under the simple coalescent we assume random mating, no population substructure, no natural selection and that the population size is constant over time. Looking back in time, each lineage (that is, direct line of ancestry) merges with another at different times at coalescence events. Importantly, coalescent theory can be used to make inferences about demographic parameters such as population sizes. Chapter 2 will introduce the Coalescent Process more formally, building on the Wright-Fisher model and Moran model. In the next section I move on to discuss methods of inferring population sizes from genetic variation more generally.

1.3 Demographic inference from DNA variation

During the first half of the 20th century, population genetics was a field driven by theoretical intuition from models such as those described previously. The problem was there was very little data to actually test these theories. However, in the mid-1960's we had the emergence of protein electrophoretic variation, a method which analyses proteins by charging the protein molecules and transporting them through a solvent with an electrical field to study variation within a population (Lewontin and Hubby, 1966). Different alleles of the protein move at different speeds, so are separated by the field and can be visualised. This only indirectly assayed variation at the DNA level.

Today, we can sequence the entire human genome in a large number of samples and can analyse this to draw inferences about many different parameters of interest, e.g., related to the demography of humans through time. This section will discuss different types of DNA data that have been used for this purpose and some methods for performing demographic inference from them.

1.3.1 Data types

Microsatellite markers

Before we developed the ability to sequence the entire genome, microsatellite markers were a powerful tool for studying molecular evolution and population genetics (Brinkmann et al., 1998). A microsatellite is a section of repeated and adjacent nucleotides in the DNA sequence. They occur at thousands of locations in the genome and have a higher rate of mutation than other areas of DNA (Jobling et al., 2014). Microsatellites range in length. The repeat unit can be from one to ten nucleotides long and the units typically are repeated between five and 50 times (Jobling et al., 2014).

In a notable study based on microsatellites, researchers were able to determine the geographical origin of humans, supporting the hypothesis that we moved out of Africa (Bowcock et al., 1994). Microsatellites have been used to infer demographic events such as a population bottleneck (Spencer et al., 2000) and parameters such as population size (Xu and Fu, 2004) and rates of gene flow (Waits et al., 2000). However, because of the high mutation rate in microsatellite markers, they are now rarely used in population genetics, as the information about a population's demography depends on mutation rate.

DNA sequence data

DNA sequencing is the process of determining the order of the base pairs T, C, A and G in a section of DNA. Sequencing the entire human *genome* (all the DNA contained in one set of chromosomes) was an important leap in genetics, but when the Human Genome Project (National

Human Genome Research Institute, 2019) was first proposed in 1985 it seemed an impossible task. Prior to this project sequencing was difficult and extremely time-consuming. In the mid-1970's, the first generation of DNA sequencing methods were developed using a process called *electrophoresis*, exploiting the movement of charged particles of DNA under the influence of an electric field. The initial methods included the Maxam-Gilbert method (Maxam and Gilbert, 1977), Sanger chain-determination method (Sanger et al., 1977) and the related terminal dye Sanger method (Smith et al., 1986).

There now exist what are called *next-generation* sequencing methods. These have developed from the methods listed above and include a collection of high-throughput methods (Reuter et al., 2015). The Human Genome Project took 13 years to sequence one human genome using Sanger methods, whereas nowadays the entire human genome can be sequenced within a day using these next-generation methods (Behjati and Tarpey, 2013) so that genetic variation at the whole genome level can be investigated and modelled in large samples.

In the early days of DNA sequencing the method was very costly, but nowadays with advances in technology, the cost has dropped dramatically from around $300 million in 1999 to around $1,000 by 2019 (National Human Genome Research Institute, 2019). With next-generation sequencing we now require large data storage, sophisticated bioinformatic systems and fast data processing methods. It is DNA sequence data that will be used in this thesis.

Single Nucleotide Polymorphisms

Single nucleotide polymorphisms (SNPs) are specific single positions in the genome where more than one nucleotide has been observed in samples from a species. Of the roughly three billion base pairs in the human genome, all but about 10 million are the same for all humans (Jobling et al., 2014). SNPs therefore occur almost once in every 300 base pairs on average (Jobling et al., 2014). To be classified as a SNP, two or more versions of the sequence must be present in at least one percent of the general population (Jobling et al., 2014).

Despite the fact that more than 99% of human DNA sequences remain unchanged between individuals, these small variations in DNA sequences can have a major effect on how certain humans respond to diseases, bacteria, viruses, drugs, etc. The challenge in whole genome association studies for scientists is to locate and identify SNPs that correlate with particular effects in humans. If this is done, reliable SNPs could be use as predictors to help inform our decisions about things such as medical care.

SNPs occur throughout both the coding and non-coding regions of the human genome. The coding region is the section of the genome that codes for proteins. Nucleotides are arranged in groups of three and each group corresponds to a particular amino acid in the chain of the protein (Hartl and Clark, 1997). Given that a SNP in the coding region would lead to a different protein being produced, they are less common in this section of the genome. This is because a change in the protein structure due to a changed nucleotide could result in the cell not developing properly,

and this could lead to the human not surviving to pass the SNP variant on to any offspring.

Any method that models SNP variation must treat the fact that SNPs have to be discovered before being assayed in samples, and so are not random positions in the genome. They may suffer from ascertainment biases, e.g. if they were determined from a panel of DNA that is not representative of the entire human population, such as from a geographically restricted sample. This ascertainment issue is largely absent from sequencing studies.

1.3.2 Statistical methods

The Coalescent Process (Kingman, 1982) provides a framework for understanding key relationships between the genealogy of a population and its demographic history, e.g., the population size as a function of time. We can now use variation in the DNA sequence data to estimate parameters of population models because changes in historic population sizes affect contemporaneous genetic diversity. One of the early methods to estimate population size relied on observing pairwise genetic differences (the number of segregating sites between pairs of sequences in the sample) and comparing these differences with those we would expect from coalescent theory (Polanski et al., 1998; Rogers and Harpending, 1992; Watterson, 1975). These methods made use of the result that the expected number of segregating sites is related to the times that the sequences diverged. Then, under genetic drift, the times of divergence is related to the population size. However, these methods ignored important information from the genealogy (the ancestral tree) and were shown to be inefficient in estimating demographic parameters when compared to methods based on the likelihood of the sequence data (Felsenstein, 1992). These methods typically try to maximise the likelihood for a set of sampled DNA sequences given specific models of mutation and population size (Felsenstein, 1981; Griffiths and Tavaré, 1994). However, such methods proved to be highly computationally intensive and difficult to implement when assuming any complex demographic models for historic population size (such as complicated changes in population size or patterns of migration).

This then opened the door to Bayesian approaches, starting with Bayesian inference from microsatellite data (Wilson and Balding, 1998). Due to developments in Markov-Chain Monte-Carlo (MCMC) algorithms, the aforementioned problem of probabilistic analysis of highly complex genetic processes could be overcome and allowed inferences about both the genealogy and population size to be drawn, by sampling the whole space of possible ancestral trees from its posterior distribution. It was found that microsatellite data contained limited information to make strong inferences such as precise estimations of the time to most recent common ancestor but uncertainties in inferences and parameters such as the mutation rate could at least be fully accounted for. This method is the stepping stone to popular Bayesian methods for inferring population sizes today.

Approximate Bayesian Computation (ABC) methods were developed specifically for dealing with computational problems in population genetics but have since had much wider application

(Csilléry et al., 2010). These approaches allow for greater flexibility since there is no longer a need to specify the likelihood function in order to determine the posterior distribution of a parameter and the use of summary statistics significantly reduces the dimensionality of the data set (Beaumont et al., 2002). ABC is very popular because it can assume complex models and can use large datasets. These benefits mean that the approach is widespread in population genetics, phylogenetics, molecular evolution, conservation genetics and infectious disease epidemiology (Lopes and Beaumont, 2010). However, rather than use the data itself, ABC requires summaries of the data, such as mean pairwise difference, etc. This then means that some of the information in the data is lost. In this work, the likelihood function is available and therefore there is no need to use ABC methods.

With the rapid accumulation of DNA data, more methods were developed in both maximum likelihood and Bayesian frameworks (Drummond et al., 2002; Kuhner et al., 1998; Pybus et al., 2000). This then led to the model of interest in this thesis: the Bayesian Skyline Plot model (Drummond et al., 2005). This model will be described in detail in Chapter 3, but in summary it allows for a piecewise constant estimate of the population size going back in time from an observed set of DNA data. This model is very popular in the field of population genetics because of its ability to account for uncertainty particularly in the phylogeny. This method assumes a flexible demographic model for population size, which reduces the risk of fitting models that do not adequately represent the truth. It does not assume a fixed reconstructed ancestral tree. As in the other Bayesian methods, it samples over the whole space of possible trees.

Other notable methods of inferring historic population sizes from variation in DNA data include the reversible jump MCMC method used to select models which allow Bayesian non-parametric estimation of the demographic history by using the reconstructed genealogy from a sample of DNA sequences (Opgen-Rhein et al., 2005). By considering coalescent times in an ancestral tree as a point process and using Gaussian Markov random field methods, population trajectories can be estimated by the conditional intensity of the point process (Palacios and Minin, 2013). The site frequency spectrum (Wakeley, 2009) can be used to infer population histories. The recent CubSFS method infers changes in population size from SFSs using a point process (Waltoft and Hobolth, 2018). Extending this idea, methods involving the joint frequency spectrum within and between populations were developed, using a composite likelihood scheme (Excoffier et al., 2013; Gutenkunst et al., 2009). A last notable method was developed for inferring the population size of two populations that diverged from a common ancestral population, using an MCMC method that allowed for the parameters to be estimated under a Bayesian or a likelihood framework (Nielsen and Wakeley, 2001).

1.3.3 Case study of non-random sampling

One particular type of non-random sampling considered in the literature, and the type of non-random sampling that this thesis will be concerned with, is choosing a sample of DNA sequences

to analyse on the basis of those sequences all belonging to a specific haplogroup. For example, researchers have used non-random sampling of this kind to draw inferences about population size at the time of humans moving around and out of Africa (Atkinson et al., 2008; Endicott and Ho, 2008; Rito et al., 2013; Soares et al., 2011). In particular, a study conducted by Rito et al. (2013) investigated the expansion of modern humans across Africa, focusing on the L0 haplogroup. This is an example of non-random sampling since the data included 42 complete mtDNA sequences of individuals belonging to haplogroup L0. In their results, the authors interpret the inferred population size from the Bayesian Skyline Plot model (my Figure 1.2) as further evidence of the spread of the L0 haplogroup from eastern to southern Africa (their Figure 1.3) because the main increase in population size of the L0 haplogroup is observed between 500 and 5,700 years ago. This result is consistent with a founder analysis they carried out on migration times of this haplogroup. The authors also reference previous work that detected a pattern of dispersal in the L3 haplogroup (Soares et al., 2011) that indicated strong Late Glacial or postglacial signals. However, this is not apparent in the inferred population size for the L0 haplogroup. The authors note that this signal is missing from the L0 haplogroup and say that one subhaplogroup L0a was probably involved in some of the said dispersal. This lack of signal in the L0 haplogroup could perhaps be because only a subtree is being analysed here. Could it be that if we were to include a more diverse selection of haplogroups we could identify changes in population size that correlate with climate changes?

The authors do supplement their results with similar analysis carried out on four samples randomly selected from a large set made from diverse haplogroups to compare with their results from the L0 haplogroup. However, as much as this is a nod to acknowledging the fact that studying the L0 haplogroup may not tell the full story, these four samples themselves do not display the desirable randomness we would require. These samples are from a set of published pan-African whole mtDNA sequences (Behar et al., 2008), not focused on any specific haplogroup, but will absolutely not be a random sample because of the nature of data collection.

The work referenced by Rito et al. (2013) has the title *The Expansion of mtDNA Haplogroup L3 within and out of Africa* (Soares et al., 2011), in which the researchers were solely concerned with investigating DNA sequences belonging to L3 and after collecting the non-random sample of sequences continues with their analysis using the Bayesian Skyline Plot model (Figure 1.3). When interpreting this population size inferred from L3, Rito et al. first focus on the increase in population size 40,000 years ago. They say that it corresponds to the emergence of haplogroups between 40,000 and 50,000 years ago since this signal is not observed in the Bayesian Skyline Plots of any of the individual branches belonging to L3 subhaplogroups (not shown). The second population increase at around 11,500 years ago corresponds to the beginning of the Holocene (the current epoch in the Earth's history, the time between the last ice age and the present). What is most striking here is that the authors refer to their supplementary material highlighting that Bayesian Skyline Plots of individual subhaplogroups of L3 do not display this signal of

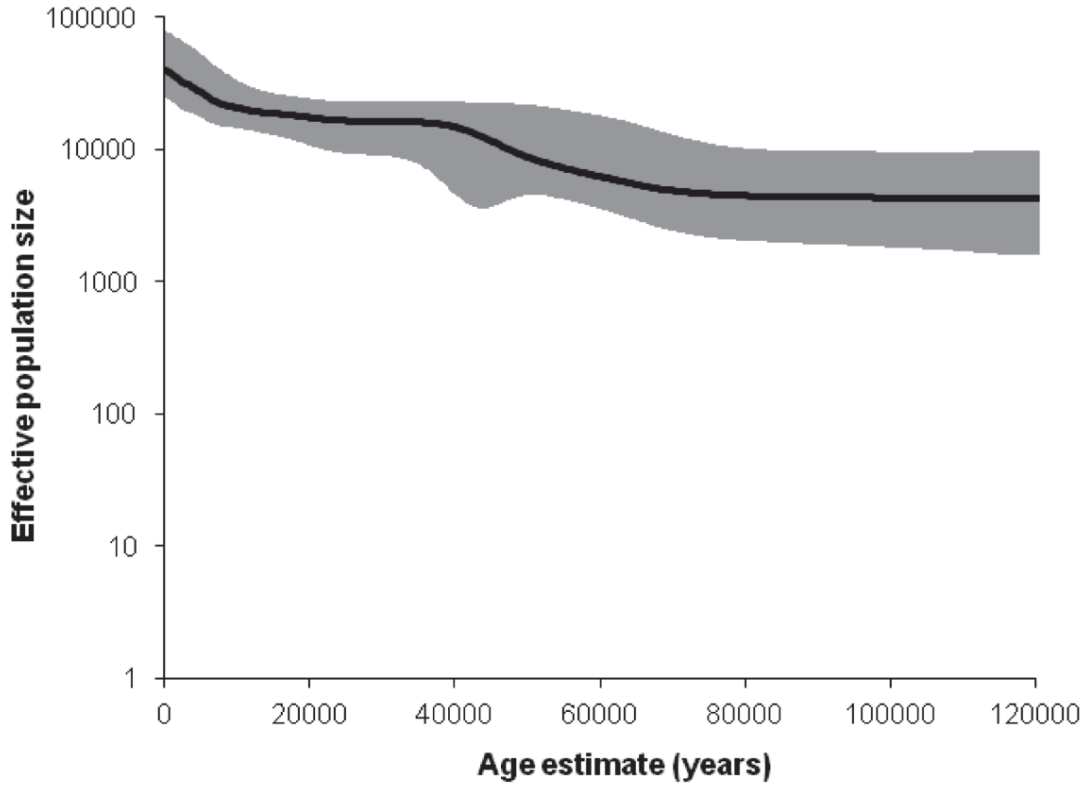


Figure 1.2: Bayesian Skyline Plot, indicating hypothetical effective population size through time, based on data form the entire L0 haplogroup. Taken from Figure 6 of the paper of Rito et al. (2013)

population size increase, but by combining all L3 subhaplogroups and analysing the full L3 haplogroup then we are able to detect this population size change.

This leaves one wondering if combining subhaplogroups presents a richer estimate of the population size over time then does this imply that when we analyse a subtree (i.e., subhaplogroup) we are left with a less informative population size estimate? And if this is the case, then will the same argument hold when we analyse specific haplogroups and not a random sample of contemporary mtDNA sequences?

One large problem exists here. I have suggested that the results from Rito et al. (2013) and Soares et al. (2011) would be more informative based on a random sample, but how would this random sample be defined? If the human population were panmictic, the meaning of a random sample is clear. But since it is not, and the Bayesian Skyline model treats it as though it were, the very definition of the 'population' from which one would sample becomes problematic.

Any real data set will be neither randomly sampled nor structureless so rather than suggest to the researcher only to proceed with population size inference using an inaccessible random sample, one approach would be to identify the extent of biases in population size estimates that

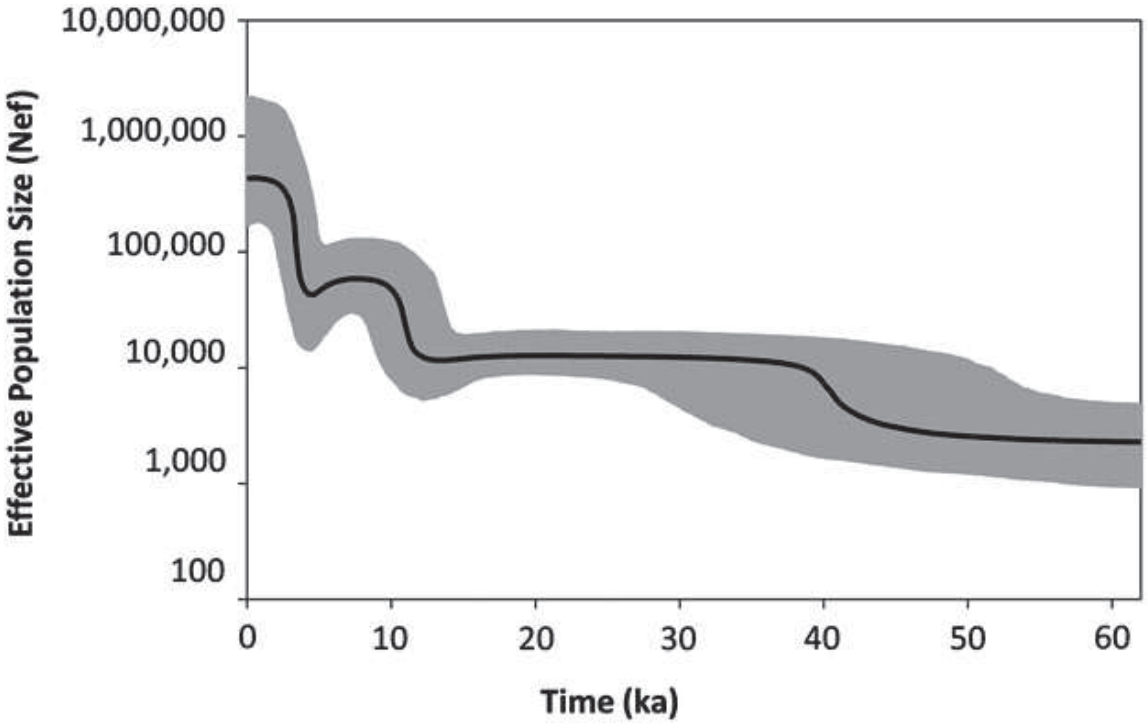


Figure 1.3: Bayesian Skyline Plot of the median of the hypothetical effective population size through time based on data from the entire L3 haplogroup except for haplogroups M and N. Taken from Figure 2 of the paper of Soares et al. (2011).

are based on assuming an idealised sample and apply a correction to them. This is the aim of this thesis, to explore the effect of non-random sampling, when random sampling is assumed in the model. I will take a simulation-based approach to make a start at doing this.

1.4 Aims and structure of this thesis

The aims of this thesis are:

1. to investigate the accuracy of inferred population size produced from the Bayesian Skyline Plot model, by simulation;

2. to determine whether or not a particular sort of non-random sample of DNA sequences provides a biased estimate of the inferred population size, and if so, to quantify the error in inferred population size; and

3. to provide a recommendation to researchers for interpreting inferred population sizes (as a function of time) from samples of DNA sequences that are not a representative random sample of the population.

The chapters are structured as follows:

Chapter 2 presents the Generative Model, including the conditions under which DNA sequences are simulated to be used to infer population sizes going back in time. It presents the theory behind the Coalescent Process. It then moves on to describe DNA sequence mutation models. These mutation models are used to simulate variability within the DNA sequences, and it is this variability that will ultimately be modelled to produce the inferred population size. Lastly, this chapter describes in detail the simulation settings for the DNA sequence data and how the random and non-random samples of DNA sequences are determined.

Chapter 3 then gives a detailed explanation of methodology behind the Bayesian Skyline Plot model, including the building blocks of the model, the Classic and Generalised Skyline Plot models, which is used in this work to produce the inferred population size from DNA sequence data.

Chapter 4 analyses the results of applying the Bayesian Skyline Plot model to data. It begins with a reanalysis of the classic data set of Ingman et al., (2000), that provides parameter estimates that are then used in the generative model for simulated DNA sequences. The results of the analysis of these simulated DNA sequences through the Bayesian Skyline Plot model are then discussed. This allows assessment of how well the model actually does in inferring population size going back in time.

Chapter 5 then attempts, in various ways, to model the inferred population sizes in such a way that a method could be proposed to convert the inferred population sizes from a non-random sample of DNA sequences into the inferred population size from what would be an equivalent randomly sampled set of DNA sequences. The conclusions are then applied to a real data set and another simulated data set with a highly variable true population size. The thesis is concluded in Chapter 6. This chapter summarises and discusses the main findings and provides warnings and recommendations for situations in which researchers analyse a sample of DNA sequences to infer population sizes from data in a subtree.

Chapter 2

The Generative Model

This chapter will introduce the theory and methods used to simulate the DNA sequence data used for analysis in this thesis, which aims to investigate the implications of non-random sampling of DNA sequences when performing inference on such data. The DNA sequences are simulated, and not sampled, because this allows particular parameters, such as the parameters of the demographic model, to be controlled which in turn allows assessment of the accuracy of inferences.

To begin with, a short summary of some key genetic ideas will be described, then the Coalescent Process (Kingman, 1982) will be introduced, which approximates a variety of models that describe a population's evolution over time (such as the Wright-Fisher and Moran models described in Section 1.2). It can be used to interpret patterns of genetic variation and, importantly, as a tool for simulating DNA sequences. Next, models of mutation will be described, the source of variability between DNA sequences. This includes the treatment of the distribution of mutation rates across a given sequence. Lastly, the simulation conditions will be discussed. These include choosing the parameter values, determining the demographic model and a thorough explanation of the algorithm used to simulate DNA sequences. A list of the mathematical symbols used throughout this thesis is presented in Appendix A.

2.1 Foundations of genetics

This section will describe briefly some of the key concepts in genetics, ones which are necessary for understanding the models of population evolution that are used later. A glossary of genetic terms used in this thesis is available in Appendix B.

The genome is the complete set of genetic material present in an organism, and consists of over 3 billion nucleotides in humans. Each section of the genome will have been inherited differently, through *recombination*, a process which creates a new chromosome from each pair of chromosomes in both parents during meiosis (Jobling et al., 2014) . Due to the nature of recombination, different sections of the genome that the offspring inherits will present a different

possible tree of ancestry, with a distinct most recent common ancestor.

To avoid this complication, one distinct section of the genome is often analysed, the mitochondrial DNA (mtDNA). The mtDNA is contained within the cytoplasm of cells and is only passed on through the female line (Giles et al., 1980). This means that any sample only has one most recent common ancestor (Cann et al., 1987) and all the nucleotides have the same history, or genealogy. However, the structure of this genealogy will not be known in general.

Mitochondrial DNA is a small circular molecule found inside the mitochondria organelles, described as the 'powerhouse' of the cell, due to the fact that they convert energy for the cell (Jobling et al., 2014). Aside from the fact that mtDNA is maternally inherited, it has a high copy number, lacks recombination and has a high mutation rate and so is a popular and convenient choice in population history studies (Giles et al., 1980; Jobling et al., 2014). The sequence has two main regions; the control region and the coding region. The control region is responsible for DNA synthesis and the coding region is responsible for the coding of proteins. This thesis will focus on modelling the variation in the coding region. Since mtDNA is only inherited through the maternal line, we are concerned with a haploid population of size N (not $2N$ as in the diploid genome (Section 1.2)), where N is the number of reproducing females.

DNA (including mtDNA) is a double-stranded helix made up of pairs of nucleotides, the pairs arising because each strand is complementary to the other. If one strand in the double helix is known, the second can be inferred and this is the basis of DNA replication. As explained in the introductory chapter, DNA is a string of four nucleotide components. These are split into two groups: *purines* and *pyrimidines*. Purines are the two-carbon nitrogen ring bases adenine and guanine (A and G, respectively) while pyrimidines are the one-carbon nitrogen ring bases thymine and cytosine (T and C, respectively). In the DNA double helix, nucleotides T and A pair together, and nucleotides C and G pair together, the so-called *nucleotide base pairs*.

Coding DNA carries information that, after a process called transcription and translation (Jobling et al., 2014), results in *proteins*. There are thousands of proteins that make up the physical characteristics of our bodies like our bone structure, eye colour and our ability to digest food. Proteins are made up of *amino acids* strung together in various combinations. There are 20 different amino acids and proteins can contain more than 100 amino acids, so the variety of combinations is huge. Amino acids are coded for by groups of nucleotides read in triplets, called codons.

The scientific theory of evolution, first proposed by Charles Darwin in the mid-19th century, observed that in successive generations members of a population are more likely to be replaced by the offspring of parents with favourable characteristics that have enabled them to survive and reproduce in their respective environments, (Darwin, 1859). Evolution is the change in heritable characteristics in populations over successive generations caused by the genes that are passed on during reproduction. The forces of genetic variation that were discussed in Section 1.1 (mutation, recombination, natural selection, genetic drift and migration) result in these changes

between generations.

The most relevant force of evolution to this particular thesis is mutation, specifically substitutions. Mutations are changes to the DNA sequence and occur throughout the genome and they produce genetic variation which can then be modelled statistically to estimate rates of mutation through time, which in turn can be used, alongside processes such as the coalescent, to infer population sizes. A special case of mutations are *substitutions* which are simply the change of one nucleotide to another. There are two types of substitution: *transitions* and *transversions*. A transition is defined to be a substitution between two purines or two pyrimidines (i.e., $T \leftrightarrow C$ or $A \leftrightarrow G$) which is an exchange between nucleotides of similar shapes (two-carbon nitrogen rings or one-carbon nitrogen rings). A transversion is a substitution between a purine and a pyrimidine (i.e., T, $C \leftrightarrow A$, G). Going forward in this thesis when describing a mutation it will be referring to the substitution type of mutation.

2.2 The Coalescent Process

To *coalesce* is to 'come together to form one mass or one whole' (Oxford English Dictionary, 2008). mtDNA is transmitted to siblings from a mother. This effectively reflects a branching of the family tree. Thinking backwards in time, the two branches join in the mother, or in terms of the mtDNA, the siblings' mtDNA coalesce in the mother's mtDNA.

The Coalescent Process describes the genetic ancestry of a sample of DNA sequences by approximating various models of the evolution of populations and makes predictions about patterns of genetic variation (Wakeley, 2009). The Coalescent Process is important in understanding the relationship between a population's demographic history and the genealogy of small segments of the genome (the ancestral tree). The reason we are interested in the Coalescent Process is because from variation in the DNA sequence, we can reconstruct the ancestral tree and by doing this, we can then infer the population size going back in time.

One way to visualise the Coalescent Process is through an example of a genealogical tree, or genealogy. Figure 2.1 shows an example of a genealogy of a sample size $n = 5$. These 5 individuals are traced back to one MRCA (along the female lines) some time in the past, through four coalescent events (which form ancestors 6, 7, 8 and 9). These five individuals in the sample (individuals 1, 2, 3, 4 and 5) are connected to their respective ancestor by *branches* in the tree. The lengths of these branches are determined by *waiting times*, the times until a coalescent event happens.

These waiting times are an important part of the analysis of DNA sequences since the expected number of mutations on a branch (which is a sum of waiting times) is proportional to its duration in time under the molecular clock hypothesis (Zuckerkandl and Pauling, 1965). The molecular clock hypothesis assumes that the mutation rate is constant throughout the tree.

The Coalescent Process is a stochastic process and is described mathematically as follows,

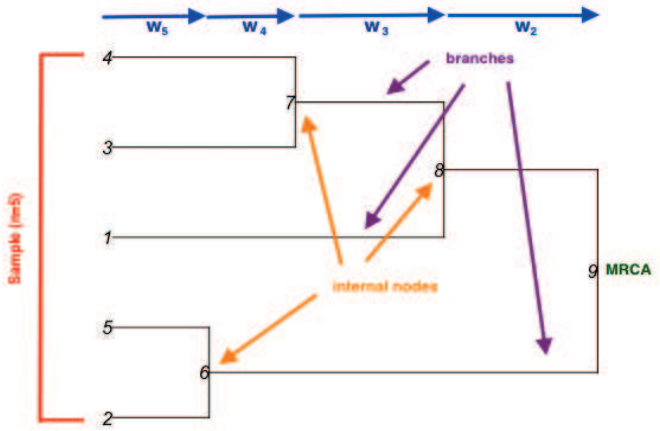


Figure 2.1: Example genealogy of a sample of size $n = 5$ in the Coalescent Process.
The sample is highlighted in red, the waiting times in blue, the branches in purple, nodes indicating coalescent events in orange and the MRCA in green.

(Kingman, 1982). Time, t, is measured in units of N_0 (the haploid population size, assumed constant) generations and on a continuous scale and *backwards*, as such the present is set at time $t = 0$. The n sequences sampled at this time are traced back to the MRCA of the sample through coalescent events. These coalescent events give rise to a death process (i.e., a non-homogeneous Poisson process). The rate of coalescence (or death) is:

$$\lambda_k = \binom{k}{2}, \tag{2.1}$$

where k is the number of lineages in the genealogy before a coalescent event and ranges from the sample size n down to 2. Note that when t is measured in units of generations λ_k is $\frac{\binom{k}{2}}{N_0}$. Approaching the root of the tree, there are only 2 lineages (or branches) left before they coalesce to form the MRCA. After each coalescent event, the rate of coalescence changes since k decreases by 1. The waiting times, W_k, $k = 2, \ldots, n$, between coalescent events are independent and exponentially distributed,

$$W_k \sim \text{Exp}(\lambda_k). \tag{2.2}$$

Therefore, to simulate the Coalescent Process in practice, the waiting times are drawn from exponential distributions for a specified n and N_0. As each two random individuals coalesce, they are removed from the population and replaced by an ancestor in the tree. The lengths of branches between nodes in the tree are then sums of the relevant waiting times.

2.2.1 Changing population size

The previous section described the Coalescent Process for a population of constant size, N_0. This is an unrealistic assumption for most populations, certainly humans and a changing population size, whether one that is growing or has a sudden jump (step) or temporary reduction

(bottleneck), needs to be accommodated. An assumption of a constant population size, N_0, implies a constant rate of coalescence per generation between a pair of lineages of $\frac{1}{N_0}$. Since this rate of coalescence depends on the population size, a method for building in a changing rate of coalescence (that corresponds to the respective population sizes) is required.

Ideally, the method might take the waiting times simulated under the constant population size Coalescent Process and transform them to allow the rate of coalescence to vary with the changing population size. This method of transforming the waiting times is called *time warping*, and effectively 'stretches' and 'squeezes' the time axis when the population size rises and falls, respectively.

Under the constant population coalescent, let $A^c(t)$ be the number of ancestors of a sample at time t (c indicates that this is under a constant population size assumption). This $A^c(t)$ is a death process (since we are moving backwards in the tree, individuals are coalescing and so $A^c(t)$ cannot increase), where the death rate, the rate of moving from state k to $k-1$, is $\binom{k}{2}$ when t is measured in units of N_0 generations, and k is the number of lineages in the genealogy at time t, as before. The relationship between the number of ancestors under a constant population size Coalescent Process and the number of ancestors under a variable population size Coalescent Process is,

$$A^v(t) = A^c\big(\Lambda(t)\big), \tag{2.3}$$

where $\Lambda(t)$ is a (generally) non-linear function of time, defined formally later in this section and $A^v(t)$ is the non-homogeneous Markov death process for the variable population size (Griffiths and Tavaré, 1994). This process starts from $A^v(0) = n$ and moves down in steps of 1 until it reaches 1, as before.

To find the transition probabilities of the process, first consider the transition probabilities under the constant size coalescent case:

$$\mathrm{P}(A^c(t+\delta t) = i | A^c(t) = k) = \begin{cases} \binom{k}{2}\delta t + O(\delta t^2) & \text{if } i = k-1 \\ 1 - \binom{k}{2}\delta t + O(\delta t^2) & \text{if } i = k \\ 0 & \text{otherwise,} \end{cases} \tag{2.4}$$

where δt is an increment of time, and note that $t + \delta t$ will come before t. Then, for the varying population size case where the population size at time t is denoted $N(t)$, consider the function λ, defined by

$$\lambda(t) = \frac{N(0)}{N(t)}, \tag{2.5}$$

which gives the changing rate of coalescence as a function of time (note that the rate of coalescence is changing because it depends on a changing population size). The transition probabilities

of the varying population size death process, $A^v(t)$, are given by

$$\mathrm{P}(A^v(t+\delta t)=i|A^v(t)=k)=\begin{cases}\binom{k}{2}\lambda(t)\delta t+\mathrm{O}(\delta t^2), & \text{if } i=k-1\\ 1-\binom{k}{2}\lambda(t)\delta t+\mathrm{O}(\delta t^2), & \text{if } i=k\\ 0 & \text{otherwise.}\end{cases} \tag{2.6}$$

Now to show that (2.3) satisfies (2.6). The transition probability to go from k to $k-1$ ancestors in the variable population case is

$$\begin{aligned}\mathrm{P}(A^v(t+\delta t)=k-1|A^v(t)=k)&=\mathrm{P}(A^c(\Lambda\{t+\delta t\})=k-1|A^c(\Lambda\{t\})=k)\\&=P\Big(A^c\Big(\Lambda(t)+\frac{\partial\Lambda}{\partial t}\delta t\Big)=k-1|A^c(\Lambda\{t\})=k\Big)+\mathrm{O}(\delta t^2)\\&=\mathrm{P}(A^c(t'+\delta t')=k-1|A^c(t')=k)+\mathrm{O}(\delta t^2),\end{aligned}$$

where $\Lambda(t)=\int_0^t\lambda(u)\mathrm{d}u$, $\delta t'=\delta t\frac{\partial\Lambda}{\partial t}$ and $t'=\Lambda(t)$. A Taylor expansion to first order is used in the second line to obtain $A^c\Big(\Lambda(t)+\frac{\partial\Lambda}{\partial t}\delta t\Big)=A^c(t'+\delta t')$. The transition probability is $\binom{k}{2}\delta t'+\mathrm{O}(\delta t^2)$, and since $\delta t'=\delta t\lambda(t)$, this gives $\binom{k}{2}\delta t\lambda(t)+\mathrm{O}(\delta t^2)$, which is the correct transition probability for the variable size.

There are three practical steps in time warping and they are as follows:

1. Generate waiting times (until coalescent events) under the constant population size coalescent, for lineages $k=n$ down to 2, where n is the sample size of DNA sequences. These waiting times are drawn independently from the exponential distribution with rate $\binom{k}{2}$ and are each denoted t_k^c where c represents the constant size case, as before.

2. Take the cumulative sum of these waiting times, $s_k^c=\sum_{i=k}^n t_k^c$. These are the transition times (for transitioning from state k to state $k-1$) in the constant population size case, measured from the present time.

3. Apply the time warping method to these constant transition times to find the transition times for the variable population size case. This is done by finding $s_k^v=\Lambda^{-1}(s_k^c)$ and the waiting times t_k^v can be found by subtraction.

To illustrate this, a population size trajectory is introduced that will serve as an example to explain the time warping method in more detail. Consider the piecewise constant population size function

$$N(t)=\begin{cases}N_0 & t_0\le t<t_1\\ N_1 & t_1\le t<t_2\\ \vdots & \vdots\\ N_{b-1} & t_{b-1}\le t<\infty,\end{cases} \tag{2.7}$$

where $t_0 = 0$ is the present, i.e., the time of the DNA sequence sample (of size n). At that time, the corresponding population size is N_0 (the initial population size). Each section of the piecewise constant population function will be called an epoch and is labelled 0 to $b-1$, where b represents the total number of epochs. The population size in epoch i is N_i for $i = 0, 1, \ldots, b-1$ and the time of the change in population size from N_{i-1} to N_i is t_i for $i \geq 1$.

Consider the following population function with 3 epochs (so that $b = 3$),

$$N(t) = \begin{cases} 1000 & 0 \leq t < 600 \text{ years} \\ 100 & 600 \leq t < 15000 \text{ years} \\ 10000 & 15000 \leq t < \infty \text{ years}, \end{cases} \tag{2.8}$$

illustrated in Figure 2.2.

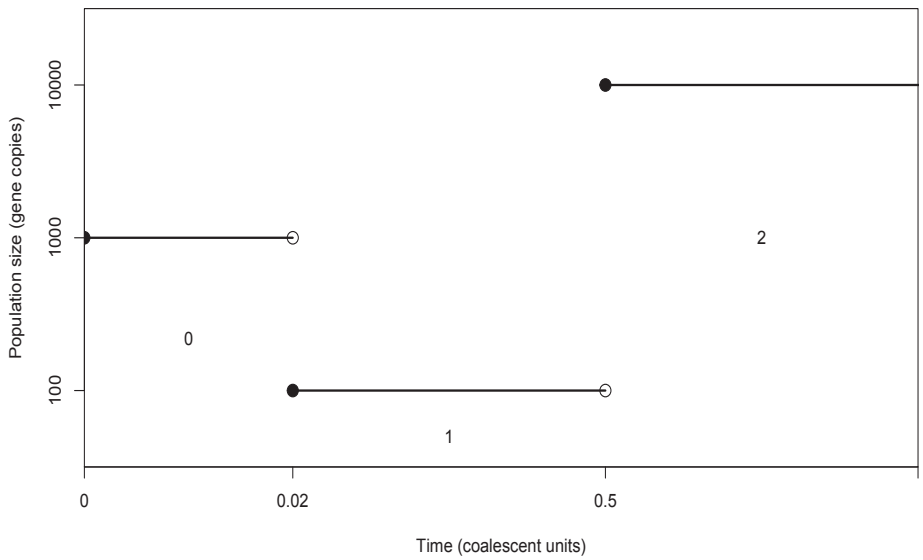


Figure 2.2: Representation of a changing population size $N(t)$, on a log-scale from Equation 2.8. Time is measured in coalescent units of N_0 generations, and the piecewise constant population function has three epochs denoted 0, 1 and 2.

As time goes backwards (moving right along the horizontal axis), the population size decreases to N_1 at time t_1. The population size remains at this value until at time t_2 it jumps to N_2, where it remains. This change in population size (dropping and then rising again) is described as a population bottleneck.

In this example the population sizes are $\boldsymbol{N} = (N_0,\ N_1,\ N_2) = (1000, 100,\ 10000)$ and the time points are $\boldsymbol{t} = (t_0,\ t_1,\ t_2) = (0,\ 600,\ 15000)$ years. These time units are transformed into coalescent units (of N_0 generations) to give $\boldsymbol{t} = (0,\ 0.02,\ 0.5)$.

Next, the function $\lambda(t)$ is introduced, which gives the changing rate of coalescence as a

function of time. It is defined by

$$\lambda(t) = \frac{N(0)}{N(t)} = \begin{cases} \frac{N_0}{N_0} & 0 \le t < t_1 \\ \frac{N_0}{N_1} & t_1 \le t < t_2 \\ \vdots & \vdots \\ \frac{N_0}{N_{b-1}} & t_{b-1} \le t < \infty. \end{cases} \tag{2.9}$$

For this example this gives

$$\lambda(t) = \begin{cases} \frac{N_0}{N_0} = \frac{1000}{1000} = 1 & 0 \le t < 0.02 \\ \frac{N_0}{N_1} = \frac{1000}{100} = 10 & 0.02 \le t < 0.5 \\ \frac{N_0}{N_2} = \frac{1000}{10000} = 0.1 & 0.5 \le t < \infty, \end{cases}$$

illustrated in Figure 2.3. Note that $\lambda(t_0) = \frac{N_0}{N_0} = 1$. When the population size falls below N_0, the rate of coalescence is greater than 1 and when the population size rises above N_0 the rate is less than 1, as in epochs 1 and 2, respectively.

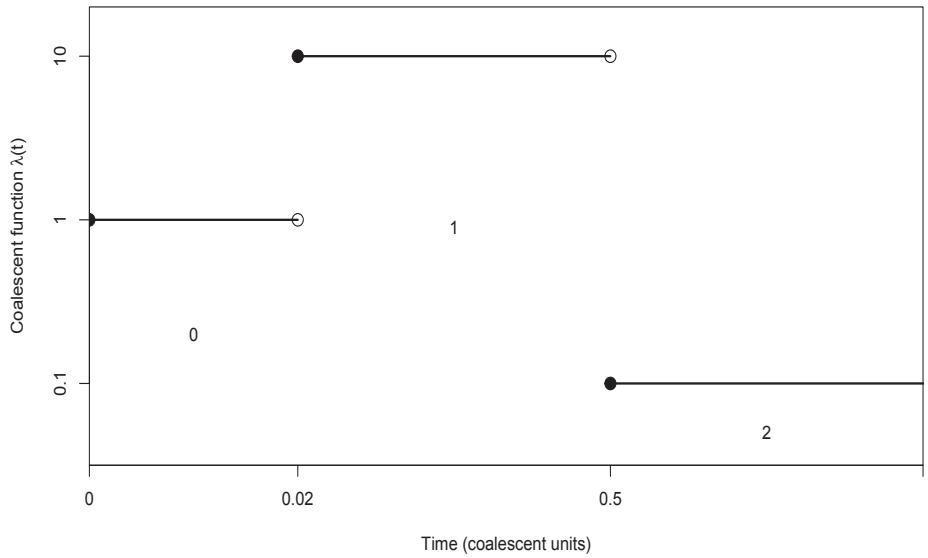


Figure 2.3: Representation of the coalescent rate function $\lambda(t)$ for the example.

Next, $\Lambda(t)$ the population-size intensity function is introduced,

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = \Big[\sum_{j=0}^{i(t)-1} \frac{N_0}{N_j}(t_{j+1} - t_j) \Big] + \frac{N_0}{N_{i(t)}}(t - t_i), \tag{2.10}$$

where $i(t)$ is the epoch that contains t. Figure 2.4 shows this population-size intensity function,

$\Lambda(t)$. Note that the slope of $\Lambda(t)$ in epoch 0 is 1, since the rate of coalescence in epoch 0 is 1. Then in epoch 1, $\Lambda(t)$ is steeper, indicating a higher rate of coalescence, and in epoch 2 $\Lambda(t)$ is less steep since the rate of coalescence is lower.

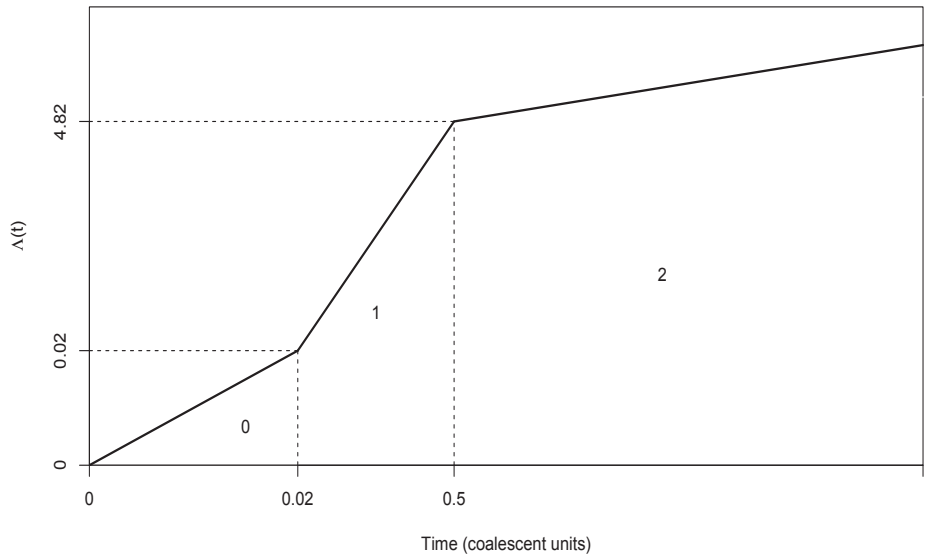


Figure 2.4: Representation of the population intensity function $\Lambda(t)$.

The last step in finding the death process $A^v(t)$ is to find the inverse function $\Lambda^{-1}(t)$ (Figure 2.5), so that the vertical axis gives the new 'warped' time points under the variable population size coalescent. These 'warped' time points are found from

$$\Lambda^{-1}(t) = \frac{N_{i(t)}}{N_0}(t - t'_{i(t)}) + t_{i(t)}, \qquad \text{where } t'_j = \Lambda(t_j). \tag{2.11}$$

As the population size changes from N_0, going back in time, the new warped time points will be 'stretched' or 'squeezed' as appropriate to account for the rising and falling population sizes, respectively.

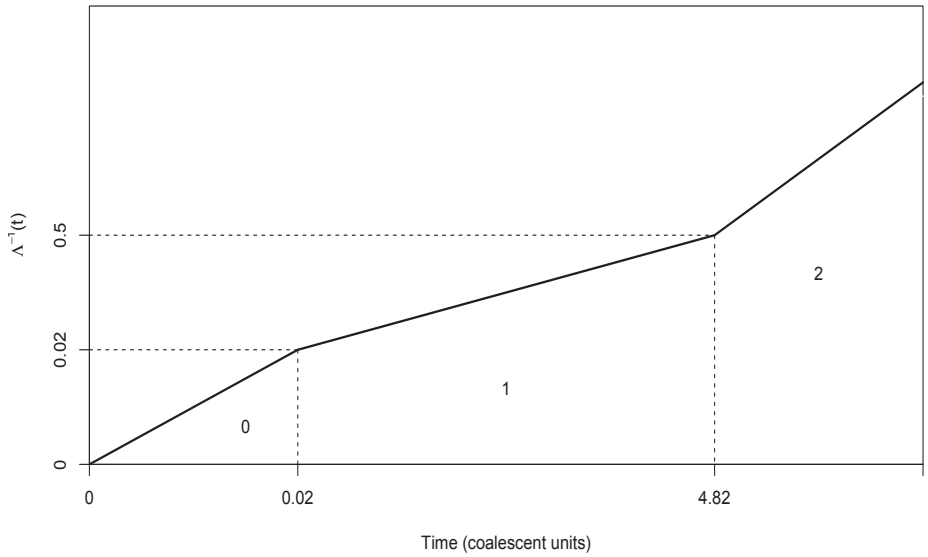


Figure 2.5: Representation of the inverse function $\Lambda^{-1}(t)$ for the potential population trajectory. Time (t) is measured in coalescent units as before.

2.3 Mutation models

In this section, models of nucleotide substitution are presented, following Yang (1993). There are many models that can be used to explain the variation within DNA sequences. However, only three of them will be described here. These three models have been selected because they give a range of options of flexibility of mutation within DNA sequences. The first two models lay the foundation of the third more complicated model which will be the assumed model of mutation in the results chapters later in this thesis.

The genetic distance between two DNA sequences in its simplest form is the proportion of different sites in the sequences. This basic measure of genetic difference can provide information about the number of substitutions in the sequence, but what it usually results in is an underestimation for two reasons: back substitutions at what is believed to be a constant site and multiple mutations at a single site that can exist throughout the genealogy. Both of these types of mutation effectively hide changes in the observed sequence. Figure 2.6 illustrates this.

The five types of substitution that can occur through a genealogy are labelled in blue in Figure 2.6. In this diagram we are interested in a section of the DNA sequence which is eight sites long. A single substitution is a simple change of one nucleotide for another somewhere between the common ancestor who existed some time in the past, and the observed sequence today. In the case of multiple substitutions, there could have been two or more mutations at a specific site in the sequence through the generations from the common ancestor to the observed sequence, but we do not observe the many mutations, only the ones found in the sample. In

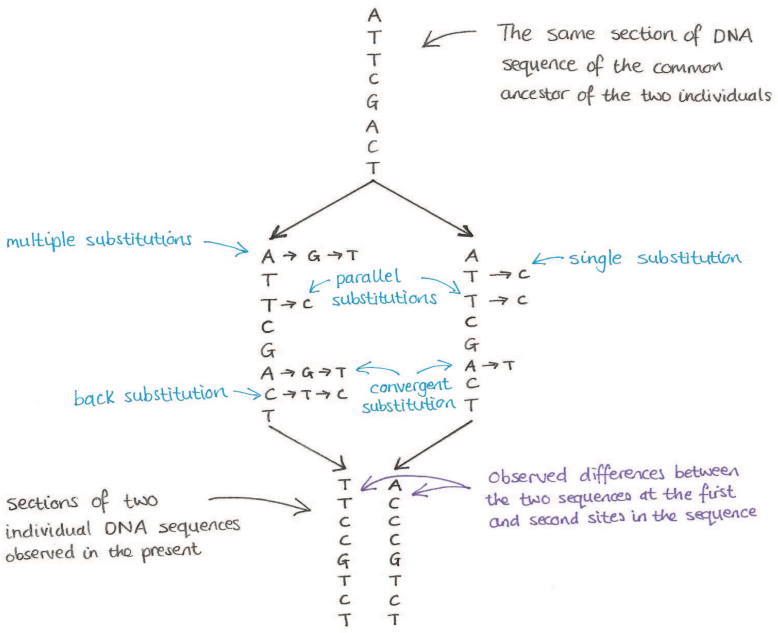


Figure 2.6: Illustration of different types of substitutions to demonstrate how an underestimate of the number of differences between sequences arises. Adapted from Yang (2006).

this example, we observe a case of multiple substitutions in the first site of the sequence of the individual on the left. Looking at the first site (labelled *multiple substitutions*), the A (from site 1 of the ancestor) changes to a G (in site 1 of the left descendant) then to a T so in the sample we only observe the T. A back substitution (site 7 in the left descendant) is similar to a multiple substitution in that a change in nucleotides happened in previous generations, but in the observed sample we note no change between the two sequences (comparing site 7 of the common ancestor to the left observed sequence). In this particular example the observed distance would be recorded as $2/8 = 0.25$ substitutions per site because in the observed sequences there exists only two differences. However, in reality there have been a total of ten substitutions throughout the genealogy (seven on the left and three on the right), which gives a *true* distance of $10/8 = 1.25$ substitutions per site.

Therefore, to account for the unobserved mutation that may have happened in the ancestry of the sample, a probabilistic model is required, one which can account for the changes between nucleotides rather than simply comparing the differences of nucleotides between samples. There are many models that can do this belonging to the class of finite-sites models (Kimura and Crow, 1964).

To model nucleotide substitution at an individual site on the sequence, a Markov chain model is used which allows for multiple mutations at each individual site (Yang, 1996). The states of the chain are the nucleotides, and the chain describes the nucleotide present at a certain position in the sequence at any time. The chain has the *Markovian property*, that it is memoryless. Given the present, the future does not depend on the past, so the probability with which the chain

jumps from one state to another state only depends on the current state and not what happened previously.

Besides the Markovian assumption, the individual nucleotide sites within the DNA sequence are assumed to be evolving independently of one another - what happens at the first site has no impact on what happens at another site in the sequence. Further constraints are often placed on the substitution (or mutation) rates themselves, and this leads to the different substitution models.

These Markov chain substitution models specify the probabilities that each nucleotide would change to any other nucleotide in time δt. Therefore, this requires a transition matrix such as,

$$T = \{p_{ij}\} = \begin{pmatrix} p_{TT} & p_{TC} & p_{TA} & p_{TG} \\ p_{CT} & p_{CC} & p_{CA} & p_{CG} \\ p_{AT} & p_{AC} & p_{AA} & p_{AG} \\ p_{GT} & p_{GC} & p_{GA} & p_{GG}, \end{pmatrix} \tag{2.12}$$

where p_{ij} denotes the probability of moving from state i to state j, i.e. mutation from nucleotide i to nucleotide j in a small time unit δt. There are two constraints on the transition probabilities within the matrix; the first is that each probability must be non-negative, $p_{ij} \geq 0$, and secondly, that all the rows in the transition matrix must sum to 1, $\sum_j^m p_{ij} = 1$. From the above transition matrix, the first row denotes the probabilities of mutating from nucleotide T to any other nucleotide, the second row denotes the probabilities of mutating from nucleotide C to any other nucleotide, and so on. The diagonal entries of this matrix account for no mutation, p_{TT} is the probability of nucleotide T changing to nucleotide T, i.e. no change.

Now, the values within this transition matrix need to relate to the overall mutation rate. The probability of a change, or mutation, from one nucleotide to another in time δt is

$$P(X(t+\delta t) \neq X(t)) = \mu \delta t, \tag{2.13}$$

where μ defines the average nucleotide mutation rate per site.

Each of the mutation models used has a known limiting distribution, $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$, which are the frequencies of each nucleotide. The equation

$$\boldsymbol{\pi} T = \boldsymbol{\pi} \lambda, \tag{2.14}$$

where λ is an eigenvalue of the value 1, gives the limiting distribution π if $Q = \frac{T-I}{\mu \delta t}$ for the identity matrix I. Therefore π is the solution to the equation

$$\boldsymbol{\pi} Q = \mathbf{0}. \tag{2.15}$$

From the probabilities of a change over a small time interval δt, one can obtain the equivalent probability for change over any time $t > 0$. This is denoted by $p_{ij}(t)$ and is the probability that

a given nucleotide i will become j over a period of time t.

The probability of a change from one nucleotide to another in time $t+\delta t$ is

$$\begin{aligned} P_{ij}(t+\delta t) &= \sum_k P_{ik}(t)T_{kj} \\ &= \sum_k P_{ik}(t)\left[I_{kj}+Q_{kj}\mu\delta t\right], \end{aligned}$$

by the law of total probability. Then, the 4×4 matrix of the P_{ij}'s satisfies,

$$\begin{aligned} P(t+\delta t) &= P(t)\left[I+Q\mu\delta t\right] \\ P(t+\delta t) &= P(t)+P(t)Q\mu\delta t \\ \frac{P(t+\delta t)-P(t)}{\delta t} &= P(t)Q\mu, \end{aligned}$$

since $P(t)=\{P_{ij}(t)\}$. $P(t)$ is then the solution to the differential equation

$$\frac{\mathrm{d}P}{\mathrm{d}t} = PQ\mu, \tag{2.16}$$

which is,

$$P(t) = e^{Q\mu t} \tag{2.17}$$

since $P(0)=I$. To find the exponential of the matrix $Q\mu t$, first perform eigendecomposition of Q as $U\Lambda U^{-1}$, where U is a nonsingular matrix and $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$, where λ_1, λ_2, λ_3 and λ_4 are the eigenvalues of Q. The columns of U and rows of U^{-1} are the corresponding left and right eigenvectors of Q, respectively.

Equation 2.17 then yields,

$$P(t) = e^{Q\mu t} = U\mathrm{diag}\{\exp(\lambda_1\mu t),\exp(\lambda_2\mu t),\exp(\lambda_3\mu t),\exp(\lambda_4\mu t)\}U^{-1}. \tag{2.18}$$

This eigendecomposition then gives a matrix of probabilities of moving from one state to another, in terms of the mutation rate. This is used for the simulation of the DNA sequences after the Coalescent Process determines the tree structure. In the following sections, each of the three mutation models will be introduced. Each subsection will discuss the assumptions of each model and the structure of the transition matrix, which, in turn, allows the derivation of the substitution rate matrix, Q.

2.3.1 Jukes and Cantor (1969) model

The Jukes and Cantor (JC69) model (Jukes and Cantor, 1969) is the most basic of mutation models and assumes that every nucleotide has the same rate of changing to any other nucleotide. The equilibrium distribution is therefore $\boldsymbol{\pi}=(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$ since the model is symmetric in per-

mutation of the nucleotide labels, so in equilibrium each nucleotide has an equal probability of appearing in the sequence. This is an unrealistic model of true DNA sequences, but provides the basis of the more complex models.

The probability of a change from one nucleotide to another in time δt is $\mu\delta t$. But, from Equation 2.13, the probability of this change from nucleotide T, say, will be

$$\begin{aligned} P(X(t+\delta t)\neq T|X(t)=T) &= P(X(t+\delta t)=C|X(t)=T)+P(X(t+\delta t)=A|X(t)=T) \\ &\quad + P(X(t+\delta t)=G|X(t)=T) \\ &= m\delta t+m\delta t+m\delta t \\ &= 3m\delta t, \end{aligned}$$

since each nucleotide has the same probability of changing to another (for bases C, A and G, respectively) and m is the mutation rate scaled to coalescent time units and across sites. Setting $3m\delta t=\mu\delta t$ and solving for m gives the following transition matrix for the JC69 model:

$$T_{JC69}=\begin{pmatrix} 1-\mu\delta t & \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t \\ \frac{\mu}{3}\delta t & 1-\mu\delta t & \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t \\ \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t & 1-\mu\delta t & \frac{\mu}{3}\delta t \\ \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t & \frac{\mu}{3}\delta t & 1-\mu\delta t \end{pmatrix}.$$

This is true because of the symmetry of this model, where all the mutation rates are constant. Then, solving for the Q matrix using the equilibrium distribution, $\boldsymbol{\pi}$, and taking out a factor of $\mu\delta t$, this gives,

$$Q_{JC69}=\frac{T_{JC69}-\mathrm{I}}{\mu\delta t}=\begin{pmatrix} -1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & -1 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & -1 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -1 \end{pmatrix}.$$

Eigendecomposition is then carried out on the Q_{JC69} matrix to find the probability of moving from one nucleotide to any other in time using 2.18 and for every site on the DNA sequence. These probabilities are then used to find the DNA sequences of each leaf of the tree found through the Coalescent Process explained in Section 2.5.4.

2.3.2 Hasegawa *et al.* (1985) model

Next is the HKY85 model (Hasegawa et al., 1985). This model allows for more flexibility in mutation rates between different nucleotides, rather than a constant rate as in the JC69 model. Primarily, the HKY85 model accounts for different mutation rates between purines and pyrimidines, as discussed in Section 2.1, since nucleotides T and C are more likely to mutate to one another, as are A and G. The HKY85 model also allows for the nucleotide frequencies to be

non-uniform, different from the JC69 model, so that $\boldsymbol{\pi}$ is not a vector of equal proportions. This accommodates a more realistic representation of DNA sequences since nucleotides do not have equal representations in the sequence. So for this model, the equilibrium distribution is $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$.

Transitions (a substitution between two purines or two pyrimidines) have a rate parameter denoted by α and transversions (a substitution between a purine and a pyrimidine) a rate parameter β. This gives the transition matrix for the HKY85 for a time δt,

$$T_{HKY85} = \begin{pmatrix} 1-\mu_1\delta t & \alpha\pi_C\delta t & \beta\pi_A\delta t & \beta\pi_G\delta t \\ \alpha\pi_T\delta t & 1-\mu_2\delta t & \beta\pi_A\delta t & \beta\pi_G\delta t \\ \beta\pi_T\delta t & \beta\pi_C\delta t & 1-\mu_3\delta t & \alpha\pi_G\delta t \\ \beta\pi_T\delta t & \beta\pi_C\delta t & \alpha\pi_A\delta t & 1-\mu_4\delta t \end{pmatrix},$$

with $\mu_1 = \alpha\pi_C + \beta\pi_A + \beta\pi_G$, $\mu_2 = \alpha\pi_T + \beta\pi_A + \beta\pi_G$, $\mu_3 = \alpha\pi_G + \beta\pi_T + \beta\pi_C$ and $\mu_4 = \alpha\pi_A + \beta\pi_T + \beta\pi_C$.

This matrix is often reparameterised in terms of the so-called the transition-transversion ratio, κ, which leads to the Q matrix given by

$$\begin{aligned} Q_{HKY85} &= \frac{T_{HKY85} - \mathrm{I}}{\mu\delta t\gamma} \\ &= \begin{pmatrix} -\kappa\pi_C + \pi_A + \pi_G & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & -\kappa\pi_T + \pi_A + \pi_G & \pi_A & \pi_G \\ \pi_T & \pi_C & -\kappa\pi_G + \pi_T + \pi_C & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & -\kappa\pi_A + \pi_T + \pi_C \end{pmatrix}, \end{aligned}$$

where γ has been taken out the matrix as a common factor and is given by $\gamma = 2\kappa(\pi_T\pi_C + \pi_A\pi_G) + 2\pi_Y\pi_R$ and $\pi_Y = \pi_T + \pi_C$, the frequency of pyrimidines, $\pi_R = \pi_A + \pi_G$, the frequency of purines and $\kappa = \alpha(\pi_T\pi_C + \pi_A\pi_G)/(\beta\pi_Y\pi_R)$.

Similarly, eigendecomposition, as in Equation 2.18, is then performed on Q_{HKY85} to obtain the probability of moving from one nucleotide to any other in time $t > 0$ at any site on the DNA sequence.

2.3.3 Tamura and Nei (1993) model

Lastly, the TN93 model (Tamura and Nei, 1993) extends the HKY85 model by allowing for a different transition rate parameter for purines and pyrimidines. These are denoted as α_1 and α_2 where α_1 accounts for the transition between pyrimidines ($T \leftrightarrow C$) and α_2 accounts for the transition between purines ($A \leftrightarrow G$). The parameter β, that accounts for transversions, stays as before. With the new parameters α_1 and α_2, the transition matrix for the TN93 model is given

by

$$T_{TN93} = \begin{pmatrix} 1-\mu_1\delta t & \alpha_1\pi_C\delta t & \beta\pi_A\delta t & \beta\pi_G\delta t \\ \alpha_1\pi_T\delta t & 1-\mu_2\delta t & \beta\pi_A\delta t & \beta\pi_G\delta t \\ \beta\pi_T\delta t & \beta\pi_C\delta t & 1-\mu_3\delta t & \alpha_2\pi_G\delta t \\ \beta\pi_T\delta t & \beta\pi_C\delta t & \alpha_2\pi_A\delta t & 1-\mu_4\delta t \end{pmatrix},$$

with $\mu_1 = \alpha_1\pi_C + \beta\pi_A + \beta\pi_G$, $\mu_2 = \alpha_1\pi_T + \beta\pi_A + \beta\pi_G$, $\mu_3 = \alpha_2\pi_G + \beta\pi_T + \beta\pi_C$ and $\mu_4 = \alpha_2\pi_A + \beta\pi_T + \beta\pi_C$. This gives the Q matrix

$$\begin{aligned} Q_{TN93} &= \frac{T_{TN93} - \mathrm{I}}{\mu\delta t} \\ &= \begin{pmatrix} -(\alpha_1\pi_C + \beta\pi_R) & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & -(\alpha_1\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha_2\pi_G + \beta\pi_Y) & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & -(\alpha_2\pi_A + \beta\pi_Y) \end{pmatrix}, \end{aligned}$$

where $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$, as before.

Again, the method of eigendecomposition, as in Equation 2.18, is used to find the probability of any substitution happening over time t for each of the nucleotides (T, C, A and G). In addition, another two parameters are introduced to account for the different ratios of transitions and transversions similar to the HKY85 model. Let $\kappa_1 = \alpha_1\pi_T\pi_C/\beta(\pi_T\pi_A + \pi_T\pi_G + \pi_C\pi_A + \pi_C\pi_G)$ and $\kappa_2 = \alpha_2\pi_A\pi_G/\beta(\pi_T\pi_A + \pi_T\pi_G + \pi_C\pi_A + \pi_C\pi_G)$ so the under the TN93 model, $\kappa = \kappa_1 + \kappa_2$.

As mentioned at the start of this section, there exist other mutation models such as the K80 model (Kimura, 1980), the F81 model (Felsenstein, 1981) and the generalised time-reversible model, known as the GTR model, (Tavaré, 1986; Yang, 1994; Zharkikh, 1994). The K80 model is similar to the HKY85 model in that it accounts for different mutation rates for transitions and transversions but still assumes equal nucleotide frequencies like the JC69 model. The F81 model extends the JC69 model by accounting for different nucleotide frequencies, rather than keeping them all equal (putting $\alpha_1 = \alpha_2 = \beta$ in the TN93 model). The GTR model assumes that, along with different nucleotide frequencies, the rate of mutation between pairs of nucleotides is governed by six rate parameters. This is an extension of the TN93 model, where there are two transition and one transversion parameters, to two transition and four transversion parameters.

Given the complex nature of the data and the relationships one is trying to estimate from these models, it could be expected that the more complex the mutation model the more robust and realistic the estimations of mutation rates. However, as Yang (2006) shows, the estimates of mutation rates are not so different between the different structures of the substitution rate matrices, Q, from each of the models described in this section. In fact, even models such as JC69 and K80, despite being gross simplifications can produce very similar results to more complex mutation models (Yang, 2006). This was observed in the analyses of simulated DNA sequences in this work, as discussed below. It is for these reasons, however, that the three models were

chosen for analyses in this thesis. The less complicated the model, the less the computational expense.

2.4 Rate heterogeneity

One important biological factor in DNA sequences in evolution is the heterogeneity of DNA itself. Indeed, the mutation rates across sites are *not* constant (Yang, 2006). One way to deal with this would be to estimate each site's own mutation rate, but this is difficult to do as there is not enough information in the data. Instead, it is often assumed that the sitewise mutation rate is Gamma distributed independently across sites (Yang, 1993). Furthermore, it is likely that some proportion of sites within in the DNA sequence are *invariant*, meaning that such a site does not mutate (or if it does mutate, the mutant is immediately eliminated by natural selection). This is termed the invariant sites model (Hasegawa et al., 1987).

A model that incorporates both a Gamma distributed mutation rate across variable sites and a proportion of invariant sites in the sequence is labelled the 'I + Γ' model, the Invariant sites plus Gamma model.

2.4.1 Gamma distributed mutation rates

Let μ_s be the mutation rate of site s ($s = 1, 2, \dots$). It is assumed that

$$\mu_s \sim \text{Ga}(\alpha_G, \beta_G), \quad \text{independently.} \tag{2.19}$$

Suppose a calibration of the mean mutation rate (per site) is available, μ, say. Since $\text{E}[\mu_s] = \frac{\alpha_G}{\beta_G}$, given α_G, the scale parameter, β_G, can be obtained from α_G/μ. So the parameter α_G characterizes the rate heterogeneity.

2.4.2 Invariant sites

Certain mutations within the DNA sequence will never be observed. For example, a mutation in a coding sequence could 'break' the protein so that the organism does not survive long enough to pass it on (see the discussion of natural selection in Section 1.1). In this instance, the site can be assumed to be *invariant*. To include the invariant sites model, sites have to be classed as *invariant* or *not invariant*. A new parameter η is introduced to represent the proportion of sites that are invariant. The remaining proportion of sites, $1 - \eta$, are variable, i.e., able to mutate, e.g., with a constant rate or with Gamma distributed rates.

An indicator variable, I_s, is introduced that is independently Bernoulli distributed,

$\mathrm{I}_s \sim \text{Bernoulli}(\eta)$ and

$$\mathrm{I}_s = \begin{cases} 1 & \text{if site } s \text{ is invariant,} \\ 0 & \text{otherwise.} \end{cases}$$

Given $I_s = 1$, $\mu_s = 0$ or equivalently

$$\mu_s | I_s = 1 \sim \delta_0, \tag{2.20}$$

a point mass at zero.

Given $I_s = 0$, $\mu_s = \mu$ (or $\mu_s \sim \delta_\mu$) in the constant rate model or $\mu_s \sim \text{Ga}(\alpha_G, \beta_G)$ in the Gamma distributed model. Then, $\mathrm{E}[\mu_s] = \mathrm{E}\big[\mathrm{E}[\mu_s | I_s]\big] = (1-\eta)\mu + \eta \times 0 = (1-\eta)\mu$, where $\mu = \mathrm{E}[\mu_s | I_s = 0]$.

2.5 Simulation settings

This section will describe the conditions under which DNA sequences are simulated. The demographic models of choice will be described, followed by a detailed discussion of selecting the 'mutation of interest' which defines the non-random sample of DNA sequences. Finally, a detailed description of the algorithm used to generate the DNA sequence sample will be explained.

Note that going forward when referring to the "population size" throughout this thesis, this is the number of reproducing females in the population because the interest is in mtDNA which is only passed on through the female line (see Section 2.1).

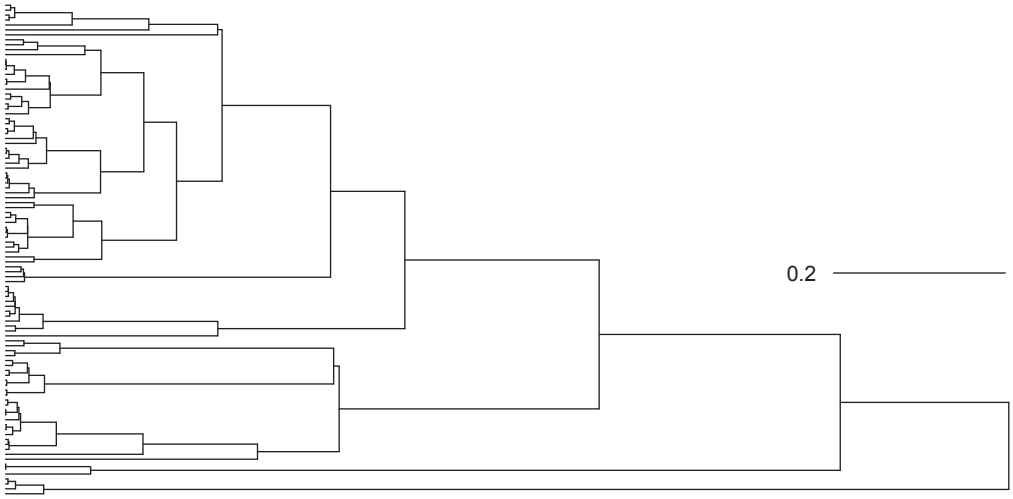
2.5.1 The demographic model

As discussed in Section 2.2.1, the demographic model needs to be chosen before simulation of DNA sequences. Once the demographic model has been chosen the Coalescent Process generates a genealogy, along which the DNA sequences can be simulated.

Two commonly used models for modelling the demography of a population of individual human DNA sequences are the simple constant population size model and the model of exponential population growth (Strimmer and Pybus, 2001). Other population demographic models can be and are used. This thesis will include a population step model and a typical population bottleneck model.

Begin with the constant population size demographic model. This model is the most simplistic and, unrealistically, assumes that the population size has stayed constant through time - from the time of the most recent common ancestor to the present day. Figure 2.7 gives an example of a genealogy simulated under a constant population size Coalescent Process of a sample of $n = 100$ individual DNA sequences with a constant population size of $N_0 = 10,000$.

(a)



(b)

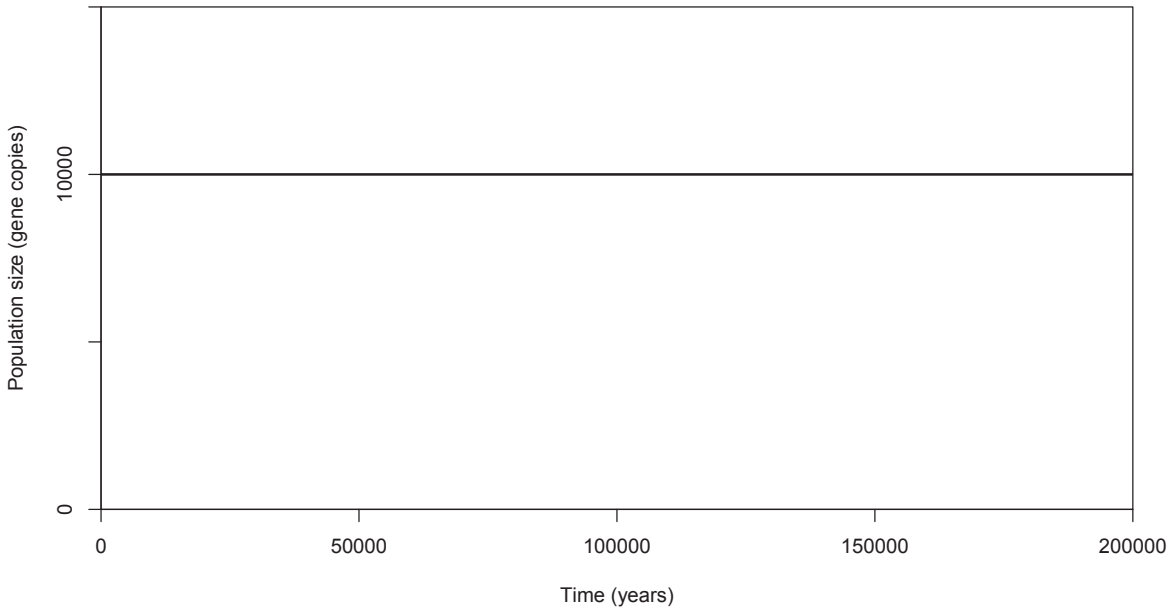


Figure 2.7: Constant population size example. a) A simulated genealogy under the constant population size coalescent, scale bar is in coalescent units of N_0 generations. b) The population size over time, where $N(t) = 10,000$.

The demographic model in this case is

$$N(t) = N_0 = 10,000. \tag{2.21}$$

With a constant population size Coalescent Process, the genealogy looks quite distinct (Figure 2.7(a)). There are lots of coalescent events early (close to the present) in the genealogy, when there are lots of individuals to coalesce. As time moves back towards the root of the tree, the rate of coalescence is slower.

Next is the exponential growth demographic model. An example of this is shown in Figure 2.8. The genealogy of a sample reconstructed from the Coalescent Process under an exponential demographic model is also distinct, featuring a comb-like structure to the branches. This is because as we move backwards in time the population size is decreasing so that we do not have to wait as long for a coalescent event, as was the case in the constant population size model. This results in long terminal branches and shorter internal branches.

The genealogy for the exponential growth population model can be simulated from a genealogy for a constant size model by time-warping as discussed in Section 2.2.1. The details for this case follow.

For the exponential population growth model

$$N(t) = N_0 e^{-rt}, \tag{2.22}$$

where r is the rate of exponential growth, N_0 is the initial population size and the negative sign in the exponential function represents the backwards time scale. Next, $\lambda(t)$, the instantaneous coalescent rate per $N(0)$ generations, as defined in Equation 2.9 is

$$\lambda(t) = e^{rt}. \tag{2.23}$$

Then Equation 2.10 gives

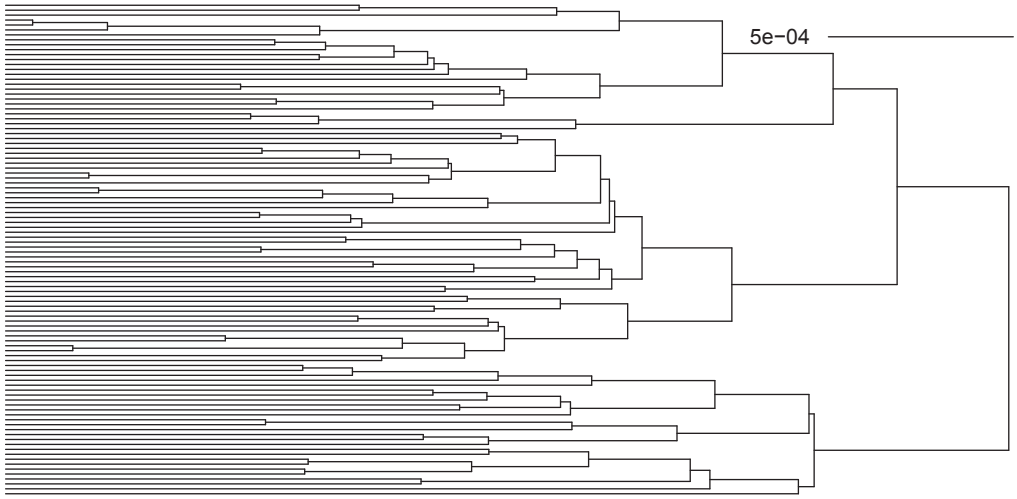
$$\Lambda(t) = r^{-1}(e^{rt} - 1), \tag{2.24}$$

and the inverse function,

$$\Lambda^{-1}(t) = r^{-1}\log(1 + rt). \tag{2.25}$$

This function, $\Lambda^{-1}(t)$ is applied to the times of constant coalescent events, to provide a set of the times of coalescent events under the exponential growth model.

(a)



(b)

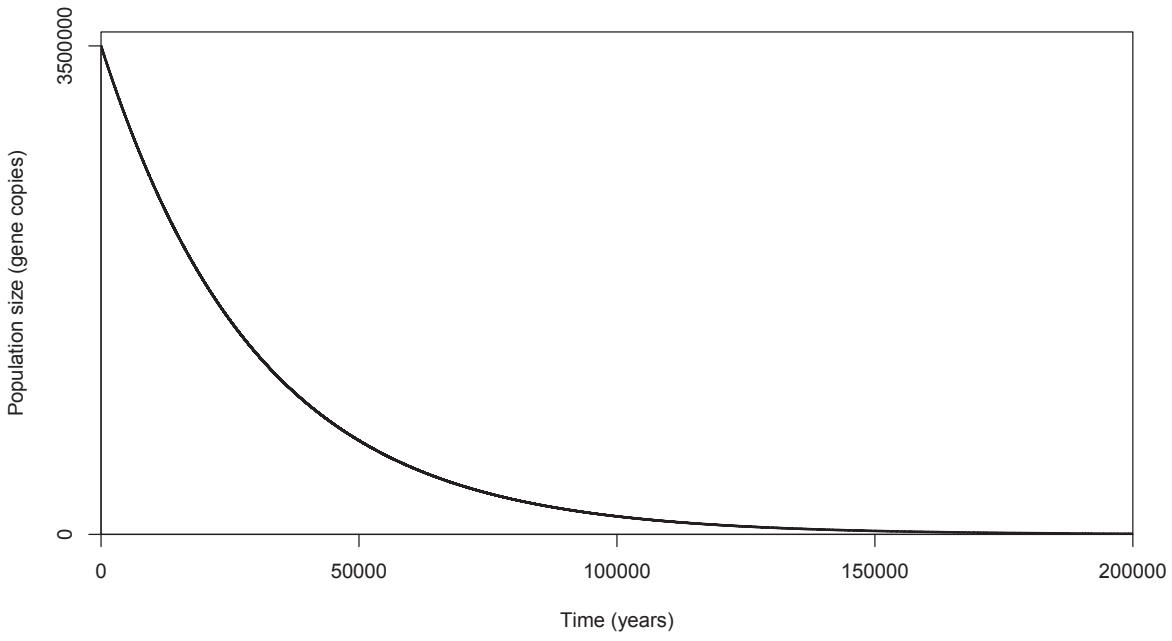


Figure 2.8: Exponential population growth example. a) A simulated genealogy under the exponential growth coalescent. The scale bar shows time in coalescent units. b) The population size over time, where $N(t) = N_0 e^{-rt}$, where $N_0 = 3.5 \times 10^6$ and $r = 3.299 \times 10^{-5}$ per year.

Next, the step and bottleneck demographic models are shown. Figure 2.9 illustrates a step in population size and the associated genealogy produced from the Coalescent Process. In this model, the population size function is

$$N(t) = \begin{cases} 100,000 & 0 \leq t < 60,000 \text{ years}, \\ 10,000 & 60,000 \leq t < \infty \text{ years}. \end{cases}$$

As shown in Figure 2.9, the population size function has two epochs, in the first the population size is 100,000 until 60,000 years ago where it drops to 10,000. The reason for choosing this step demographic model is to reflect the hypothesis that humans expanded rapidly in size after the out-of-Africa event, approximately 60,000 years ago (Soares et al., 2011).

Lastly, a population bottleneck will be considered. Figure 2.10 shows an example of a genealogy from the Coalescent Process assuming a bottleneck demographic model.

The population size function for this demographic model is

$$N(t) = \begin{cases} 100,000 & 0 \leq t < 23,000 \text{ years}, \\ 1,000 & 23,000 \leq t < 26,500 \text{ years}, \\ 10,000 & 26,500 \leq t < \infty \text{ years}. \end{cases} \tag{2.26}$$

The changes in population size were chosen for this model to correspond to the hypotheses of the times and impacts of the Last Glacial maximum, (Clark et al., 2009), a period of time where the ice sheets were at their greatest extent. As a result, human population sizes would have undoubtedly been influenced. The population function (Equation 2.26) attempts to reflect this.

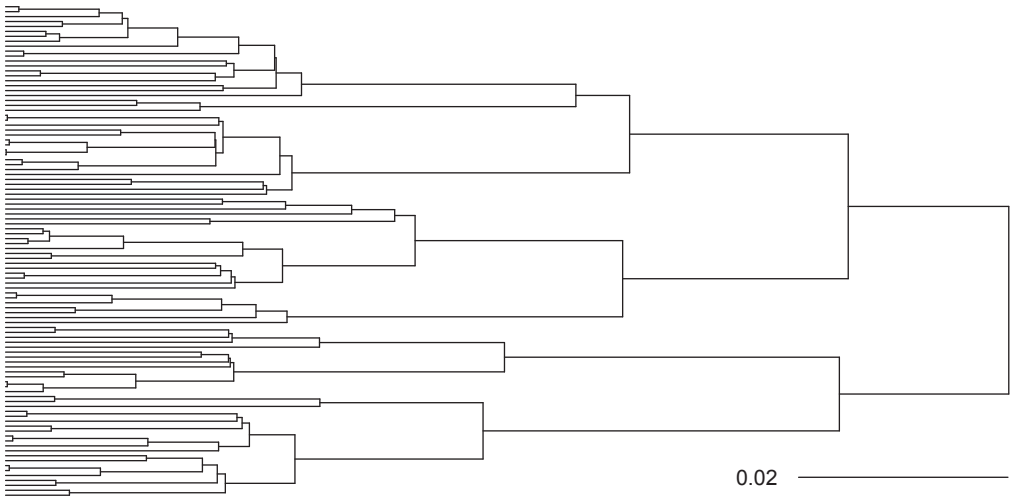
The dramatic change in population size at 23,000 years ago results in lots of coalescent events in the genealogy. As in the cases of the exponential and step demographic models, the times of coalescent events can be obtained by time-warping. Assuming each of these four demographic models, DNA sequences will be simulated under the Coalescent Process and then used for inferring population sizes as a function of time.

2.5.2 Determining the non-random sample

To investigate any bias in non-random sampling of DNA sequence data, a *random* sample of DNA sequences was simulated first. The *non-random* sample was then a subset of this random sample of DNA sequences. To differentiate between these two samples, n is used to denote the number of *randomly* sampled DNA sequences, and n_s denotes the size of the subsample of *non-randomly* sampled DNA sequences.

To select the subsample of DNA sequences, chose to select a branch in the genealogy that would be long enough to contain at least one mutation in expectation. This is to reflect the idea

(a)



(b)

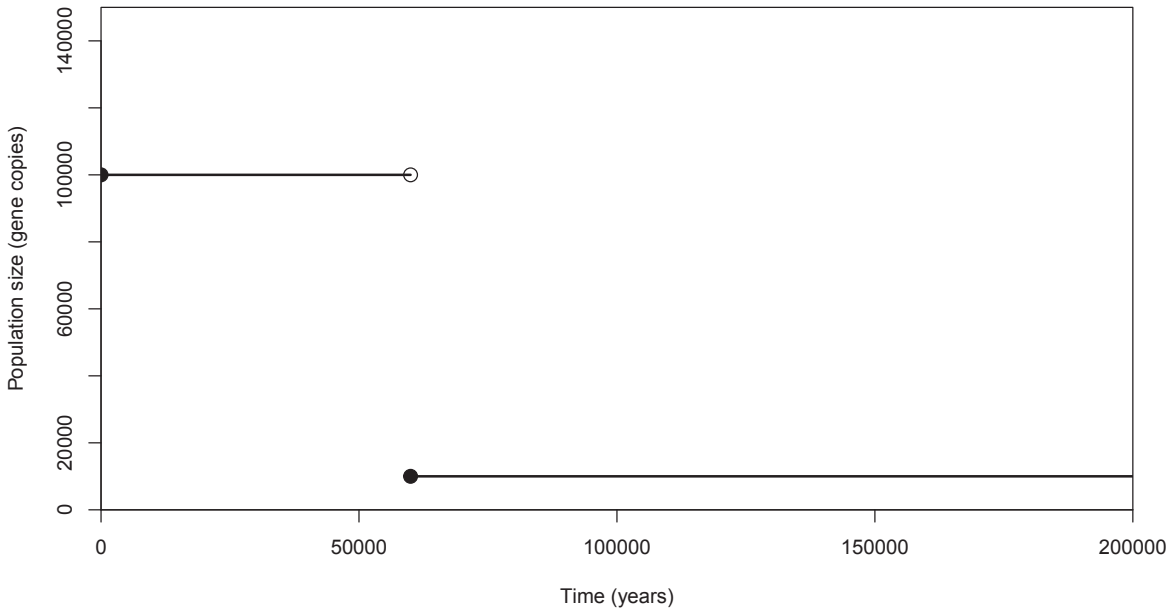
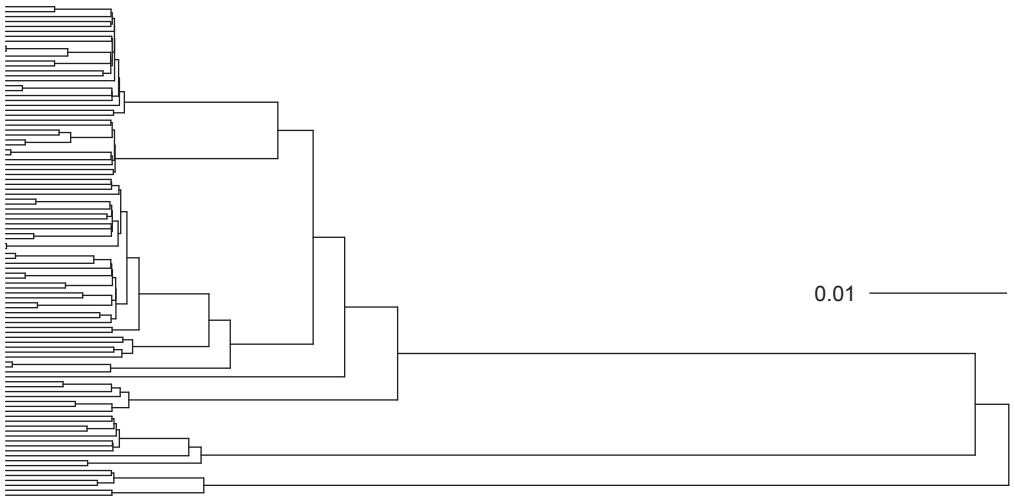


Figure 2.9: Step demographic model example. a) A simulated genealogy from the Coalescent Process. b) Population size over time.

(a)



(b)

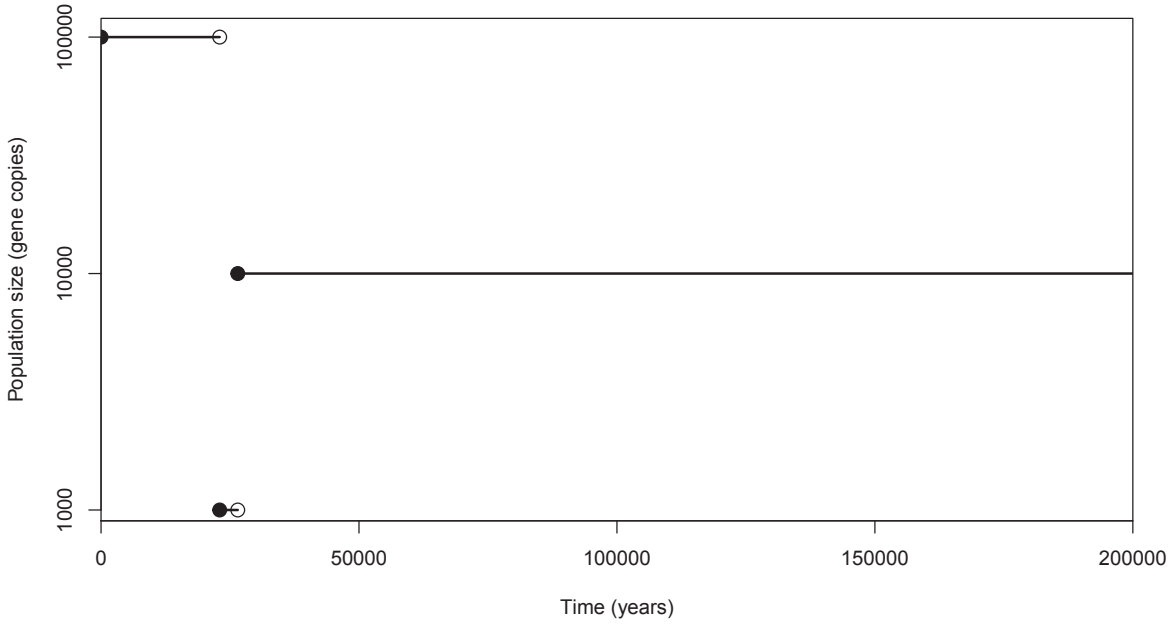


Figure 2.10: Bottleneck population size example. a) A simulated genealogy under this demographic model. b) Population size over time.

that the subsample might be determined by a so-called 'mutation of interest', e.g., a mutation defining a mtDNA haplogroup. This minimum branch length is denoted by ε and is calculated as follows.

Suppose the mutation rate of a full sequence per coalescent unit of time is $\frac{1}{2}\theta$ (the $\frac{1}{2}$ is conventional). With $\theta = 2\mu l$ for μ the average mutation rate per site and l the length of the sequence. The number of mutations K on an edge of the genealogy of length t is often modelled as

$$K \sim \text{Po}\Big(\frac{1}{2}\theta t\Big), \tag{2.27}$$

with $\text{E}[K] = \frac{1}{2}\theta t$. Equating this to 1 when $t = \varepsilon$ gives $\varepsilon = \frac{2}{\theta}$.

Any branch in the genealogy that is as long as or longer than this length ε is then a candidate for the branch containing the 'mutation of interest'. The set of possible branches is collected and then one is randomly chosen and the genealogy *cut* on that branch. The descendants then make the subsample of DNA sequences and since they all contain a common mutation (in expectation), they are a non-random set of sequences. Since θ depends on the initial population size, this value of ε changes depending on the underlying demographic model.

2.5.3 Defining parameters

The Ingman dataset consists of 53 complete mtDNA sequences from 53 humans of diverse origins (Ingman et al., 2000). This was used to estimate parameter values for the mutation models, the initial population size for the constant and exponential demographic models and the parameters in the $\Gamma + I$ model.

The parameters to be chosen are listed below.

- The sample size of DNA sequences, n.
  This was taken to be $n = 100$. This is primarily to allow the simulations to be carried out in a reasonable amount of time, alongside the need for the subsample (which will be from the full sample, n) to be large enough to ensure that there is enough information from DNA sequences to appropriately compare the full and subsamples. In the literature, sample sizes range from quite small, at around $n = 20 - 50$, (Ingman et al., 2000; Wiuf, 2001), to $n > 100$ and $n > 1000$, (Atkinson et al., 2007; Cann and Wilson, 1983; Rito et al., 2013; Soares et al., 2011; Vijayraghavan et al., 2018).

- The initial population size, N_0.
  The values of the initial population size changes depending on the demographic model assumed and are described in Section 2.5.1.

- The average nucleotide mutation rate per site, μ.
  This was assumed to be $\mu = 1.26 \times 10^{-8}$ (Mishmar et al., 2003) per site per year.

- The number of years per generation, g.
  The number of years per generation is taken to be the difference between the average age of mothers at the birth of their first child and the average age of mothers at the birth of their last child. The number of years per generation was taken to be $g = 30$ (Fenner, 2005; International Society of Genetic Genealogy, 2015).

- The length of the DNA sequence, l.
  The length of the DNA sequence was $l = 15,446$ which is the length of the coding region of mitochondrial DNA (Anderson et al., 1981).

- The shape parameter of the Gamma distribution in Section 2.4.1, α_G.
  The shape parameter of the Gamma distribution, α_G takes different values under each of the different mutation models (JC69, HKY85 and TN93) and it also depends on whether or not the Invariant Sites model is included. This was estimated from the Ingman data set.

- The proportion of invariant sites as described in Section 2.4.2, η.
  As with the parameters of the Gamma distribution, this parameter was estimated from the Ingman data set and also changes depending on the underlying model.

- The equilibrium frequencies of each of the base pairs, $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$.
  The base frequencies, $\boldsymbol{\pi}$, also depend on what models (mutation models and the Gamma and Invariant sites models) are being assumed when simulating the DNA sequences. Under the JC69 model, the base frequencies are equal, $\boldsymbol{\pi} = (1/4, 1/4, 1/4, 1/4)$ but under the HKY85 and TN93 models, they are not equal, with the bases C and A being more frequent, then the base T, with the base G being the least frequent.

- The transition-transversion ratio, κ under the HKY85 model and κ_1 and κ_2 under the TN93 model.
  The transition transversion ratio, κ is estimated from the Ingman data for the HKY85, as are κ_1 and κ_2 for the TN93 model. Again, the values for these parameters all depend on whether or not the Gamma and/or the Invariant sites models are being assumed when simulating the DNA sequences.

- The minimum length of branch containing the 'mutation of interest', ε.
  The value of ε was chosen as explained in Section 2.5.2.

2.5.4 Algorithm for simulating DNA sequences

To simulate the DNA sequences in conditions that represent 'real life' as much as possible, the steps are as follows. The simulation of DNA sequences was carried out in `R` (R Core Team, 2018), using Version 5.1 of the `ape` package (Paradis et al., 2004).

1. First of all, set the parameter values listed in Section 2.5.3. This step also includes choosing the appropriate mutation model, and specifying the demographic model.

2. Simulate the Coalescent Process to produce a genealogy for the random sample of size n DNA sequences.

   (a) Find the waiting times between coalescent events drawn from Exponential distributions, with rate parameters set to the respective rate of coalescence as shown in Section 2.2.

   (b) Find the times of coalescent events by summing the waiting times.

   (c) i. If using the constant population size demographic model, these coalescent times are used.

   ii. If using the exponential population growth demographic model, the coalescent times are warped (Section 2.5.1) to give the appropriate warped coalescent times under the exponential growth model.

   iii. If using a varying population demographic model, the coalescent times are warped (Section 2.2.1) to give the appropriate warped coalescent times, under different population sizes.

   (d) At each coalescent event randomly select two individuals in the sample to coalesce, remove these two individuals from the sample and replace with their ancestor until there is only one individual left. This is the Most Recent Common Ancestor (MRCA).

   (e) Calculate the length of branches in the genealogy by subtracting relevant coalescent event times.

3. The `ape` package in `R` can be used to visualise the genealogy (if the topology and branch lengths are coded in a 'phylo' object) under the Coalescent Process and using the simulated branch lengths. This is the simulated genealogy and DNA sequences will be simulated at each node in the tree.

4. Next, the branch containing the 'mutation of interest' has to be selected (Section 2.5.2). Starting at the root of the tree, the algorithm moves systematically down the tree until it finds a branch with a length meeting all the requirements (Section 2.5.2). The sample of individuals at the leaves of the subtree defined by that branch are taken as the non-random subsample of DNA sequences.

5. Next, randomly generate a DNA sequence. A vector of length l the length of the sequence is created and each position in the vector represents a site in the sequence. Moving through each site, a nucleotide is randomly chosen from the set (T,C,A,G) with probabilities $(\pi_T,\pi_C,\pi_A,\pi_G)$. This will be the DNA sequence of the MRCA, the root of the tree.

6. Lastly, the simulated DNA sequence of the MRCA is 'evolved' throughout the genealogy, using the mutation model to create the matrix giving the probability of each nucleotide changing to another (Section 2.3). This evolution algorithm follows a series of sequential steps as follows.

   (a) Firstly, create a vector of length l to store the Gamma distributed mutation rates μ_s for each site in the sequence.

   (b) Starting at site 1 in the sequence the indicator variable I_1 is drawn from a Bernoulli distribution and one of two events happens.

   - If the site is invariant ($I_1 = 1$) then the mutation rate at this site is zero and the two descendants of the MRCA inherit her nucleotide at site 1, their descendants inherit the same nucleotide at site 1, and so on throughout the tree. This results in every node and leaf in the tree having the same nucleotide as the MRCA at site 1.
   - If the site is not invariant ($I_1 = 0$) then the mutation rate at this site is non-zero and will be either
     - constant across sites and so the mutation rate is μ, or
     - Gamma distributed across sites, in which case the mutation rate is μ_1.

     Two transition matrices are calculated using the mutation rate and the respective branch lengths joining the MRCA to each of her descendants. The resulting nucleotide in site 1 of each descendant is then sampled from the set of nucleotides (T,C,A,G) with the relevant row of the transition matrix.
     This process is then repeated throughout the tree (keeping the mutation rate constant but calculating different transition matrices for each branch), finding the descendants of each node until the leaves of the tree are reached.

   (c) Once all the nucleotides at site 1 are found, the process then repeats from part (b) moving through each site in the sequence until a full set of DNA sequences are collected from the leaves of the genealogy.

Once the DNA sequences have been simulated under the desired parameter settings, demographic and mutation models, the data are then ready to be analysed. The population size as a function of time is what is to be estimated. Chapter 3 introduces the Bayesian Skyline Plot model as a way to do this.

Chapter 3

Methodology of the Skyline Plot Models

This chapter introduces the Skyline Plot model. This model describes random samples of DNA sequences and permits inference of population size, going back in time to the estimated time of the most recent common ancestor of the locus. To investigate the consequences of non-random sampling of DNA sequences defined by a mutation of interest, one approach is to compare the demographic parameter estimates between the full *randomly* sampled set and the *non-randomly* sampled set of DNA sequences. This can be done using the Skyline Plot model. The analyses in this thesis were carried out using the methods of the Bayesian Skyline Plot model (Drummond et al., 2005). This model was built extending two previous models; the Classic Skyline Plot model (Pybus et al., 2000), and the Generalized Skyline Plot model (Strimmer and Pybus, 2001). All three of these models will be described in this chapter.

As discussed in Chapter 1, DNA sequences hold information about their ancestors and the demography of the population from which they were sampled. Exploring the demography of a population can provide some understanding of the history of that population, including characteristics such as the size of the population. When talking about a past population size in these demographic models we usually refer to the *effective* population size. This is defined to be the size of an idealized population, one which displays similar properties, e.g., of genetic diversity, to a population obeying the Wright-Fisher model (and its assumptions including random mating, non-overlapping generations and constant population size through time) and is denoted N_e. Understanding the relationship between N_e and census population size can be challenging. However since data *simulated* from the Wright-Fisher model (or rather the Coalescent Process) will largely be treated here, the distinction can be side-stepped.

Models for estimating a demographic history from a sample of DNA sequences are mostly based on coalescent theory, (Section 2.2), where the rate of coalescence is intrinsically linked to population size. So, if one can estimate the rate of coalescence in the genealogy of a sample then inference can likely be drawn about the population size. One such method proposes the use of lineages-through-time (LTT) plots that display the rate of coalescence in a reconstructed genealogy through time (Nee et al., 1995). This is done by plotting the times until lineages

coalesce and from this inferring the associated population size via an estimated coalescence rate. Unfortunately this model assumes a known demographic model and tests the hypothesis that the population actually follows that model, which is problematic since the true past population size is unknown. Another method estimates the history of a population through an observed distribution of pairwise differences, (Polanski et al., 1998). This method assumes the population size is monotonically increasing. However, this method has been criticised (Felsenstein, 1992), since it ignores any information from the genealogy. Perhaps more seriously, neither of these methods provide any measure of uncertainty around the population size estimates.

This prompted the creation of the Skyline Plot model (similar to LTT plots). This method used for estimating a demographic history from a sample of DNA sequences relies on a simple parametric model, $N(t)$, which describes the population size through time. As before, time at the present is indicated by t being zero, and increases back into the past, to the most recent common ancestor of the segment of the genome under analysis, or to the root of the genealogy. Two simple models are commonly used to describe the demography of a population: a constant population size and a population size featuring exponential growth forward in time (as discussed in Section 2.4.2). With the constant population size assumption, $N(t) = N_0$ there is one demographic parameter, N_0. With the exponential growth case, $N(t) = N_0 e^{-rt}$, there are two demographic parameters, r the exponential growth rate and N_0, the present day population size. However, both of these models are very simplistic and there is little evidence to assume either of these demographic models for humans. Thus, a more flexible model is desirable.

The Skyline Plot model gets its name from the output of the model, a plot made of piecewise constant estimates of the population size through time creating a graphic not dissimilar to a city skyline. This output is a reconstruction of a variable population size of past populations given a sample of DNA sequences. To obtain the population size estimates the model requires information from both the variability between DNA sequences in a sample and for the Classic and Generalized models the reconstructed genealogy of those samples.

There are three distinct and useful comparisons that could be made using the Skyline Plot model with DNA sequences simulated with the methods described in Chapter 2. These are:

1. to compare estimates of population size of a set of *randomly* sampled DNA sequences with the true demographic model from which the DNA sequences were simulated;

2. to compare the estimates of population size of a set of *non-randomly* sampled DNA sequences with the true demographic model from which the DNA sequences were simulated; and

3. to compare the estimates of population size of the *randomly* sampled DNA sequences with the *non-randomly* sampled DNA sequences.

The following sections will describe the Classic, Generalized and Bayesian Skyline Plot models. Each section will describe the theory of each model and show an example of the model

output.

3.1 The Classic Skyline Plot Model

Prior to the development of the Classic Skyline plot model, methods had been developed to use information in the observed DNA sequences to infer historic population sizes. One possible approach is to estimate the likelihood of the demographic hypothesis given the observed DNA sequences. This was difficult because to find the likelihood of the demographic hypothesis for the sampled DNA sequences one has to integrate over the space of all possible trees. Given a sample size, n, of DNA sequences, the number of possible bifurcating rooted trees that could represent the relationship between ancestors back to the most recent common ancestor (MRCA), (Felsenstein, 2004), is given by

$$n_{trees} = \frac{(2n-3)!}{2^{n-2}(n-2)!}, \tag{3.1}$$

which increases rapidly with sample size n. Although this likelihood can be estimated using Monte Carlo estimation, it becomes computationally intensive and expensive (Griffiths and Tavaré, 1994; Kuhner et al., 1998). The classic skyline plot model attempts to overcome this problem by taking the genealogy to be the genealogy which maximises the likelihood (under some mutation model).

Suppose this maximum-likelihood genealogy $\hat{G}$ has been found from a set of n DNA sequences under some mutation model and assuming the molecular clock hypothesis (Zuckerkandl and Pauling, 1965), giving estimates for the times of the interval nodes of the tree.

Then, $\hat{G}$ determines the set of $k-1$ ordered intervals between coalescent events (coalescent intervals) $I_n, I_{n-1}, \ldots, I_2$ where the subscript here denotes the number of lineages present during each interval. The derivation of interval length I_k is represented by $\hat{w}_k$, the estimated waiting time between coalescent events (as in Section 2.2). In the Coalescent Process the waiting times W_k are independently exponentially distributed with rate λ_k. This gives the expected waiting time (in generations) to be

$$\mathrm{E}(W_k) = \frac{1}{\lambda_k} = \frac{N_0}{\binom{k}{2}}, \quad k = n, \ldots, 2. \tag{3.2}$$

Now, the flexible demographic model $N(t)$ assumes that the population size during each interval I_k is a local constant M_k. This piecewise constant function, $N(t)$, has $n-1$ independent variables $M_n, M_{n-1}, \ldots, M_2$. Setting the expected waiting time to the next coalescent event (Equation 3.2) equal to $\hat{w}_k$ and solving for M_k (replacing N_0 in 3.2) gives a method-of-moments estimator $\hat{M}_k$ for the population size during each interval I_k:

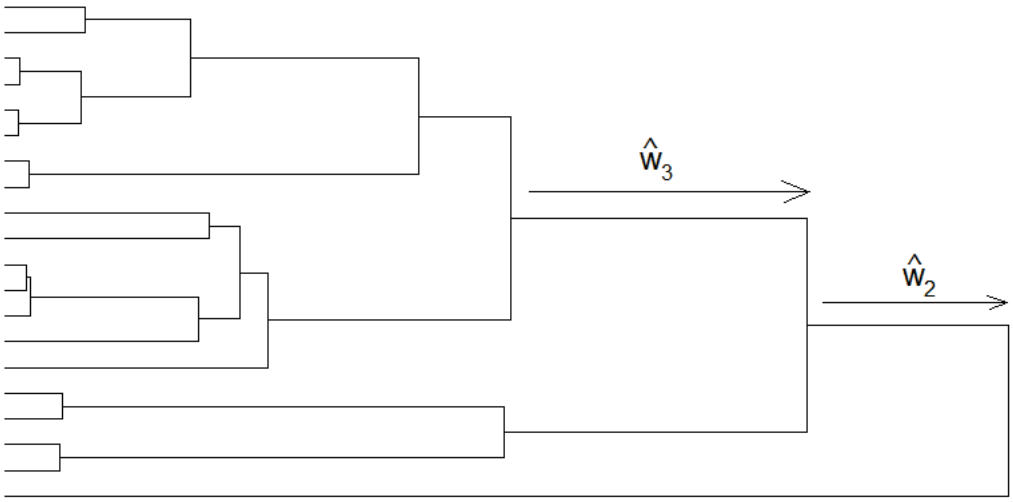
$$\hat{M}_k = \hat{w}_k \frac{k(k-1)}{2}, \quad k = n, \ldots, 2. \tag{3.3}$$

Equation 3.3 gives the Classic Skyline plot estimate for the population size during time interval I_k (Pybus et al., 2000). An example of the Classic Skyline plot is shown in Figure 3.1 for a small sample of $n = 20$ observed DNA sequences.

Each time there is a coalescent event, there is a piecewise change in population size. The waiting time between events is related to the population size: the longer waiting times correspond to a higher population size, because genetic drift occurs more slowly in larger populations. This is particularly obvious closer to the root of the tree, when there are fewer coalescent events, compared to nearer the present day, where there are lots of coalescence events happening and, as a result, lots of changes in population size. The waiting time $\hat{w}_3$ is much longer than the waiting time $\hat{w}_2$, indicated in Figure 3.1, and the resulting skyline plot estimates a larger population size for the longer time interval and a smaller population size for the shorter.

One clear limitation of the approach is the assumption made around the genealogy of the sample. As much as the *most likely* genealogy is assumed this is not the true genealogy. There will ultimately be some uncertainty (in the genealogy) unaccounted for as a result, and the development of the skyline plot model will address this clear issue in Section 3.3.

(a)



(b)

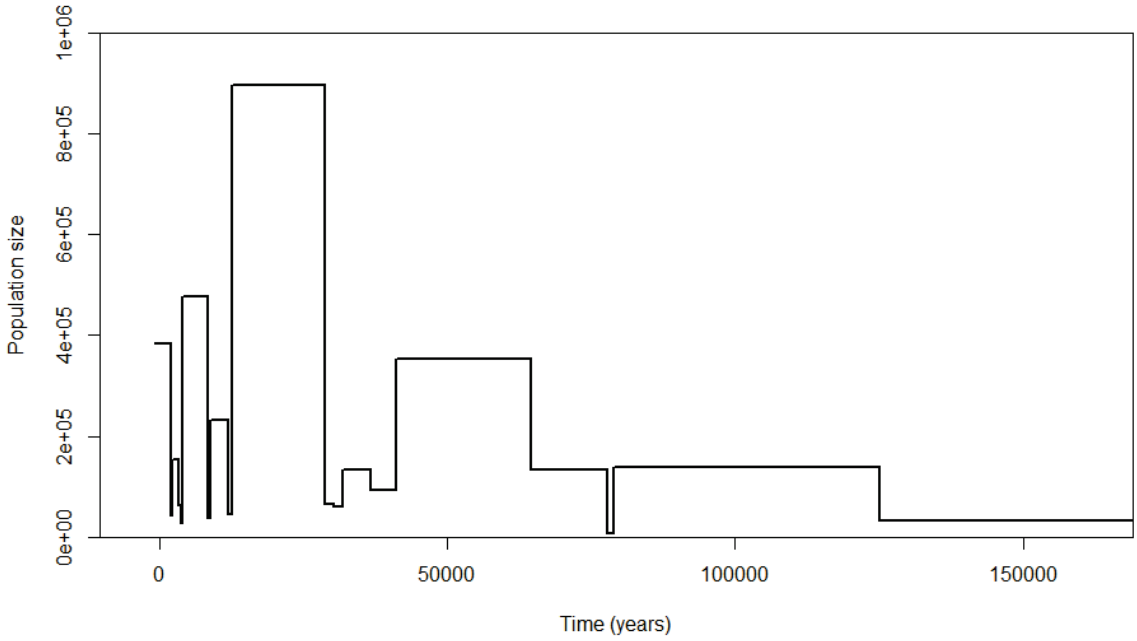


Figure 3.1: Example of Classic Skyline Plot. a) A genealogy for a sample of $n = 20$ DNA sequences and b) the reconstructed skyline plot.

3.2 The Generalized Skyline Plot Model

Very quickly after the development of the Classic Skyline plot model, it was improved to the Generalised Skyline plot model (Strimmer and Pybus, 2001). As with the Classic Skyline plot model, this model provides a flexible estimate of the historic population size of a sample of DNA sequences based on the sample's reconstructed genealogy. This model expands on the Classic Skyline Model by allowing *grouping together* of coalescent intervals, particularly useful for coalescent intervals that are short.

The durations of short intervals are poorly estimated, since they will contain no or very few mutations, so their associated population size estimates will be very noisy. Pooling adjacent coalescent intervals dampens some of that noise. A small-sample Akaike information criterion is used to choose the optimal grouping strategy objectively. The one disadvantage to this method is that some of the temporal structure in the data is ultimately lost.

The Generalised Skyline plot model is defined as follows. Here, the time interval (the waiting time between coalescent events) becomes $I_{k,e}$ being the interval with k lineages at the start (near the present) of the interval and e is the total number of coalescent events taking place during the interval. This interval, $I_{k,e}$ has an observed duration $\hat{w}_{k,e} = \hat{w}_k + \hat{w}_{k-1} + \cdots + \hat{w}_{k-e+1}$ and corresponding random duration $W_{k,e}$ (note that e implicitly depends on k). Assume a locally constant population size $M_{k,e}$ during this interval, then a method-of-moments estimator for $M_{k,e}$ is

$$\hat{M}_{k,e} = \hat{w}_{k,e}\frac{k(k-e)}{2e}, \tag{3.4}$$

which follows since

$$\mathrm{E}[W_{k,e}] = M_{k,e}\frac{2e}{k(k-e)}.$$

Equation 3.4 gives the Generalized Skyline Plot estimate for interval $I_{k,e}$. It should be noted the Classic Skyline plot is a special case when all the e's are 1.

To choose which adjacent coalescent intervals within the genealogy, $\hat{G}$, should be *pooled* together, the following steps are worked through.

1. First, generate the duration of the standard internode intervals, $I_n, I_{n-1}, \ldots, I_2$, from the genealogy, $\hat{G}$.

2. Choose a threshold parameter ζ: any interval shorter than ζ is considered as *small* and is pooled with the interval above it in the genealogy closer to the root.

3. Working systematically through the genealogy from present back to the root, looking at each interval I_n to I_2, each *small* interval is pooled with its neighbouring interval, the one which lies further back in time. If this pooled interval is similarly *small*, then pooling continues until the multiple pooled intervals together form an interval longer than ζ.

This leaves the parameter ζ controlling how much temporal structure from the data is preserved and ultimately the degree of smoothness in the Generalized Skyline Plot. There are two contrasting factors to bear in mind when choosing the value of ζ.

- It should be large enough to remove noise coming from the randomness of the mutational process.
- It should be small enough to capture the time-varying demographic signal in the data.

Clearly, one method of choosing ζ would be to visually inspect population plots with different values of ζ and select the value that subjectively provides the appropriate level of smothing A more objective method is desirable and so Strimmer and Pybus suggest one such method based on statistical model selection. The log-likelihood function of a generalised skyline plot derived from a genealogy, $\hat{G}$, with n leaves, evaluated at the estimated values of the $M_{k,e}$ population size parameters is

$$\log(L) = \sum_{i=2}^{n} \left[\log\left(\binom{i}{2} / \hat{M}_i \right) - \binom{i}{2} \hat{w}_i / \hat{M}_i \right], \tag{3.5}$$

where $\hat{M}_i$ is equal to the value of the estimated population size of the pooled interval that contains the i^{th} coalescent interval.

Let P denote the number of inferred parameters, i.e., the number of composite intervals in the skyline plot, and let $S = n - 1$ denote the number of coalescent events in the genealogy. To compare skyline plots with various values of ζ, the log-likelihood function (3.5) is penalised using the AIC_c correction, which penalises skyline plots that overfit the data. AIC_c is an extension of the well-known and commonly used AIC criterion (Akaike, 1974) which seeks to minimise:

$$\text{AIC} = 2P - 2\log(L), \tag{3.6}$$

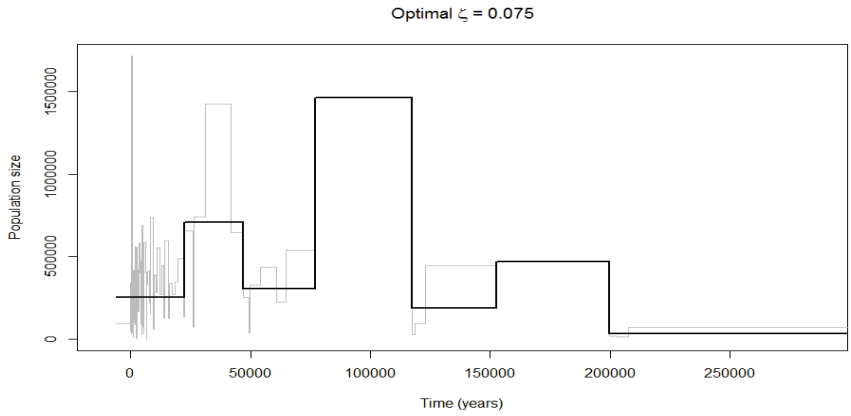
which is valid for larger sample sizes and when the ratio of the sample size over the number of inferred parameters (S/P) is greater than around 40. AIC_c is given by

$$\text{AIC}_\text{c} = \text{AIC} + \frac{2P(P+1)}{S-P-1}, \tag{3.7}$$

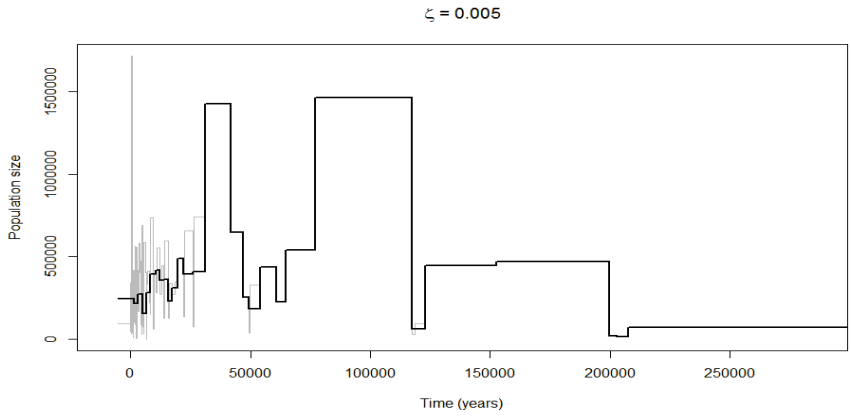
and is valid for small sample sizes (Burnham and Anderson, 1998) as is likely here. Given that the value of P depends on the value of the threshold parameter, ζ, which controls the pooling and smoothness of the Generalized Skyline Plot, then the optimal value of ζ is chosen to be the value which minimises AIC_c.

Figure 3.2 shows some examples of the Generalised Skyline plot model for different values of ζ. Figure 3.2 (a) shows the optimal value of ζ, as chosen using AIC_c while (b) and (c) show extremes of ζ, (b) with a very small value and (c) with a larger value. Comparing the three plots, we see that in the extreme case of $\zeta = 0.2$, almost all of the temporal structure is removed: the

(a)



(b)



(c)

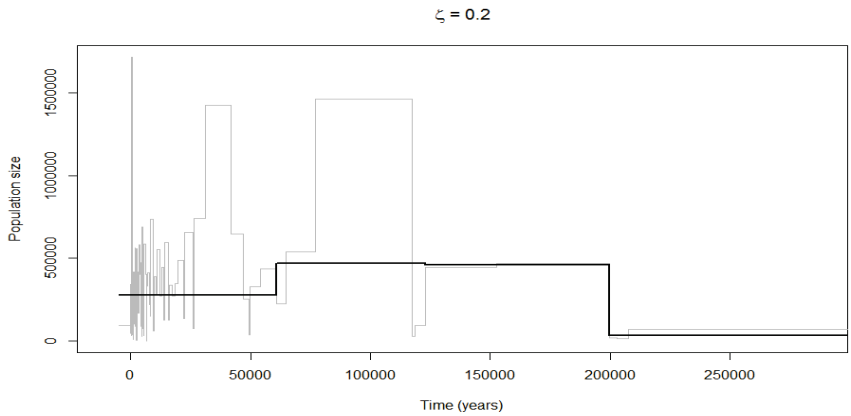


Figure 3.2: Examples of Generalised Skyline Plots. Each plot shows an estimate of the population size $N(t)$ and the corresponding ζ, so that in each case there is a different number of pooled intervals: in (a) there are 7, in (b) there are 28 and in (c) there are 4. The Classic Skyline Plot ($\zeta = 0$) is shown in each panel (grey line). These population sizes were estimated from a genealogy with $n = 100$.

population size function $N(t)$ hardly varies through time. In contrast, when $\zeta = 0.005$, there is a lot of (presumed) noise, particularly close to the present day. Lastly, the optimal value of $\zeta = 0.075$ shows a population size estimate that captures most of the temporal structure but removes the noisy population size estimates, between the present and 50,000 years ago.

Again, as with the Classic Skyline Plot model, this method depends on $\hat{G}$, the point estimate of the genealogy of the DNA sequence samples. Therefore, this model also ignores any error associated with the phylogenetic reconstruction. This leads us to the Bayesian Skyline plot model.

3.3 The Bayesian Skyline Plot Model

Until the Bayesian Skyline plot model was developed (Drummond et al., 2005), the Skyline Plot model depended on $\hat{G}$, the point estimate of the genealogy reconstructed from a set of observed DNA sequences, ignoring any uncertainty in $\hat{G}$. The inference of population size could have been repeated with alternative $\hat{G}$s to explore robustness, but this was rarely, if ever, done in practice.

The Bayesian Skyline Plot model takes into consideration the error associated with reconstructing the genealogy by sampling over the space of possible genealogies through MCMC methods. Like the two previous Skyline Plot models, the Bayesian Skyline Plot model uses a flexible demographic model. It combines inference of demography and genealogy into one larger scheme, allowing for the uncertainty in both, as well as uncertainty in the parameters of the mutation model.

3.3.1 Prior specification

The Bayesian Skyline Plot model models $N(t)$ as a piecewise constant function as before. The model parameters are listed below.

Define the following terms:

- G represents the genealogy of n contemporaneous DNA sequences, i.e. DNA sequences all sampled at the same point in time.

- P represents the number of grouped and ordered intervals in the genealogy, $(1 \leq P \leq n-1)$, as in Section 3.2. In the Bayesian Skyline Plot model, the hyperparameter, P, is the number of pooled coalescent intervals or groups and can be chosen a priori or be sampled in a hierarchical model via reversible-jump MCMC. Drummond et al. (2005) choose P a priori, and it is usually selected to be 10. In this thesis, $P = 10$. Note that in the Generalised Skyline Plot model, the value of P is determined by the choice of ζ.

- $\Theta = \{\theta_1, \theta_2, \ldots, \theta_P\}$ is a vector and represents the population sizes within each grouped interval.
- $A = \{a_1, a_2, \ldots, a_P\}$ is a vector and represents the number of coalescent events in each interval, $(a_i > 0, \sum_{i=1}^{P} a_i = n-1)$.
- Ω is a vector containing all relevant parameters from the assumed mutation model of the data, such as α_G, κ, etc.

Next, let $f_G(G|\Theta,A)$ be the probability density of genealogy G given the demographic parameters, where the vectors Θ and A define a piecewise constant demographic history with $2P-1$ demographic parameters. Then, the log likelihood of the piecewise demographic model is given by Drummond to be

$$\log\big(f_G(G|\Theta,A)\big) = \sum_{i=1}^{n-1} \Big(\log\frac{k_i(k_i-1)}{2\theta_{h(i)}}\Big) - \frac{k_i(k_i-1)\Delta u_i}{2\theta_{h(i)}}, \tag{3.8}$$

where $\Delta u_i = u_i - _{i-1}$ is the waiting time between each coalescent event, and $k_1, k_2, \ldots, k_{n-1}$ denotes the number of lineages present during each Δu_i. A mapping function, $h()$, is introduced. This provides a mapping from the indices of $\boldsymbol{u}$ (the vector of times at which coalescent events occur from u_1 to u_{n-1} where at the leaves of the genealogy $u_0 = 0$) to the indices of $\boldsymbol{w}$ (the vector of times at which each grouped interval ends from w_1 to w_P and is a subset of $\boldsymbol{u}$). This mapping function is defined by

$$h(i) = \begin{cases} 1, & \text{if } i \le a_1, \\ j, & \text{if } \sum_{k=1}^{j-1} a_k < i \le \sum_{k=1}^{j} a_k. \end{cases} \tag{3.9}$$

A simple smoothing is also introduced on the components of θ to represent the belief that population size is autocorrelated through time, via the prior distribution

$$\theta_j|\theta_{j-1} \sim \text{Exp}(\theta_{j-1}), \quad j = 2, \ldots, P. \tag{3.10}$$

A scale-invariant Jeffrey's prior on the first element is also added, $f_{\theta_1} \propto \frac{1}{\theta_1}$, to indicate that the prior belief is invariant to timescale. Then, the multivariate prior distribution on θ is

$$f_\Theta(\Theta) \propto \frac{1}{\theta_1} \prod_{j=2}^{P} \theta_{j-1} e^{-\theta_j/\theta_{j-1}}. \tag{3.11}$$

The authors do not specify the prior distribution on the vector A (Drummond et al., 2005), but since A contains group sizes that define the number of coalescent events a_i in each interval $(i = 1, \ldots, P)$ then it could be that each vector with $a_i > 0$ $(i = 1, \ldots, P)$ and $\sum_{i=1}^{P} a_i = n-1$ has equal probability.

Lastly, each of the independent prior distributions on these parameters within Ω are as follows. Sensitivity analysis was carried out on all of these parameters, and it was found that initial values and the prior hyperparameters were not sensitive. Two common parameters to estimate under all mutation models are the shape parameter of the Gamma distributed mutation rate across sites, α_G, and the proportion of invariant sites, η (Section 2.4). An Exponential prior was placed on α_G due to the fact that the shape parameter of a Gamma distribution must be positive. In particular,

$$\alpha_G \sim \text{Exp}(2),$$

with mean 0.5. The proportion of invariant sites must lie between 0 and 1 (given that it is a proportion) and so a Uniform distribution is a natural choice of prior for this parameter,

$$\eta \sim \text{Uniform}(0,1).$$

There are no further parameters to estimate under the simple JC69 mutation model, but for both the HKY85 and TN93 mutation models, the nucleotide frequencies are now estimated rather than fixed in equal proportions. Again, given that these parameters $\boldsymbol{\pi} = (\pi_T, \pi_C, \pi_A, \pi_G)$ are also proportions, a Dirichlet prior, uniform on the simplex $\pi_T + \pi_C + \pi_A + \pi_G = 1$ is chosen:

$$\boldsymbol{\pi} = \text{Dirichlet}(1,1,1,1).$$

Next are the parameters that account for transitions and transversions. Under the HKY85 model there is only one new parameter, κ. This parameter represents a ratio of transitions to transversions and so must be greater than zero. Since only a lower bound on this parameter is needed a log Normal prior is chosen.

$$\kappa \sim \text{logNormal}(1, 1.5625),$$

with mean 1 and standard deviation 1.25. In the TN93 model transitions and transversions are represented by two parameters κ_1 and κ_2, respectively. Again, both of these parameters are ratios and so must be greater than zero, and so the log Normal prior is applied here too,

$$\kappa_1 \sim \text{logNormal}(1, 1.5625),$$
$$\kappa_2 \sim \text{logNormal}(1, 1.5625),$$

again, each with mean 1 and standard deviation 1.25.

3.3.2 Sampling from posterior distribution using MCMC

MCMC methods are then used to sample parameters from the posterior distribution of the Bayesian skyline plot model using the Metropolis-Hastings algorithm (Drummond et al., 2002).

The posterior distribution sampled is

$$f(\Theta, A, \Omega, G|D, \mu) = \frac{1}{z} \mathrm{P}(D|\mu, G, \Omega) f_G(G|\Theta, A) f_\Theta(\Theta) f_A(A) f_\Omega(\Omega), \tag{3.12}$$

where D denotes the data of sampled DNA sequences, z is the normalising constant and f_Ω contains the independent prior distributions of the parameters under the relevant substitution model. The term $\mathrm{P}(D|\mu, G, \Omega)$ represents the probability of the observed DNA sequences given the mutation model and sampled genealogy. The mutation rate, μ, is a parameter assumed known which scales the genealogy from units of mutations per site to units of time.

Given the complexity of the parameters to be sampled, e.g., genealogies, this is a challenging process and so many different proposal mechanisms are used. For the Markov Chain X_n, $n = 0, 1, 2 \ldots$ with $x = (\Theta, A, \Omega, G)$, suppose that $X_n = x$. A value for X_{n+1} is found by the following algorithm where a given move may alter any one of Θ, A, Ω or G. Introduce $m = 1, 2, \ldots, M$ which labels the different move types (or proposal mechanisms).

1. For the random operation m acting on state x, propose a move to x' from the probability density $q_m(x'|x)$.

2. Determine the ratio of posterior densities as

$$P(x, x') = f(x'|D, \mu) / f(x|D, \mu),$$

   and the ratio of the densities for proposals $x' \to x$ and $x \to x'$ as

$$Q_m(x, x') = q_m(x|x') / q_m(x'|x).$$

3. The type of move m is chosen according to a fixed probability distribution on all the M move types.

4. A value for the proposed state x' is drawn from the density $q_m(x'|x)$.

5. Calculate the acceptance probability

$$a_m(x, x') = \min\big(1, P(x, x') Q_m(x, x')\big).$$

6. Then, set $X_{n+1} = x'$ with probability $a_m(x, x')$ and set $X_{n+1} = x$ with probability $1 - a_m(x, x')$.

Each of the M proposal mechanisms have different acceptance probabilities that are accounted for in $Q_m(x, x')$. The move labelled $m = 1$ is a *scaling move* which scales the times of ancestral nodes in the tree but does not alter the leaf node times (Green, 1995). For $m = 2$

this is a *Wilson-Balding move* that is based on the branch-swapping move (Wilson and Balding, 1998) and is implemented here by randomly selecting a subtree and moving it to a new branch in the genealogy. For $m = 3$ this is called a *subtree exchange* and involves swapping the subtrees descended from two child nodes of a chosen parent node. When $m = 4$ this is called a *node age move* which chooses an internal node at random and replaces its time with some other new time. Note that in both cases of $m = 1$ and $m = 4$ the root node is included, but a different proposal is placed on drawing the new root node time compared to other internal nodes within the genealogy. Lastly, when $m = 5$ these are *random walks* for all other parameters in the model. The variance of the random walk proposals are tuned to achieve good acceptance probabilities. Drummond et al. (2002) describe these M proposals in the case of a constant population size scenario. In a variable population size case further updates must be done. The updates on the tree parameters ($m = 1,\ldots,4$) are multivariate moves, since they change more than one of the parameters that describe the tree (e.g., the node age move will affect the branch lengths). Another multivariate move changes the number of coalescent intervals in each group. The remaining parameters in the model ($m = 5$), including the population sizes associated with each group, the nucleotide substitution rate parameters and the parameters characterising mutation rate heterogeneity, are updates separately, with univariate random walks.

3.3.3 Convergence

Inference will only be valid if the Markov chain has converged to its target distribution, for which the chain needs to be long enough. The chain should also mix well, i.e., the sampler should move rapidly between regions of high probability. There are a number of ways to check convergence of the chain. One informal and common approach is to examine trace plots for each parameter in the model. These plots should show no trend in the sample after the burn-in period, which is first part of the chain that has not converged and is discarded before any inferences are drawn around parameters since the sampled states do not represent the true posterior distribution.

The Gelman-Rubin statistic is one statistic which identifies non-convergence between multiple MCMC sequences with over-dispersed starting points. If convergence has been reached in chains with over-dispersed starting points then the variance within the chains should be the same as the variance between the chains (Gelman et al., 1992). The within- and between- chain variances for one parameter are defined by

$$W = \frac{1}{K}\sum_{j=1}^{K} s_j^2 \quad \text{and} \quad B = \frac{M}{K-1}\sum_{j=1}^{K}(\bar{\theta}_j - \bar{\theta})^2, \tag{3.13}$$

respectively, where K is the number of chains, M is the number of samples within each chain

and s_j^2 is the variance of the j^{th} chain, given by

$$s_j^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\theta_{ij}-\bar{\theta}_j)^2, \tag{3.14}$$

for parameter θ where $\bar{\theta} = \frac{1}{K}\sum_{j=1}^{K}\bar{\theta}_j$, $\bar{\theta}_j = \frac{1}{M}\sum_{i=1}^{M}\theta_{ij}$ and θ_{ij} is the parameter value in the i^{th} sample of the j^{th} chain. The estimated variance of the stationary distribution is then a weighted average of W and B,

$$\widehat{Var}(\theta) = \left(1-\frac{1}{M}\right)W + \frac{1}{M}B. \tag{3.15}$$

The potential scale reduction factor (PSRF) is defined by

$$\hat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}}. \tag{3.16}$$

The value of $\hat{R}$ should be approximately one if chains have run long enough to reach their stationary distributions.

The effective sample size (ESS) measures the autocorrelation between the draws of the Markov chain and can be thought of as the number of effectively independent draws from the posterior distribution to which the Markov chain is equivalent. This should also be checked to ensure that the chain is mixing well. The formula for finding the ESS is

$$S_{\text{eff}} = \frac{M}{\kappa}, \tag{3.17}$$

where $\kappa = 1 + 2\sum_{k=1}^{\infty}\rho_k$ and ρ_k is the lag k autocorrelation, i.e., the correlation between the draw at iteration i and the draw at iteration $i+k$ for $k = 1, 2, \ldots$.

3.3.4 BEAUti, BEAST and Tracer

The pipeline of a typical Bayesian Skyline Plot analysis is implemented using three programs (Rambaut et al., 2018; Suchard et al., 2018):

1. BEAUti

   This is a graphical user-interface (GUI) application for generating BEAST XML (extended markup language) files (`http://beast.community`).

2. BEAST

   This program uses an XML command file as input and returns 'log files' as output. These log files record a sample of the states that the Markov chain encountered. It is this output that is used to produce estimates of the parameters of interest, in this case population sizes (`http://beast.community`).

3. Tracer

   This is a graphical application that is used to explore the output from BEAST (`http://tree.bio.ed.ac.uk/software/tracer/`).

The process begins by loading a set of n DNA sequences into BEAUti (Bayesian Evolutionary Analysis Utility) in FASTA (Lipman and Pearson, 1985) or Nexus (Maddison et al., 1997) format. From here, the user selects a range of modelling options, including the nucleotide substitution model (JC69, HKY85, TN93, etc.), the site heterogeneity model (none, Gamma, invariant sites or both), the number of Gamma categories, the coalescent prior on the tree (assuming the Bayesian Skyline Plot, a constant population size, exponential growth, etc.) and prior distributions and values to initialise the Markov chain for each parameter in the model. Properties of the MCMC chain (number of steps in the chain T and level of thinning) are specified. BEAUti then generates an XML file containing all this information along with the DNA sequences themselves. Figure 3.3 shows a screenshot of the user interface.

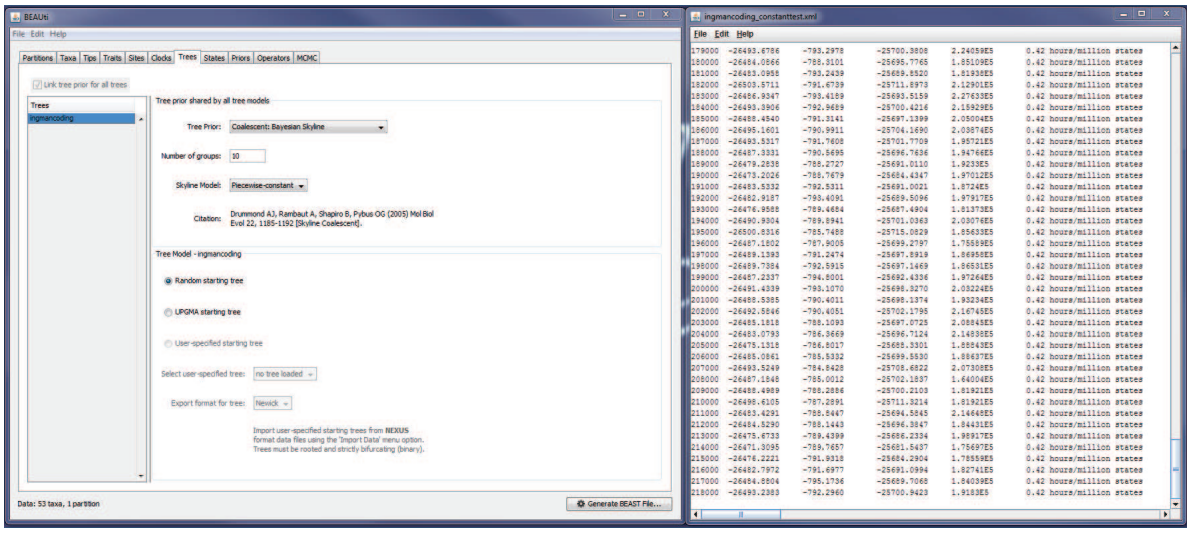


Figure 3.3: Screenshots of the user interfaces for BEAUti specifying the tree prior (left) and a snapshot of BEAST running the MCMC chain (right).

Once this XML file is generated, it is then loaded into BEAST and the MCMC chain is run (also shown in Figure 3.3). BEAST saves a log file containing each parameter value at each step in the chain after thinning (Θ_j, Ω_j and A_j) and a tree file that contains the sampled tree at each step in the chain (G_j) (the latter saved in Newick format including tree topology and branch lengths). These two files are needed to reconstruct the Skyline Plot in Tracer. Figure 3.4 shows a screenshot of the post-analysis that can be carried out using this program.

The demographic history of the sample, $N(t)$, is a piecewise constant function of time at each of the T steps of the chain. A list of T states is obtained, each state with an associated genealogy, a set of mutation model parameters and demographic parameters, Θ_j, Ω_j, A_j and G_j. The marginal posterior distribution of $N(t)$ is calculated at a series of times on a grid. At

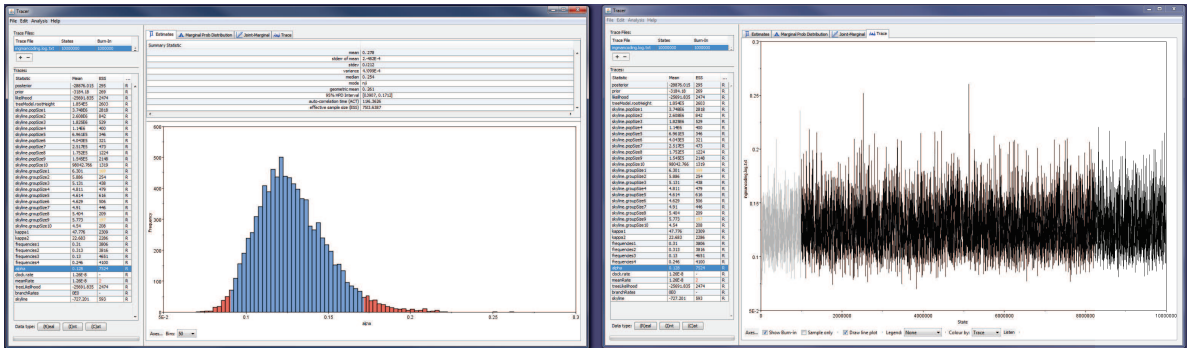


Figure 3.4: Screenshots of post-analysis that can be carried out in Tracer. The panel on the left shows a histogram approximating the posterior distribution of the α_G parameter with summary statistics above and the panel on the right shows the trace plot (with burn-in shown dimmed). The list on the left hand side shows all possible parameters that can be explored.

each time the values of $N(t)$ taken through the chain form a sample from the marginal posterior distribution of the population size at that time. From these the mean, median and 95% Highest Posterior Density intervals can be calculated at each t and this results in the skyline plot of the estimated population size through time with an associated measure of uncertainty (3.5).

This pipeline is relatively straightforward. However it is not designed for large simulation studies where we are looking to analyse hundreds of sets of DNA sequences simultaneously. It is completely impractical to do so using the interfaces and so this had to be done from the command line. This is possible for both BEAUti and BEAST where a template XML file was run through `R` and a model XML created. This model XML file contained all the information about the mutation model, site heterogeneity, prior distribution, etc., as before but allowed for the sets of DNA sequences to be loaded in separately. Furthermore, it allowed more flexibility, e.g., in the number of Gamma categories for the site heterogeneity model since there is not the limitation of the drop-down menu.

Then, as before, the log and tree files are saved and this provides all the required information for reconstructing the population sizes. Another limitation with the pipeline is that Tracer can only display one Bayesian Skyline Plot at a time but for this work multiple curves were required on one plot. So this was implemented in `R`. That is, for every MCMC run, the information of the population size and end points of each group were knitted together from the log and tree files across the steps of the chain. This allowed me to calculate posterior summaries easily for each simulation. Exact details of this process will be presented in Section 4.2.2.

As Figure 3.5 illustrates, the Bayesian Skyline Plot provides a smooth estimate of the population size going back in time. Note however that this is an artefact of the choice of summary. The model itself is piecewise constant. Now, with the DNA sequence data generated and the population size inferred from the Bayesian Skyline Plot model, Chapter 4 will turn to the main issue, the effect of the sampling scheme.

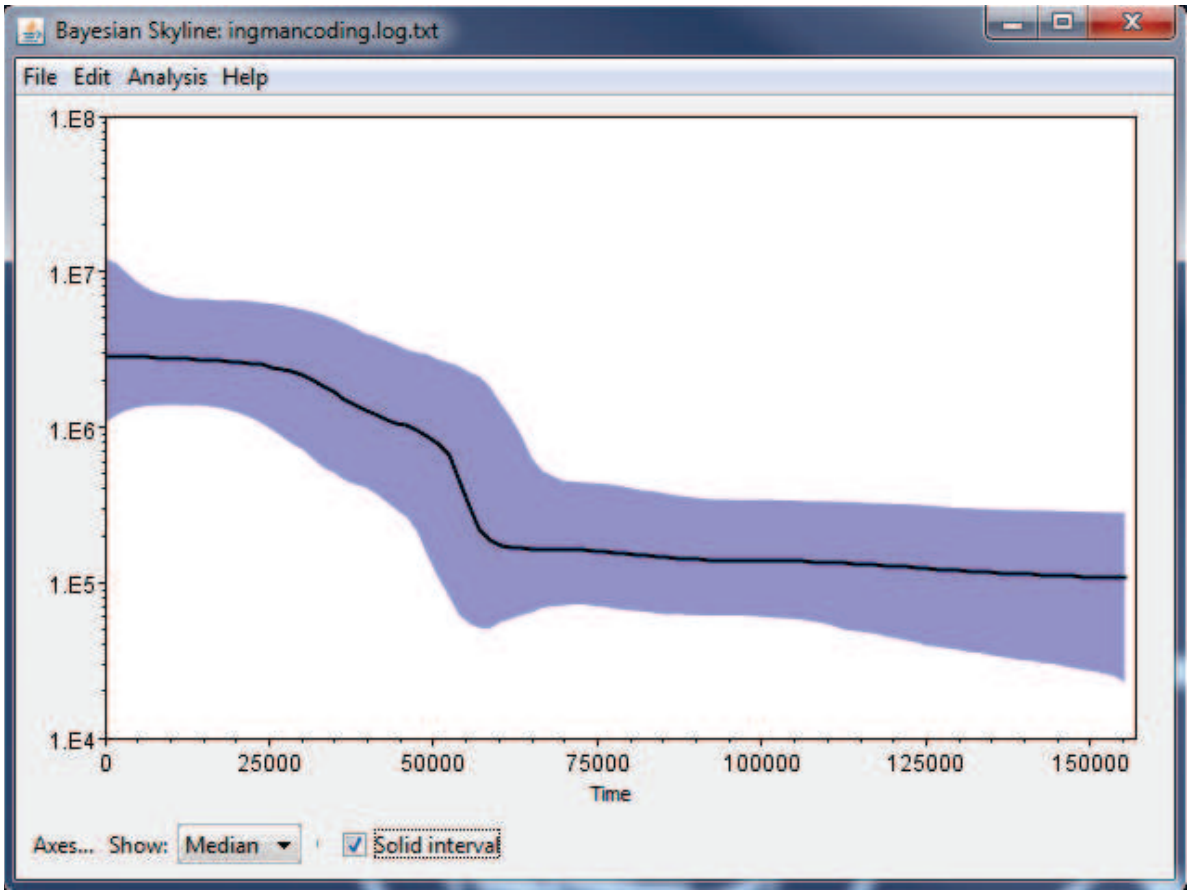


Figure 3.5: Screenshot of the Tracer output of the Bayesian Skyline Plot reconstruction of posterior median population size. Population size is displayed on the y-axis and time in years on the x-axis. The shaded area is the pointwise 95% highest posterior density band.

Chapter 4

Performance of the Bayesian Skyline Plot Model

This chapter presents properties of the genealogies and DNA sequences simulated from the Generative Model in Chapter 2 and the inferred population sizes of these sequences from the Bayesian Skyline Plot model, as discussed in Chapter 3. The main aims of this chapter are:

1. To test whether properties of the simulated DNA sequences correspond to those we expect in theory, and are close to actual sequences, as represented by the Ingman data set of 53 diverse human mitochondrial DNA sequences (Ingman et al., 2000). This allows us to confirm that the simulated sequences are as realistic as possible.

2. To test the reliability of the Bayesian Skyline Plot model using simulated randomly sampled sets of DNA sequences and their true demographic models.

3. To present and compare the results of the Bayesian Skyline Plot model for simulated randomly and non-randomly sampled sets of DNA sequences.

Using the generative model described in Chapter 2, DNA sequences are simulated under the four demographic models: a constant population size, exponential growth, a step in population size and a population bottleneck model. Before comparing the inferences of population size for a non-randomly sampled set of DNA sequences with a randomly sampled set, this chapter will investigate the inferred population size from a random sample of DNA sequences is unbiased. Comparing the inferred population size from the Bayesian Skyline Plot (BSP) model with the true underlying demographic model gives us a sense of how well the model can do at reconstructing historic population sizes from the set of DNA sequences.

The BSP model has been widely applied, likely because of a few factors. It uses a flexible demographic model, properly accounts for uncertainty, and is implemented in a well designed suite of software (discussed in Section 3.3). However, there have been few reports on the model's ability to infer accurate population sizes. This is surprising since a Google Scholar search on the

term 'Bayesian Skyline Plot' (on November 21st 2019) returned 6,910 results and on the same day, the paper introducing the model (Drummond et al., 2005) had been cited 2,370 times.

The little investigation that has been conducted into the limitations of the model treats a related but different issue to the one that will be presented in this chapter (Heller et al., 2013). Heller et al. acknowledge that the Bayesian Skyline Plot model does not account for violated assumptions of the Coalescent Process, particularly the assumption of random mating (panmixia). In an attempt to show that this could lead to inaccurate inferences of population size they conduct a simulation study in a similar manner to this work. However, they focus on the effect of population structure and not with the effect of sampling. They conclude (by comparing population curves) that population estimates from the Bayesian Skyline Plot model are inaccurate when compared to the true underlying population model for a constant population size and a changing population size island model.

Before any conclusions can be drawn on the inferred population size of a *non-random* sample of DNA sequences, the inferred population size of a *randomly* sampled set of DNA sequences that adhere to all assumptions of the model should be compared with the true population model. Once the ability of the model has been checked, this chapter then goes on to infer the population size of a corresponding set of the non-randomly sampled DNA sequences. The chapter begins with the analyses of the Ingman data set (Ingman et al., 2000), to provide parameter estimates for the generative model. Then, two sets of population estimates will be presented and compared with the true underlying demographic model.

4.1 Ingman data analysis

The Ingman data set of complete mitochondrial DNA sequences (mtDNA) of 53 humans of diverse origins provides a small sample of the worldwide level of mtDNA variation (Ingman et al., 2000). This data set has been widely analysed and is highly cited in the literature (Bandelt et al., 2006; Gabriel et al., 2002; Jobling et al., 2014; Kong et al., 2003; Maca-Meyer et al., 2001), possibly due to the diversity of the individuals in the sample and because it was the first large, reliable set of complete mtDNA sequences to be published (Bandelt et al., 2006), i.e., the sequences did not contain any clear signs of missequencing. This makes it an attractive starting point for this analysis, to provide parameter estimates to tune the simulation so that is mimics reality. Much larger data sets are available now, but this one is large enough for our purpose.

The sequences were aligned by eye (since they are not very divergent) in Bioedit (Hall, 2011). The coding region was isolated (the region between nucleotides 577 and 16023 according to the revised Cambridge Reference Sequence (Andrews et al., 1999)), producing a 53×15446 matrix of nucleotides and alignment gaps.

The data were analysed under three mutation models: JC69 $\Gamma+$ I, HKY85 $\Gamma+$ I and TN93 $\Gamma+$ I of increasing complexity from unrealistically simple (JC69) to adequately more parameter

rich (TN93), without being too complex, so that computation of the BSP was not too expensive. Going forward, the results of the TN93 model will be presented to avoid repetitiveness, since the results from the other two mutation models were not very different.

During the early stages of analysis of this data set, an identifiability issue in the mutation model was discovered, which will be discussed now.

4.1.1 Identifiability of α_G and η

While sampling from the posterior distribution for the Ingman data set, it became clear that there was an issue with convergence particularly of the shape parameter of the Gamma distribution of mutation rates across sites, α_G, and the proportion of invariant sites, η. As discussed in Section 2.4, α_G and η both describe a biological property of sites on the DNA sequence and the two parameters are somewhat related. The shape parameter of the Gamma distribution, α_G, accounts for the heterogeneity of mutation rate across sites of the sequence that can mutate, while η is the proportion of invariant sites. For example, if $\eta = 0.4$ this indicates that 40% of the sites on the sequence are invariant, while the other 60% of sites do mutate, and do so with a mutation rate drawn from the Gamma distribution.

Figure 4.1 shows draws from the joint marginal posterior distribution of both parameters over the chain with 10 million iterations and a burn-in of 1 million.

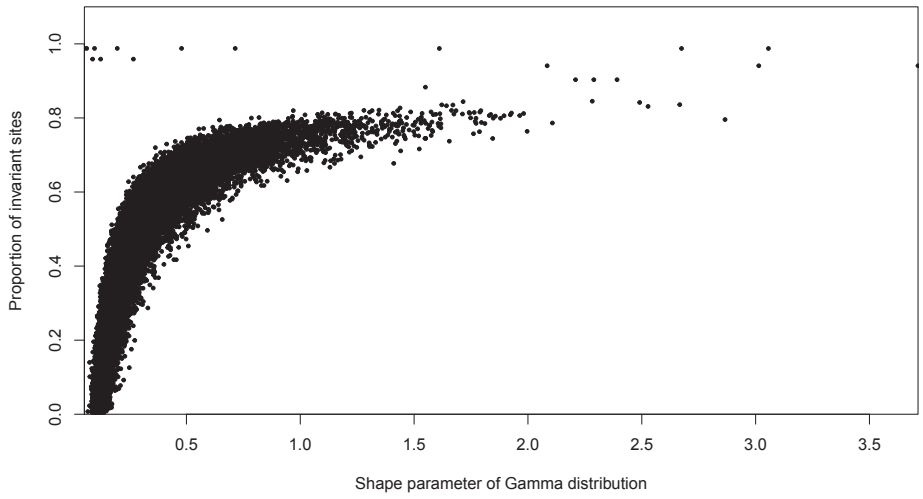


Figure 4.1: Samples from the joint marginal posterior distribution of α_G and η.

It was difficult to obtain convergence for these parameters and as Figure 4.1 shows, it is clear why. When α_G is small/large, the proportion of invariant sites η tends to be small/large also. For this reason, it was decided that the proportion of invariant sites should be removed from the model. Allowing for a small α_G (less than one) gives a distribution peaked at zero. Thus, the

effect of invariant sites is essentially present as a large majority of sites will have a mutation rate close to zero (whilst some sites will be randomly allocated a high mutation rate and become very variable). Figure 4.2 illustrates the Gamma distribution with a range of values of α_G. When $\alpha_G < 1$ there is a divergence at zero.

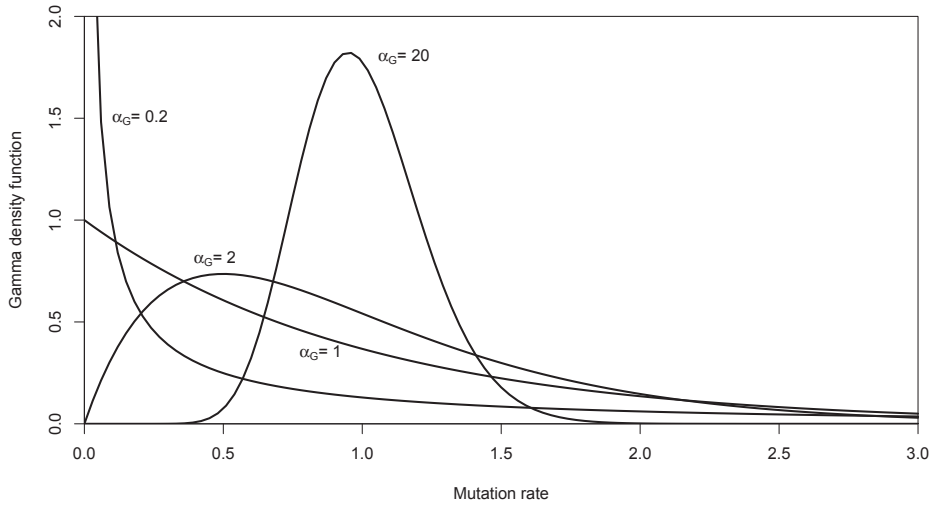


Figure 4.2: Probability density function of the Gamma distribution $\text{Ga}(\alpha_G, \alpha_G)$ for $\alpha_G = 0.2, 1, 2, 20$ representing variable mutation rates across sites.

Therefore, given that having both α_G and η in the model results in neither of them being reliably estimated and that having an α_G of less than one will approximately account for invariant sites, η was put equal to zero. Yang discusses the detrimental effect of having both parameters in the model and concludes that an α_G leading to a mutation rate distribution with the typical 'L' shape is desirable (Yang, 2006).

In addition, it has been shown that the estimate of η from real data is very sensitive to the number of DNA sequences in the sample and to the length of each of these sequences. The value of η will never be more than the observed proportion of constant (unchanging) sites in the sample of DNA sequences, so when more sequences, especially sequences with more genetic variation, are included in the sample, the estimate of η tends to drop along with the number of constant sites. It has been suggested that, to overcome this, one could use a Gamma mixture model (Mayrose et al., 2005), where the mutation rate for any site in the sequence is drawn from a mixture of two Gamma distributions with two sets of different parameters (one of the distributions peaking near zero). This has been found to be more stable than the $\text{I}+\Gamma$ model, but when estimating genealogies and branch lengths both the mixture model and the single Gamma distribution model produce similar results (Yang, 2006). Therefore, in the remainder of this work only a simple Gamma distribution will account for the variable mutation rates across sites. In fact, when analysing the 53 DNA sequences from the Ingman data set, the posterior mean of

α_G was estimated to be 0.1279 with a Highest Posterior Density interval of (0.0903, 0.1715), under the TN93 mutation model. A plot of this distribution (with mean set to 1) is shown in Figure 4.3. This estimated value of α_G is indeed less than one and has a large peak at zero, satisfying the condition that it needs to account for 'invariant' sites.

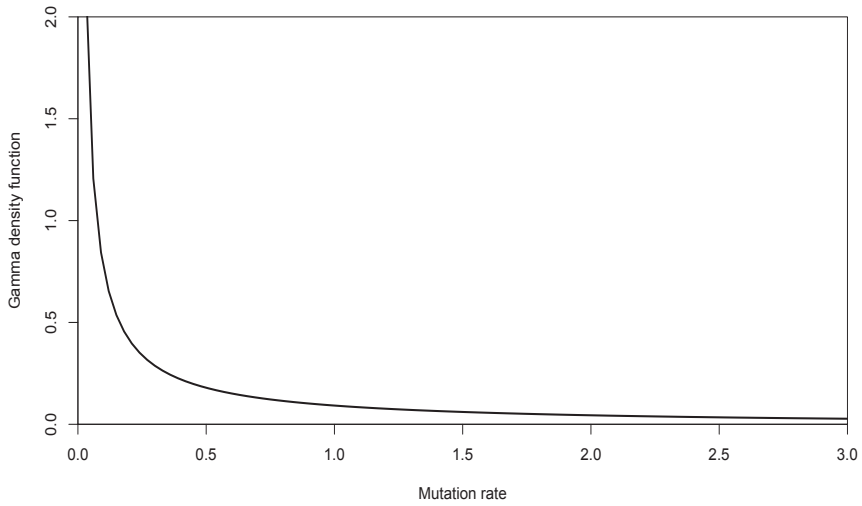


Figure 4.3: Probability density function of the Gamma distribution $\mathrm{Ga}(0.1279, 0.1279)$ representing variable mutation rates across sites.

4.1.2 Discrete gamma distribution of mutation rate and categories

One known problem with the simulation of DNA sequences is again with the shape parameter of the Gamma distributed mutation rate across sites within the sequence. The issue arises from simulating DNA sequences with mutation rates from a continuous Gamma distribution but then estimating the shape parameter of a discrete Gamma distribution.

The discrete Gamma model of mutation rates across sites uses C equal probability categories to approximate a continuous gamma distribution. The mean rate in each category represents all the rates in that category. This makes estimating α_G computationally easier and it has been shown that approximating the Gamma distribution with only four categories gives a good approximation to the continuous distribution (Yang, 2006). Figure 4.4 shows an example of the continuous Gamma distribution estimated by a discrete Gamma distribution with four categories.

The boundaries of the categories are at the 25th, 50th, and 75th percentiles, cutting the density into four categories each of probability $\frac{1}{4}$. The four mutation rates selected to represent the categories are the mean of each of these categories. In this example, the mean rates in the four categories are 0.0325, 0.0484, 0.0852, and 0.4841. As much as this number of categories is a common choice, it seems a little small to capture the very skewed distribution of rates ($\alpha_G \ll 1$) that occurs often in practice. So the number of categories was selected to be $C = 25$.

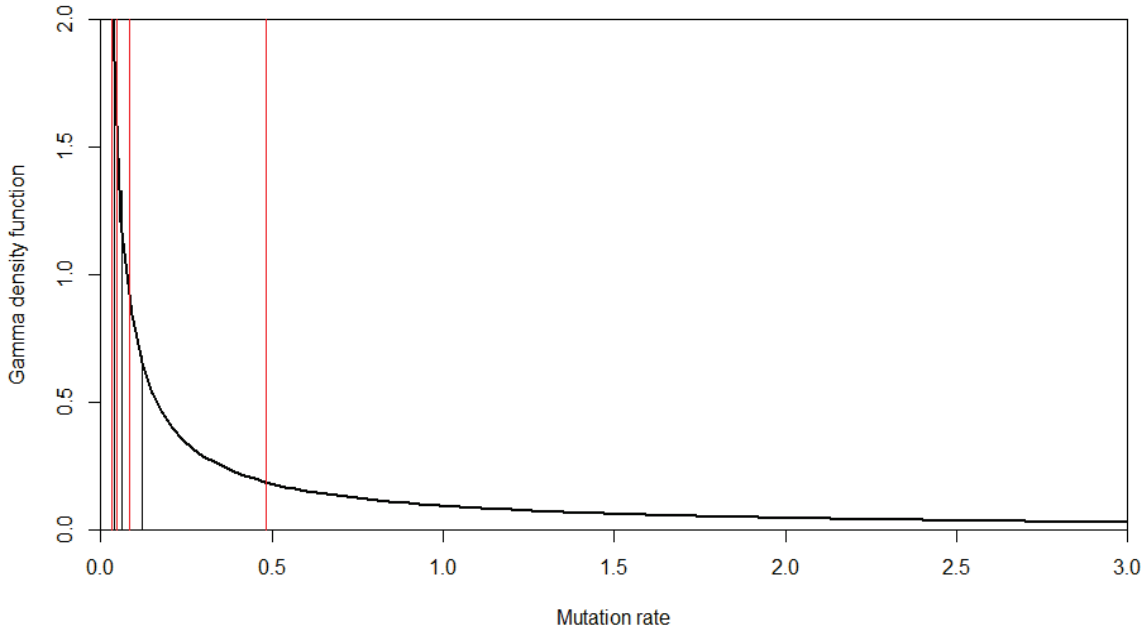


Figure 4.4: The discrete Gamma model of mutation rates across sites with four equal probability categories to approximate the continuous Gamma distribution. This plot shows a Gamma distribution (thick solid line) with shape parameter 0.1279 as estimated from real data and mean 1. The four vertical black lines are at the 25th, 50th, and 75th percentiles of the distribution. The four red lines have abscissae equal to the four possible mutation rates (located at the mean rate within each category.)

In principle, one would want to use as many categories as possible so that the discrete distribution approximates the continuous distribution as accurately as possible. However, there is computational cost to this. The larger the value of C, the longer the chain takes to run. Another reason for choosing a larger C is because the discrete case leaves less variation in rate among sites than the continuous case. As a result, estimates of α_G using the discrete method are often smaller than those estimated from the continuous distribution. Hence, to err on the side of caution a larger C than is recommended by Yang was chosen for this work. It has been shown however, that for data sets with hundreds of sequences (like our simulated data), many more categories are desirable (Mayrose et al., 2005).

4.1.3 Model convergence

Convergence was checked for every parameter in all three mutation models, the genealogy and all population size and group size parameters, but results only for the TN93 model are presented here to avoid repetition. The other cases behaved similarly (data not shown). Over-dispersed starting values for repeated chains (five chains for each combination of demographic and muta-

tion model) were used for each parameter and the chain length was set to 10^7 with a burn in of 10^6 iterations. The length of the chains was decided through trial and error, making sure chains appeared to have converged and that effective sample sizes were large enough for a representative independent sample. The results were thinned by recording only every 1,000$^{\text{th}}$ iteration of the chain. In all cases, convergence was quickly reached and the trace plots showed the desired rapid up-and-down variation with no patterns.

The Gelman-Rubin statistic for the α_G parameter was close to 1 revealing no convergence issues. For parameters κ_1 and κ_2, the value of $\hat{R}$ was also close to 1 for both of them. The Gelman-Rubin statistic was calculated for all population size and group size parameters in each model, as was the tree length parameter. All of these parameters were very close to 1 too.

The effective sample sizes for each parameter in the model (including parameters from the mutation model as well as parameters such as group size and population sizes from the demographic model) were checked for all chains. The BEAST documentation suggests that an ESS of less than 100 is too low and prefers a value greater than 200. This is a fine balance, because having a very high ESS (say greater than 10,000) could be a waste of computational resources. The ESS of the tree length always had large ESS values (between 3,000 and 5,000). The frequencies of nucleotides, the site heterogeneity and the transition/transversion parameters were always comfortably between 1,000 and 10,000 indicating good mixing of the MCMC. Population sizes for each group were always in the hundreds, sometimes thousands and the group sizes were a little lower, mostly just over 100. These were all sufficiently large enough indicating that the chains all mixed well.

4.1.4 Bayesian Skyline Plot model output for Ingman data

This section will present the results from the Bayesian Skyline Plot analysis of the Ingman data set using the TN93+Γ mutation model. Analysis was carried out using JC69 and HKY85 too and the results were very similar (results not shown). Figure 4.5 shows the inferred population size as a function of time. The plot shows a rapid increase in population size from around 60,000 years ago consistent with some hypotheses about human evolution. It has been argued that population size increased around this time due to advances in human dispersal, due to climate change and due to the widely accepted hypothesis of human migration out of Africa (Klein, 1992; Macaulay et al., 2005; Mellars, 2006; Soares et al., 2011). The error bands around this estimate of population size in the past 60,000 years are very large. This is a common feature of the Bayesian Skyline Plot model and is a reflection of the limited amount of information in contemporary samples of DNA sequences about past population size changes. Nevertheless, even accounting for the error bands, a strong signal of expansion remains. The timing of this population increase was also motivation for choosing the demographic parameters for the step model back in Section 2.5.1. Another feature of this model output is that the population size stays roughly constant at around 10,000 prior to this dramatic expansion. This population size

	Mutation Model					
Parameter	**JC69**		**HKY85**		**TN93**	
α_G	0.113	(0.081, 0.153)	0.117	(0.081, 0.155)	0.128	(0.090, 0.172)
π_T	0.25	—	0.244	(0.237, 0.250)	0.246	(0.240, 0.253)
π_C	0.25	—	0.311	(0.303, 0.318)	0.313	(0.306, 0.321)
π_A	0.25	—	0.313	(0.305, 0.320)	0.311	(0.303, 0.318)
π_G	0.25	—	0.132	(0.128, 0.138)	0.130	(0.125, 0.135)
κ	—	—	33.297	(23.052, 44.826)	—	—
κ_1	—	—	—	—	47.416	(32.303, 64.178)
κ_2	—	—	—	—	22.603	(15.500, 30.558)

Table 4.1: Posterior mean parameter values and corresponding 95% HPD intervals using the Ingman data.

was motivation for the value of $N(t)$ in the constant demographic model (Section 2.5.1).

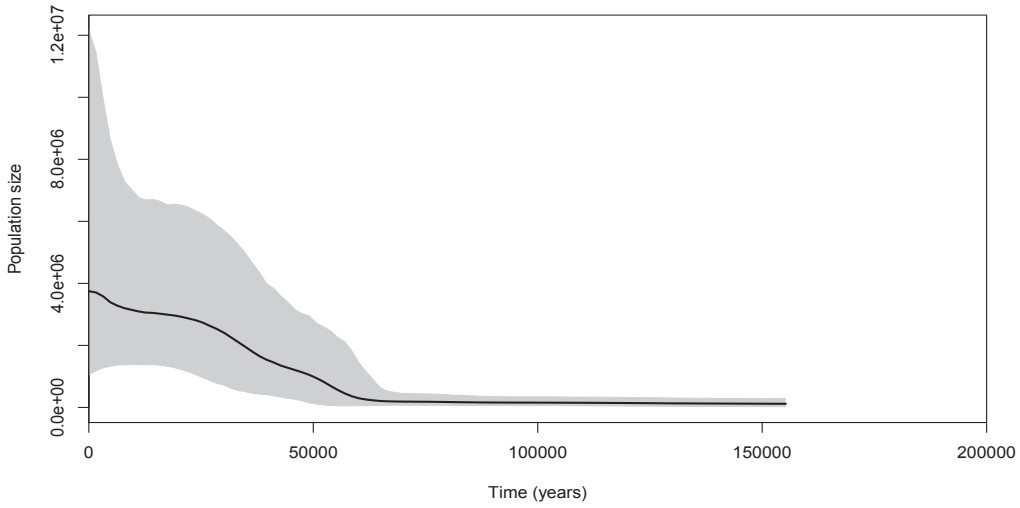


Figure 4.5: Posterior mean population size from Bayesian Skyline Plot model for the Ingman data, with pointwise 95% HPD intervals.

Next, the posterior distributions of each mutation model parameter were used to provide values of the parameters for the mutation model in the simulation (Sections 2.3 and 2.4). These are presented in Table 4.1. Note that in all three models, these estimates were consistent across multiple MCMC chains to three decimal places.

The parameter estimates in Table 4.1 were then taken to be the parameter values for the simulated DNA sequences. The last parameter to estimate was the rate parameter for the exponential growth demographic model (Equation 2.22). An exponential growth demographic model was fitted to the Ingman data set using BEAST. This gave a growth rate of $r = 3.29 \times 10^{-5}$ per year with a 95% HPD interval of $(2.336 \times 10^{-5}, 4.29 \times 10^{-5})$. The initial population size N_0 was estimated to be 3.5×10^6 with a 95% HPD interval of $(1.7 \times 10^6, 5.5 \times 10^6)$. Lastly, to reiterate,

the overall mutation parameter μ was determined, as in Section 2.5.3, to be 1.26×10^8.

4.2 Simulated DNA Sequences

This section will discuss the simulated DNA sequences and their properties. Firstly, comparing properties of the simulated genealogies with theoretical results provided a useful means of detecting errors in the `R` code that performs the simulation. Secondly, comparing a measure of genetic variation within the Ingman DNA sequences to the genetic variation in the simulated sequences was another way of checking that the simulation settings were as realistic as possible. This section will also discuss some problems that had to be addressed when using the Bayesian Skyline Plot model. I will then compare the inferred population size to the true underlying demographic model to assess any bias in the BSP.

For each of the four demographic models (constant population size, step in population size, bottleneck population and an exponential growth in population size), one hundred trees with $n = 100$ leaves were simulated from the Coalescent Process. DNA sequences were evolved along each of these trees under each of the three mutation models. In total there were 12 sets of simulated DNA sequences, each set representing 100 *randomly* sampled DNA sequences. (To be clear, only four sets of 100 trees were simulated. The trees stayed constant across mutation models.) Then, according to the scheme discussed in Section 2.5.2 where each tree was cut on a branch long enough to contain at least one mutation in expectation, a corresponding set of *non-randomly* sampled DNA sequences was obtained.

4.2.1 Properties of the simulated sequences

A convenient check that the simulations are accurate is to compare them to exact summaries of the tree. As explained in Section 2.5.4, the first step of the simulation process is to simulate a tree from the Coalescent Process. The two properties of the tree that are used here are the length of the tree and the depth of the tree. Figure 4.6 shows a tree simulated under the constant coalescent model and will be used to illustrate the properties introduced in this section.

The depth of the tree (T_{MRCA}) is the time from tips to the root along the tree. The length of the tree (L), on the other hand, is the sum of the lengths of all branches in the tree. By the properties of the Coalescent Process in Chapter 2, since the waiting time T_k until k lineages coalesce to $k-1$ is exponentially distributed (Equation 2.2) with rate $\binom{k}{2}$, the expectation and variance of that time are

$$\mathrm{E}(T_k) = \binom{k}{2}^{-1} \quad \text{and} \quad \mathrm{Var}(T_k) = \binom{k}{2}^{-2}.$$

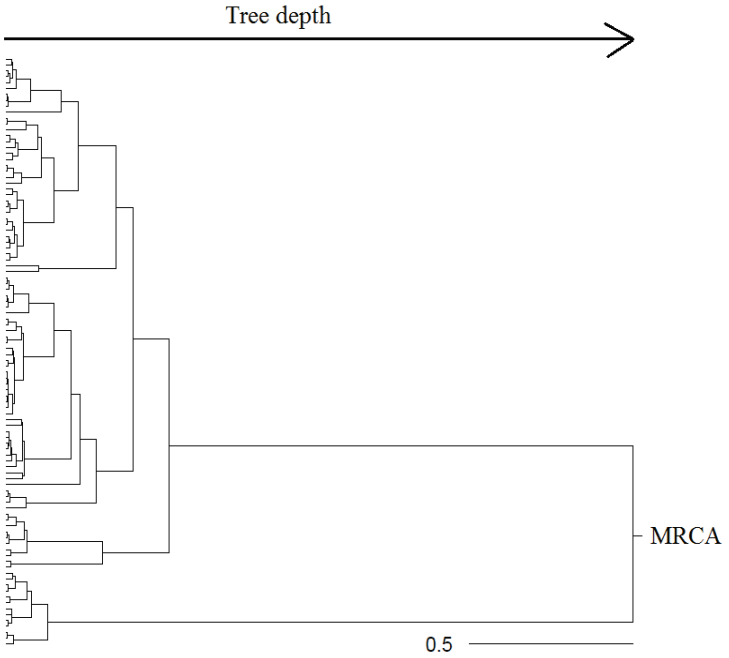


Figure 4.6: A simulated tree under the constant population size demographic model with a sample size of 100. A scale bar has been added and shows time in coalescent units.

Thus, the expected time to the most recent common ancestor, T_{MRCA}, is

$$\mathrm{E}(T_{MRCA}) = \sum_{k=2}^{n} \binom{k}{2}^{-1} = 2\sum_{k=2}^{n} \frac{1}{k(k-1)} = 2\sum_{k=2}^{n} \left[\frac{1}{k-1} - \frac{1}{k}\right] = 2\left(1 - \frac{1}{n}\right), \tag{4.1}$$

and as the sample size n increases, this tends to 2 in units of N_0 generations. The length of the tree can be written as $L = \sum_{k=2}^{n} kT_k$ and so the expected length of the tree under the Coalescent Process is

$$\mathrm{E}(L) = \sum_{k=2}^{n} k\mathrm{E}(T_k) = \sum_{k=2}^{n} k\frac{2}{k(k-1)} = 2\sum_{k=2}^{n} \frac{1}{k-1} = 2\sum_{k=1}^{n-1} \frac{1}{k}. \tag{4.2}$$

Given that in this simulation study, the sample size is $n = 100$, the expected depth of the tree should be about $2N_0$ generations, and the expected total length of the tree should be about $10N_0$ generations. It should be noted that these theoretical expectations only hold for the constant demographic model. Figures 4.7 and 4.8 show boxplots of the tree depths and lengths, respectively, over demographic models.

In both Figures 4.7 and 4.8, on average over the 100 simulated trees assuming a constant population size demographic model, the tree depths and lengths correspond to what would be expected from the Coalescent Process. The mean tree depth over 100 trees and under the constant model was 1.824 and the tree length 10.101 (in coalescent units). Comparing these to the expected tree depth of 1.98 and expected tree length of 10.355 shows that the properties of the simulated trees are very similar to those that would be expected. A one-sample t-test compared these values and both the tree length and the tree depth were not significantly different from

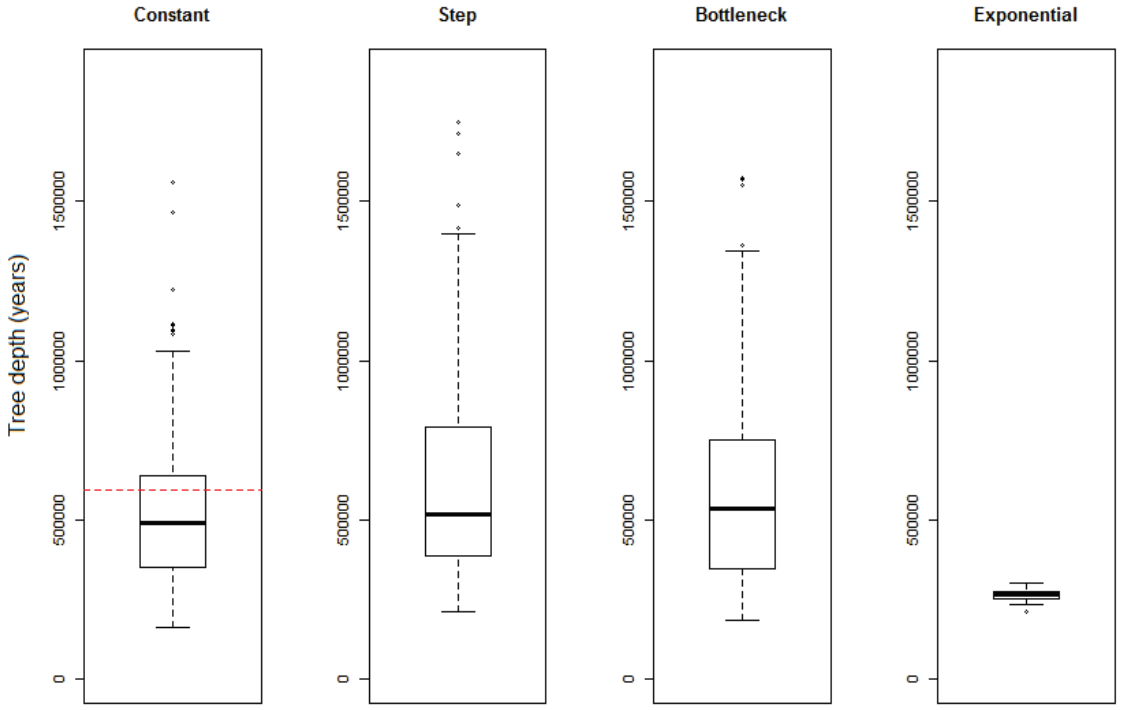


Figure 4.7: Boxplots of the tree depths of 100 trees for each demographic model. The red dotted line shows the expected tree depth under the Coalescent Process.

those expected at 5% significance level.

Under the Coalescent Process time is scaled by N_0 generations and under each demographic model the value of N_0 changes (though the number of years per generations stays constant). Since our unit of time then depends on N_0, our scale of time changes depending on this value: constant $N_0 = 10,000$, step $N_0 = 100,000$, bottleneck $N_0 = 100,000$ and exponential $N_0 = 3,500,000$. As such, the boxplots show the results converted back to units of years so that they are comparable across demographic models. We have no theoretical result from the Coalescent Process to compare the simulated tree lengths and depths to when the population size is not assumed to be constant. The tree depths of the exponential model are somewhat smaller on average than the other models (which are comparable) and less variable. There is much more variation in tree lengths between models, with the exponential model having a much higher tree length corresponding to the more comb-like genealogy in this case.

Next, the genetic variation of the 100 simulations in each demographic model was calculated and compared to the genetic variation in the Ingman data set using the mean pairwise difference. This statistic is the mean number of differences between pairs of sequences in the sample (Tajima, 1983), defined by

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}, \tag{4.3}$$

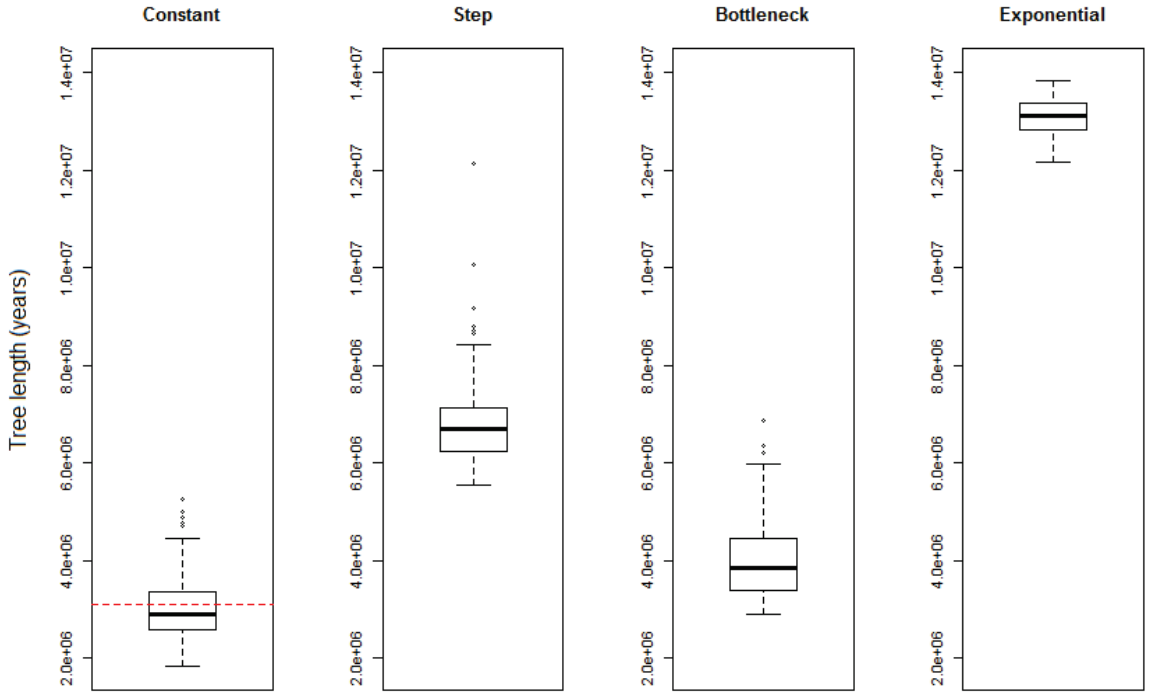


Figure 4.8: Boxplots of the tree lengths of 100 trees for each demographic model. The red dotted line shows the expected tree length under the Coalescent Process.

where n is the sample size and d_{ij} is the number of nucleotide differences between sequence i and sequence j.

The mean pairwise difference of the coding region of Ingman sequences is 46.14. Under the constant population size coalescent the expected mean pairwise difference of a sample is

$$\mathrm{E}(\pi) = 2\mu, \tag{4.4}$$

where μ is the mutation rate per sequence per N_0 generations (Wakeley, 2009). This gives an expected mean pairwise difference of 116.77. Figure 4.9 shows the distributions across simulations of the mean pairwise differences of 100 samples for each demographic model, under the TN93 model.

The mean of the simulated mean pairwise differences under the constant population size model is 97.15 - just slightly below the value that would be expected under the Coalescent Process, no significant difference from a one-sample t-test at 5% significance level. The distributions of the mean pairwise differences for the other four models are spread around this value, with the exponential growth model having a slightly lower and less variable distribution. The mean of the mean pairwise differences were 118.19, 101.86 and 83.84 for the step, bottleneck and exponential demographic models, respectively. There is almost twice as much genetic variation in our simulated data than there is in the realistic data set. However, given that there is not more variation than what would be expected under the properties of the constant Coalescent

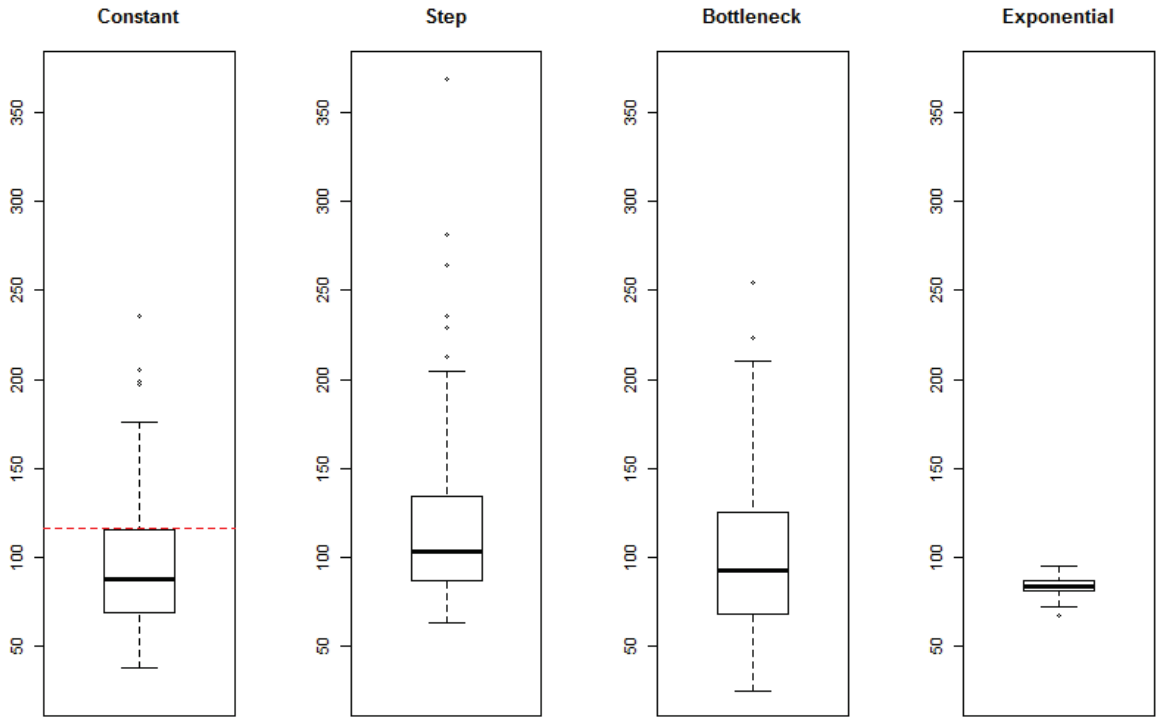


Figure 4.9: Boxplots showing the distribution of mean pairwise differences over 100 sets of 100 simulated DNA sequences for each of the four demographic models. The red dotted line shows the expected mean pairwise difference under the constant-size Coalescent Process.

Process then this meets our requirements for a *random* sample of DNA sequences.

Lastly, the distribution of the sample sizes of the non-random samples for each demographic model are presented in Figure 4.10. As these histograms show, each demographic model has a fairly similar distribution of subsample sizes. The minimum subsample size in each model is 34, and the maximum is 99 for the constant and step models and 97 for the bottleneck and exponential models. The mean subsample sizes are 57.67 in the constant model, 69.25 in the step, 70 in the bottleneck and 65 in the exponential model, and the median sizes are 55.5, 50, 52 and 49, respectively.

Consistently over all four models, there is a higher frequency of subsamples with sizes that lie between 35 and 50. This is due to the conditions around the subsample whereby the branch containing the mutation of interest has to be of a certain length (see Section 2.5.2).

The next stage in the simulation process was to use the Bayesian Skyline Plot model and infer the temporary changes in $N(t)$ from the randomly sampled sets of DNA sequences. Before presenting these results, the next three sections of this chapter will outline some issues in both the simulation process and the Bayesian Skyline Plot model. Once these issues have been addressed, the chapter will continue with inferring the population size from the simulated sets of DNA sequences.

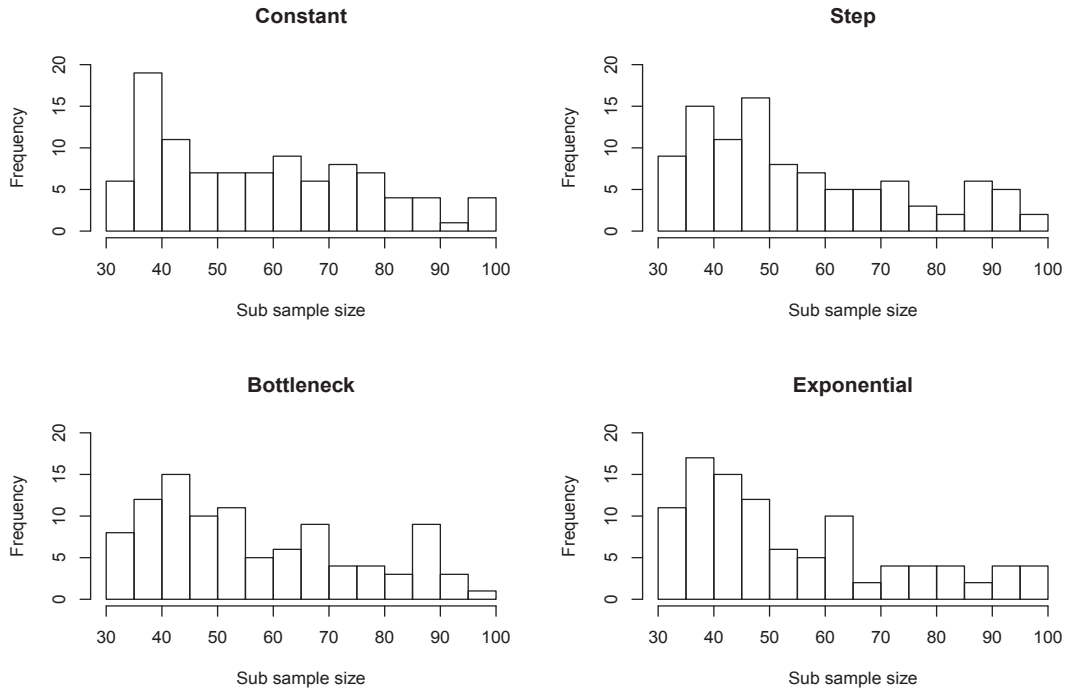


Figure 4.10: Histograms showing the distribution of subsample sizes for each demographic model.

4.2.2 A common time axis

To visualise the results of the Bayesian Skyline Plot model software called Tracer is used (Rambaut et al., 2018). However, given the large number of sets of DNA sequences to be analysed at one time, the raw MCMC samples from the posterior distribution were used. For each iteration in the chain, the model provides estimated population sizes for each of the ten groups of coalescent intervals. As well, each iteration in the chain has a sampled genealogy (consisting of the branching structure and the edge lengths). This information comes together and gives what are effectively *change points*, ten points in time where there is a change in the inferred population size that correspond to ten grouped coalescent intervals. Figure 4.11 illustrates this using five randomly chosen MCMC steps for one set of DNA sequences.

Note the variation between iterations in the population size estimates and the times at which population sizes change. To be able to calculate pointwise summaries, like the posterior mean or median at any time, there needs to be a common grid of time points at which to evaluate these.

This leaves two decisions to be made: how far back in time the time grid will go and how fine this grid will be. In Tracer, the user can use the posterior distribution of the sampled trees to determine this maximum time, by selecting the 2.5^{th} or 97.5^{th} percentiles, mean or median of the posterior distribution of depths of the sampled trees. Since, under simulation, the time genealogies are known, then the 95^{th} percentile of the depths of these trees would be a plausible *maximum time*. Due to the maximum time depending on the simulated genealogies, the

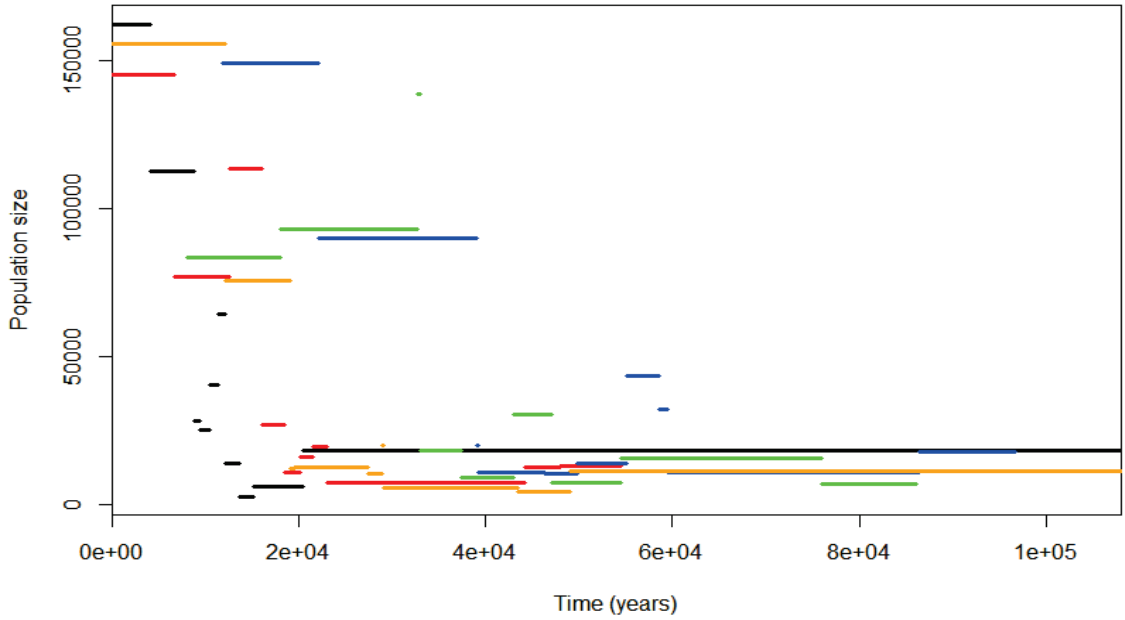


Figure 4.11: Example of Bayesian Skyline Plot model output for five randomly selected steps in the MCMC chain each showing different population sizes and change points. Each colour represents a different step.

time axes were then slightly different between demographic models; the maximum time for the constant population size model was 1.1×10^6 years, for the step model 1.4×10^6 years, for the bottleneck model 1.3×10^6 years and for the exponential model 2.9×10^5 years. These maximum time points are very long since it is believed that modern humans would not have existed more than 200,000 years ago (Cann et al., 1987). Therefore, once the inferred population size from the Bayesian Skyline Plot model was constructed, analysis continued only up to 200,000 years ago.

Once the maximum time point was decided, the number of grid points along the time axis has to be chosen. A 1,000 grid point was chosen to provide a fine resolution. At each grid point, the population sizes at that time was identified (it could belong to any of the groups) and the mean calculated, to give an estimate of the posterior mean population size at that time.

4.2.3 A truly random sample of DNA sequences?

Note that the '*random*' sample of DNA sequences is never truly random. This is because it technically depends on the non-randomly sampled DNA sequences due to the condition that the tree must contain

1. a branch that is long enough to contain at least one mutation in expectation and,

2. the subsample defined by that branch must be large enough to be worth performing inference on.

The first condition was in place to define the non-randomly sampled set of DNA sequences under the assumption that they would (likely) share a mutation of interest, to represent a sampling scheme that happens in reality, when DNA sequences are, for example, chosen from a haplogroup. To select this non-random sample in the simulation process, a full tree is simulated under the Coalescent Process and a branch long enough to contain the mutation of interest (at least in expectation) is identified. It is on this branch that the tree is cut and the present-day descendants form the non-random sample.

The second condition was in place firstly because the Bayesian Skyline Plot model requires there to be more nodes (or coalescent events) in the tree than there are groups (A from Section 3.3) in this case at least 10 (Drummond et al., 2005). Secondly, in the early stages of this work, it was found that very small samples (of less than around 30) produced very uninformative population estimates. Finally, sample sizes in the literature almost always at least 30 (Section 2.5.4) so it was decided that the non-random sample size be at least $n/3$, meaning that $n_s \geq 34$ here.

In the simulation process (described in Section 2.5), any tree that did not meet these requirements was discarded, and only trees that satisfied the two conditions above were kept and analysis continued assuming this a random sample. However, this meant that the *randomly* sampled trees were in fact dependent on the conditions defining the non-random sample. There should be a concern around treating this *quasi-random* sample (that meets the conditions above) as a *truly random* sample. However if the properties of the trees simulated under the quasi-random sample conditions were similar enough to properties of the trees simulated under truly random conditions, then these concerns would be relieved. Statistical tests were carried out to compare the properties of these trees and to compare the genetic variation between the two samples. The results are presented in this section.

Figure 4.12 shows boxplots comparing three properties there might be interest in. The first two boxplots show depths and lengths of 100 truly randomly generated trees and 100 quasi-randomly generated trees. The third boxplot shows the mean pairwise difference that accounts for the genetic variation in the 100 DNA sequences from the truly randomly generated trees and the 100 DNA sequences from the quasi-randomly generated trees.

Overall, there is only a small difference between the two samples. The quasi-random sample has a somewhat smaller tree depth, tree length and mean pairwise difference than what would be expected under the coalescent. In comparison, the truly random cases are more similar to the expected values from coalescent theory. The mean tree length was 10.70 for the truly random trees and 10.10 in the quasi-random case. The mean tree depth was 2.08 in the truly random and 1.82 in the quasi-random case. Lastly, the mean of the mean pairwise difference was 103.81 in the truly random and 97.15 in the quasi-random case. None of these differences are significant

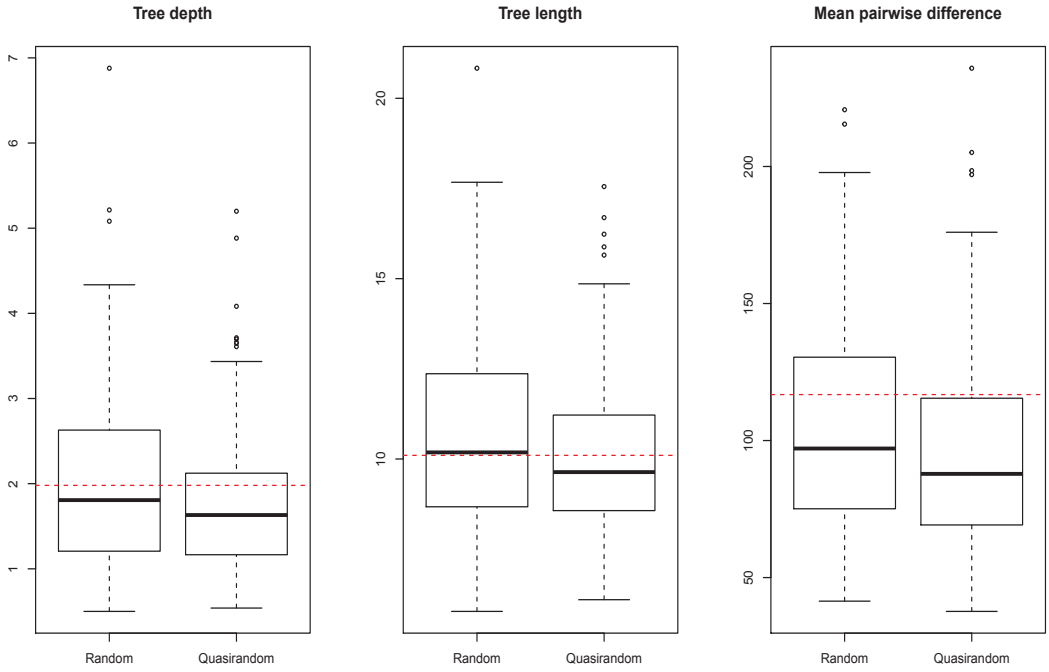


Figure 4.12: Boxplots comparing three properties of the truly randomly and quasi-randomly simulated samples. Each of these were simulated under a constant demographic model and the TN93 mutation model. Time is measured in coalescent units. The theoretical expectation of a random sample is shown (red dotted line).

(two-sample t test at 5% significance level).

At this point there is a convincing argument that there is little difference between the two samples. However, it made sense to check the results of the inferred population sizes from DNA sequences simulated from the truly random trees and compare these population curves to the quasi-random case. The population curves shown in Figure 4.13 are the mean of 100 posterior mean population sizes for each of the truly random samples and the quasi-random samples.

The inferred population sizes from both sets of samples (truly random and quasi-random) are very similar. Both population curves lie very close to each other all throughout the time axis, and the central 95% of pointwise posterior mean population sizes overlap significantly. The population size for the truly random sample is a little higher, probably due to these trees being a little deeper (see Figure 4.13) but the difference is much smaller than the width of the confidence bands.

Obtaining a completely random sample of DNA sequences to analyse is, ultimately, tricky. The nature of the work in this thesis needs a randomly generated coalescent tree from which we can obtain a random sample of DNA sequences and a corresponding non-random sample. To investigate any bias in the results of the inferred population sizes, it is very useful to be able to pair these two samples together. Therefore, this work continues using the quasi-random sampling method. There is no evidence that properties of the trees and the genetic variation

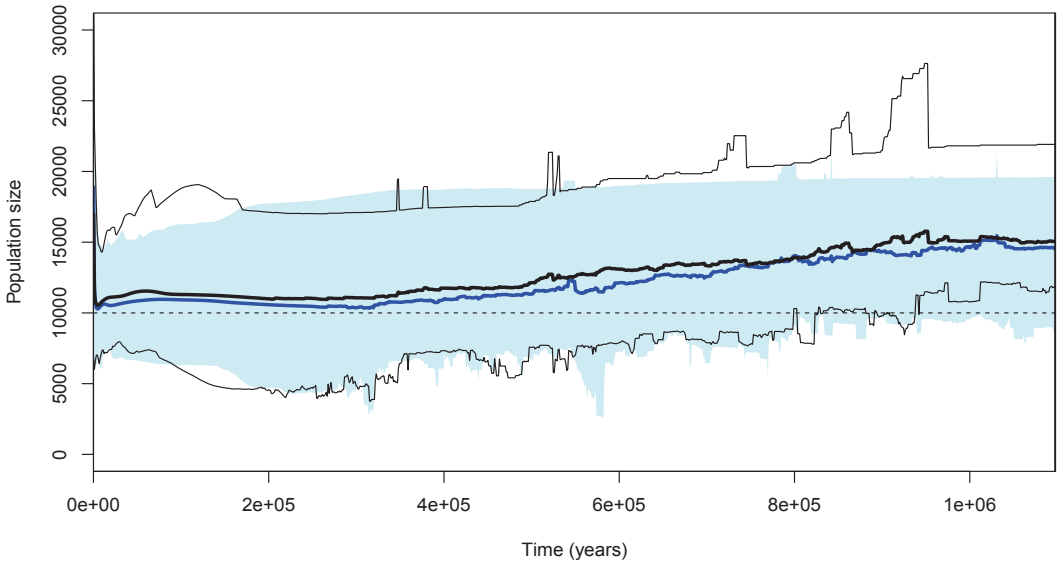


Figure 4.13: Population size curves from the Bayesian Skyline Plot model for DNA simulated under the constant demographic model. The plot shows the truly random sample's population curve (black solid line), the quasi-random sample's population curve (solid blue line) and the true population size (thin dotted line). The pale blue shaded area represents the 2.5^{th} and 97.5^{th} percentiles of the posterior means for the quasi-random samples, and the thin black lines the same but for the truly random samples.

between the truly random and quasi-random samples are meaningfully different.

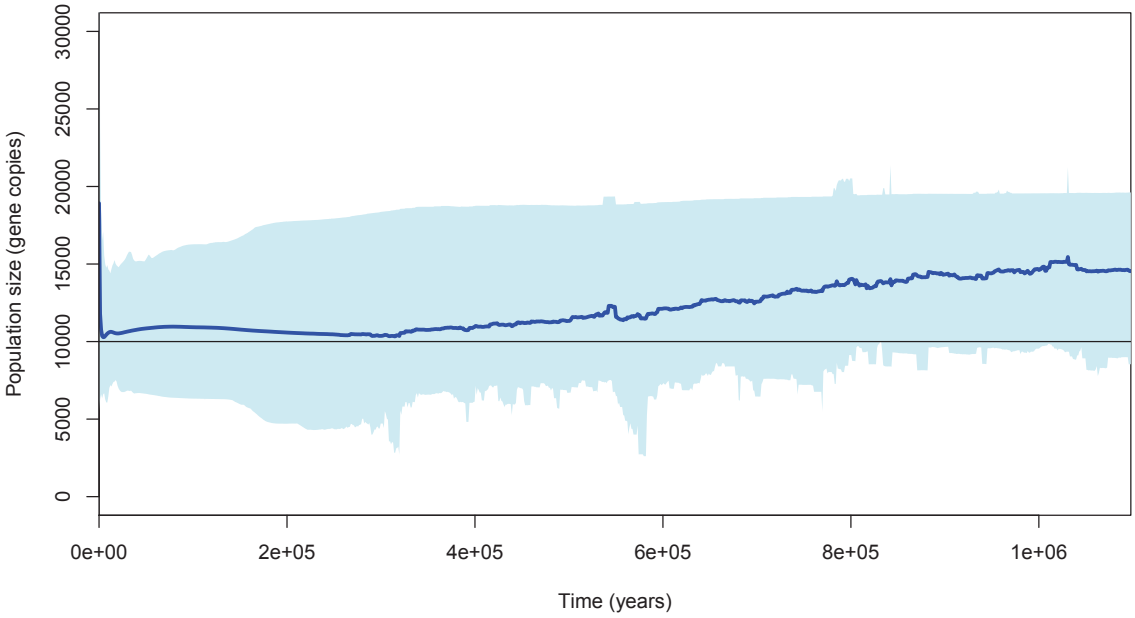
An analogous analysis was carried out on the step, bottleneck and exponential growth. The results are consistent with the constant case and are presented in Appendix C to avoid repetitiveness. Therefore, in the following sections, the term *randomly sampled DNA sequences* refers to this quasi-random sample, which we will assume does indeed not depart much from a random sample of sequences.

4.2.4 Bias in the population size trajectory

The last issue to overcome before the results of the Bayesian Skyline Plot model are presented highlights a small flaw in the BSP model itself. Figure 4.14 shows a summary of 100 posterior mean population size trajectories for random and for non-random samples (for data simulated from the constant size model with $N(t) = 10000$). In both of these plots there are peculiar features that cause one to question whether or not these are reasonable estimates of the population size going back in time.

First of all, in both cases, the further back in time we go, the higher the mean population size climbs. Secondly, in each case, although more so in the case of the subsample (4.14b), as we move back in time the uncertainty around the population size estimate becomes smaller. This

(a)



(b)

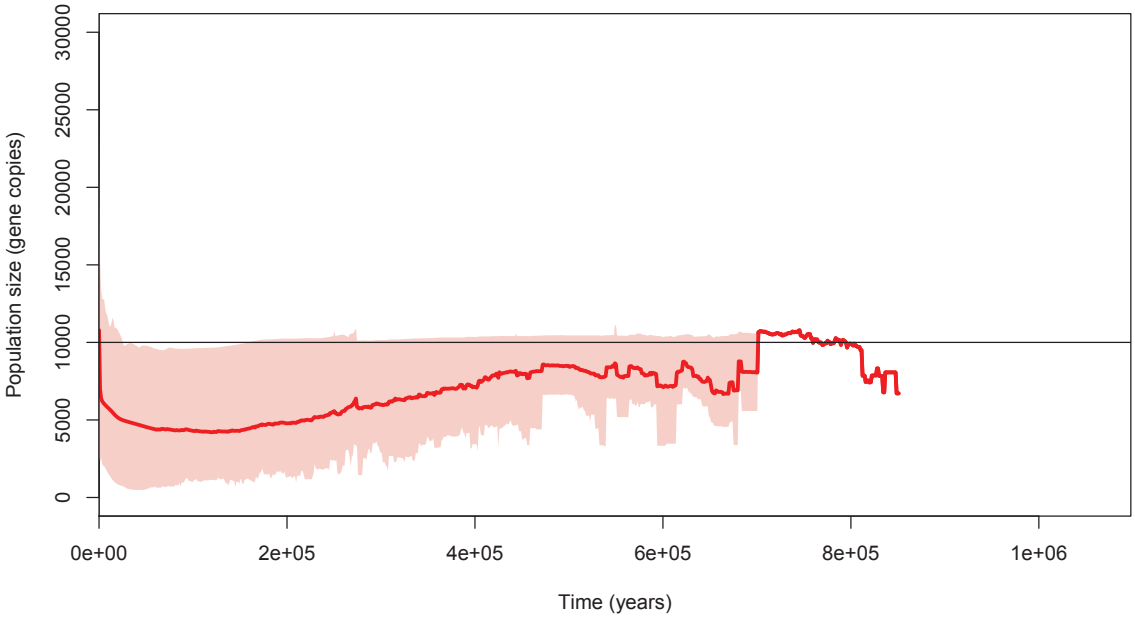


Figure 4.14: Mean inferred population size from the Bayesian Skyline Plot model for DNA sequences simulated under a constant demographic model and the TN93+Γ mutation model. a) The blue solid line represents the population size estimates for sets of randomly sampled DNA sequences and the blue shaded area is the 95% confidence interval. b) The red solid line represents the population size estimates for sets of non-randomly sampled DNA sequences and the red shaded area is the 95% confidence interval. In both plots the thin black line is the true population size.

seems highly unlikely. So, looking at these results there are two questions to answer:

1. Why does the mean population size increase as t increases?

2. Why does the population estimate become more accurate as t increases?

To begin with, consider the properties of our point estimate, $\hat{N}_i(t)$, the posterior mean population size as a function of time in the i^{th} simulation. The natural properties to look at are those which measure the quality of said estimator, e.g., the bias and root mean square error.

The bias of a point estimator, e.g. $\hat{N}_i(t)$, as a function of time is the difference between the expected value of the parameter and the true value of the parameter, in our case the expected value of the population size estimate at time t and the true population size at time t. It can be estimated by replacing expectation by an average over N_s independent simulations:

$$b(t) = \frac{1}{N_s}\sum_{i=1}^{N_s}\hat{N}_i(t) - N(t), \tag{4.5}$$

where $N_s = 100$ and represents the number of simulations and the function $N(t)$ is the true demographic model. The method is unbiased if $b(t)$ is close to zero at all t. The bias function calculated for both samples is shown in Figure 4.15a.

The second property of the point estimator is the root mean squared error. This measures the average distance between the estimator $\hat{N}_i(t)$ and the true parameter $N(t)$. The root mean squared error as a function of time is estimated from

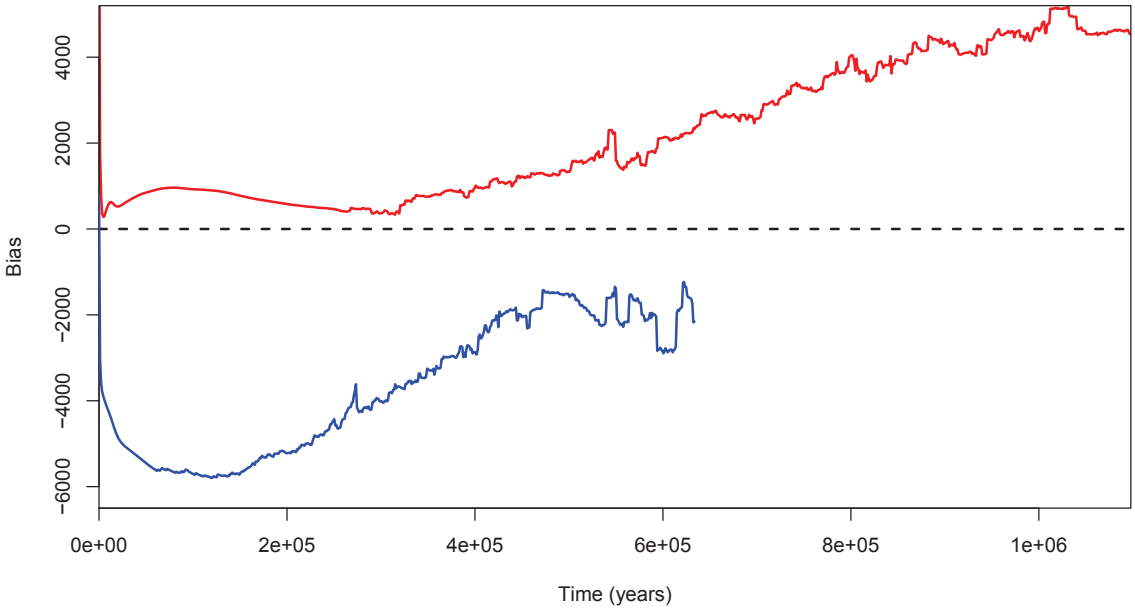
$$rmse(t) = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s}\left(\hat{N}_i(t) - N(t)\right)^2}. \tag{4.6}$$

Figure 4.15b shows the root mean squared error as a function of time for inferred population sizes for each sample.

Figure 4.15a tells a similar story to the plots in Figure 4.14. With the exception of the full sample for the first few hundred thousand years both estimates are biased, the bias of the estimator in the full sample case increasing as t increases. The population size estimator for the subsample case is consistently underestimating the true value of the population size. The error for the full sample very slowly increases as t increases, after a rapid initial drop. In the subsample case, curiously, the error generally decreases as t increases.

The explanation of this behaviour is as follows. Some trees sampled by the MCMC sampler will reach their most recent common ancestor before the end of the time grid. These trees can provide no population size estimate beyond that time: these values are missing. But they are not missing at random. They are missing when the tree is short. Short trees are what are expected for lower population size (higher genetic drift). So MCMC samples with lower population estimates

(a)



(b)

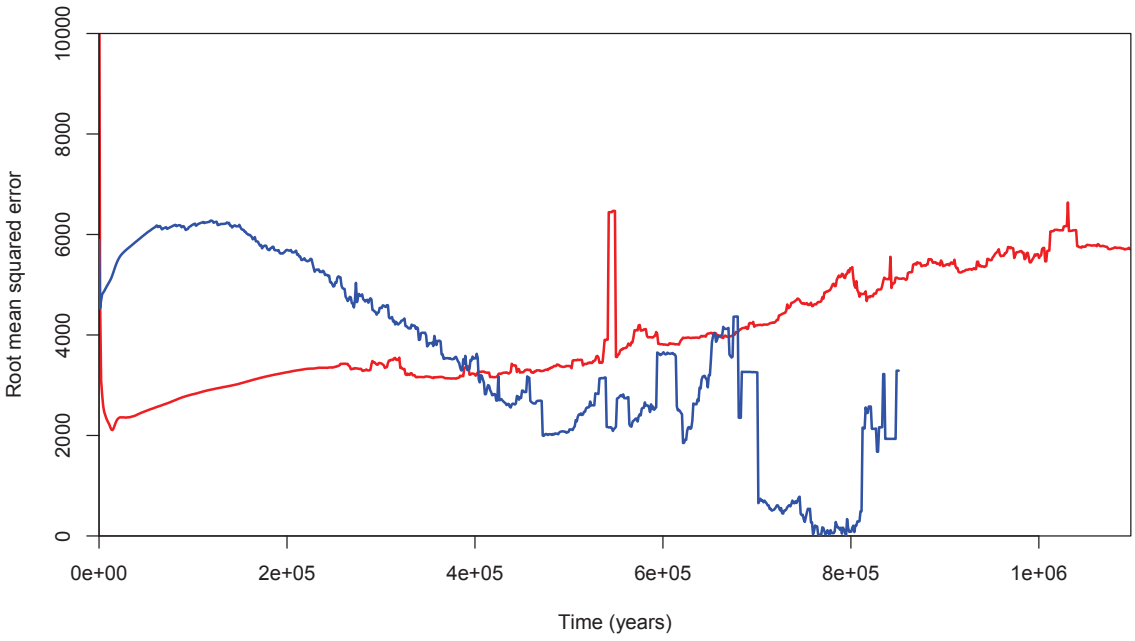


Figure 4.15: (a) The bias of the population size point estimator as a function of time and (b) the root mean squared error as a function of time. In both cases, the red line represents the full sample and the blue line the subsample.

are missing (for large times). As a result the estimates of population size is inflated. The larger the time, the more values are missing and the greater the bias.

On a higher level, when the summary over 100 simulations is calculated (e.g., when calculating the mean of the posterior means), there will also be missing values for some of the simulations the further back in time we travel. So not only is bias induced in each simulation as explained above, but also when summarising the 100 simulations. Figure 4.16 shows the number of simulations with missing population sizes moving back in time.

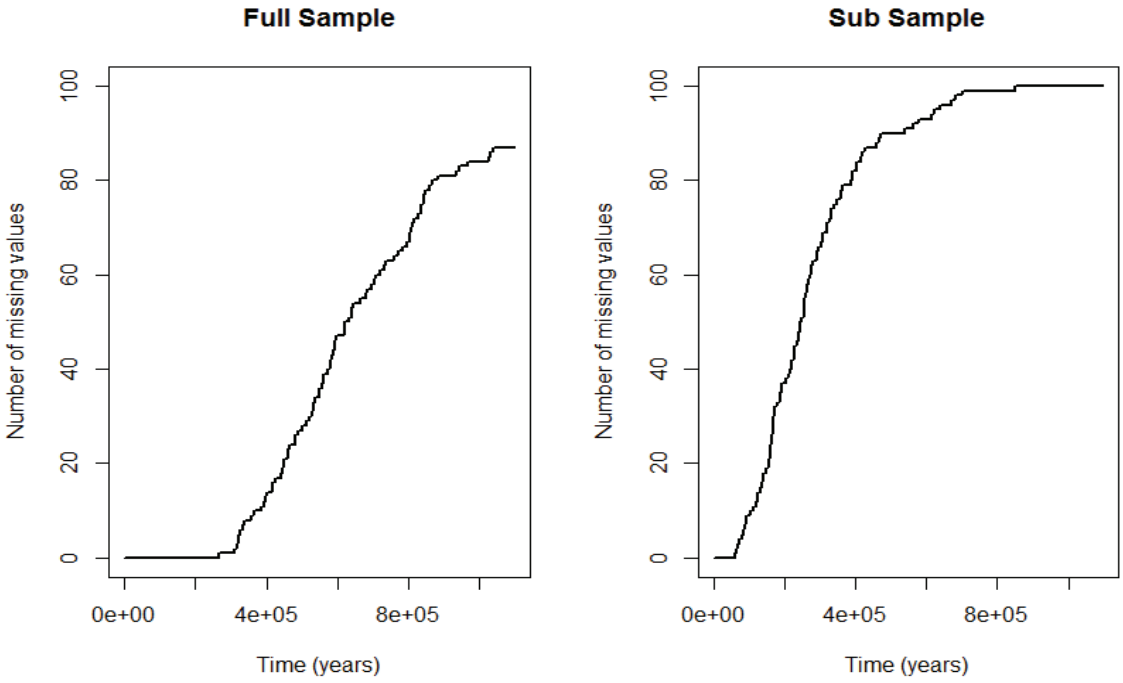


Figure 4.16: Number of missing population estimates from Bayesian Skyline Plot model over time from sequences simulated under a constant demographic model and the TN93 mutation model.

As seen from Figure 4.16, very early in time there begins to be some missing population sizes in the subsample but it is not until around 300,000 years ago that this happens for the full sample. So, the 'true' subtrees being shorter than the fuller trees is reflected in the inferences drawn by the Bayesian Skyline Plot model, as it has many missing values very early in time. This makes sense, since the chain is sampling over the range of possible trees, naturally some should be shorter than the true tree, but is potentially problematic for small subsamples where the tree is very short and then the population estimates are skewed (like in Figure 4.14b).

Two solutions to this problem are proposed.

1. For each MCMC step, if there is no population size estimate from a certain time point further back in time, because the MRCA has been reached, then the last estimated population size extends over these missing time points to give a constant estimate of population size

further back in time to the chosen maximum time point (Section 4.2.2). This corresponds to a model where the population size beyond the last group is assumed equal to that of the last group.

2. Another 'group' is introduced representing the time beyond the MRCA. Its population size is a priori correlated with the population size of the previous group in the same way as all the original groups (Equation 3.11). This means that the first missing population size will be replaced by a draw from the Exponential distribution with mean equal to the last estimated population size for that iteration. Then, the population size stays constant at this value until the maximum time point.

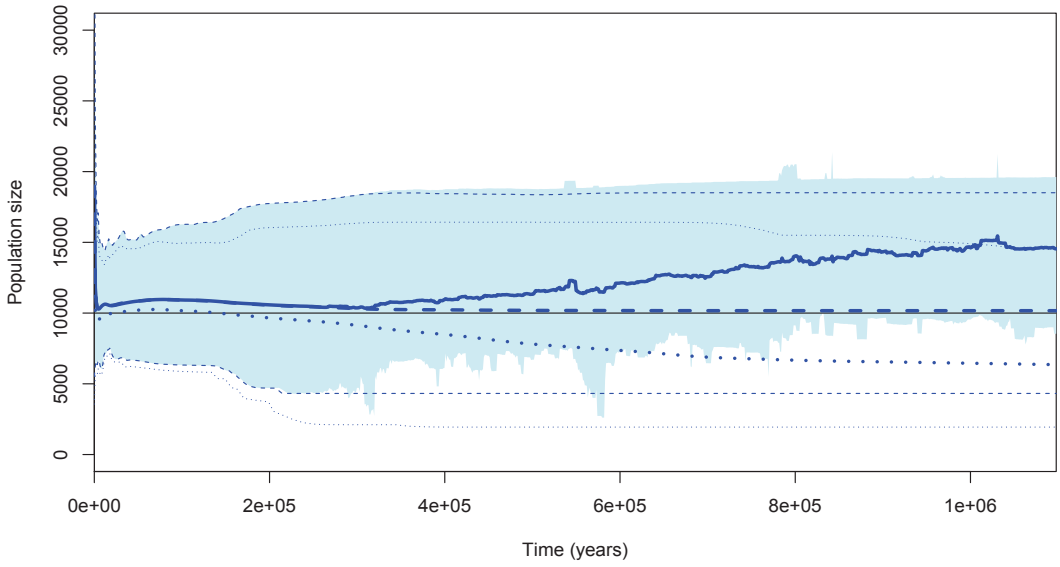
Figure 4.17 show both of these solutions implemented for the full sample and subsample, respectively. Figure 4.17(a) shows the resulting population size estimates for the full sample of DNA sequences. The original population size estimate is shown as before (in Figure 4.14(a)) and the proposed population sizes for the two solutions are superimposed for comparison. Solution 1 provides a very accurate estimate of the true population size. Solution 2 does not provide as accurate an estimate. Also noteworthy are the error bands, shown by the thinner lines. In particular for Solution 1 (the long-dashed lines) the error bands get wider the further back in time, as would be expected.

Similarly, Figure 4.17(b) shows some promising results. The uncertainty around the population size estimates remain relatively wide throughout, and the estimated population size displays no upward trend. In both the random and non-random sample cases, each population estimate produced by these solutions are much smoother since the high variability due to missing values has been eliminated.

Lastly, the bias and root mean squared error were calculated for these two new population estimates and compared to the originals in Figure 4.18. Given that Solution 1 results in the population estimate for the full sample being almost completely unbiased, and performing generally better than Solution 2 in both samples, this indicates that Solution 1 would be an improvement on the current population estimate from the Bayesian Skyline Plot model. Ultimately, it was decided that analysis would continue applying Solution 1, i.e., assigning a population size to the period beyond the MRCA equal to the population size of the last group.

It should be noted that this bias induced by the model is concentrated around population size estimates very far in to the past for the full sample of DNA sequences. Realistically, one would not be inferring modern human population sizes from one million years ago. However, it does pose a solution for the smaller subsample of DNA sequences that represent the non-randomly sampled set. Due to their trees being necessarily shorter than their corresponding full sample trees, there may be many *missing* population size estimates during relevant time periods.

(a)



(b)

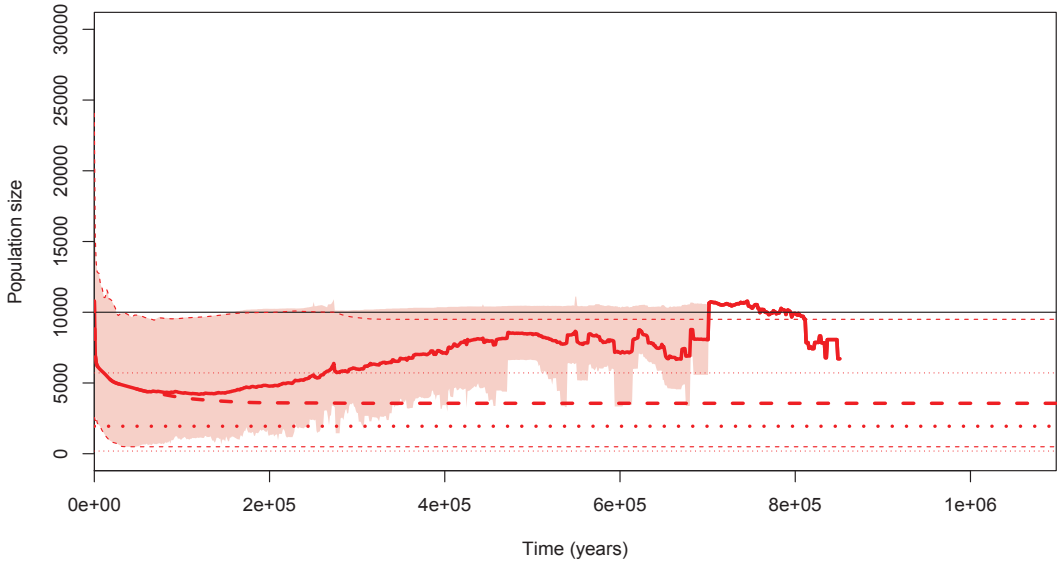
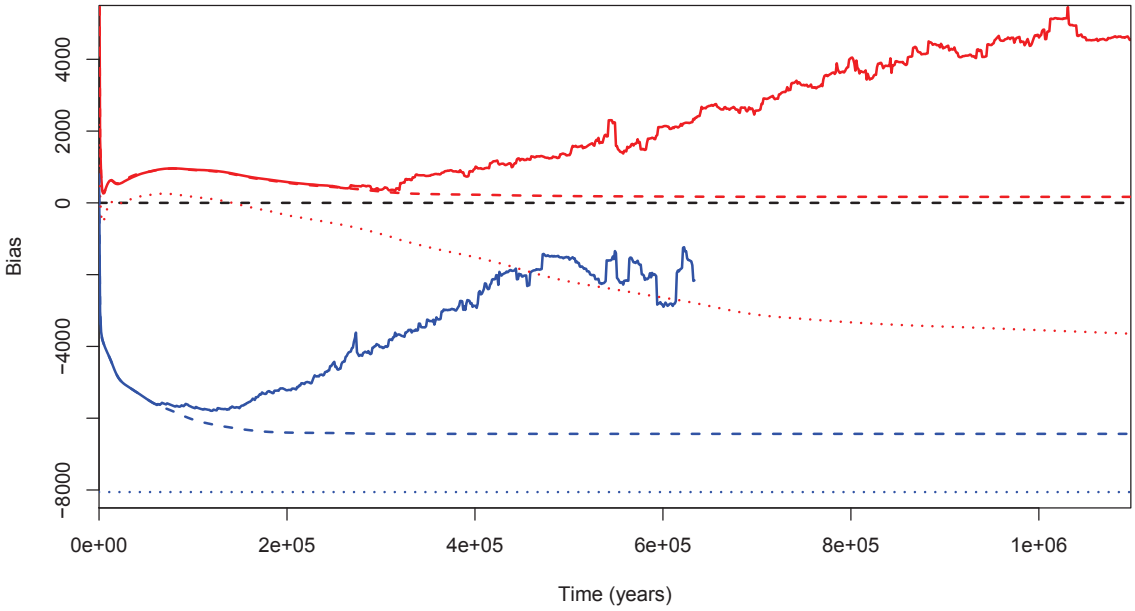


Figure 4.17: Means of the posterior mean population size estimates from the BSP model over 100 simulations for both random and non-random samples of DNA sequences. In both plots the true population is shown by the thin black line. (a) The solid blue curve is the BSP output for the random sample with blue shaded 2.5^{th} and 97.5^{th} percentiles. The blue long-dashed line is Solution 1, with corresponding percentiles shown by the blue thin long-dashed lines. The blue short-dashed line is Solution 2, with corresponding percentiles shown by the blue thin short-dashed lines. (b) The solid red curve is the BSP output for the non-random sample with red shaded 2.5^{th} and 97.5^{th} percentiles. The red long-dashed line is Solution 1, with corresponding percentiles shown by the red thin long-dashed lines. The red short-dashed line is Solution 2, with corresponding percentiles shown by the red thin short-dashed line.

(a)



(b)

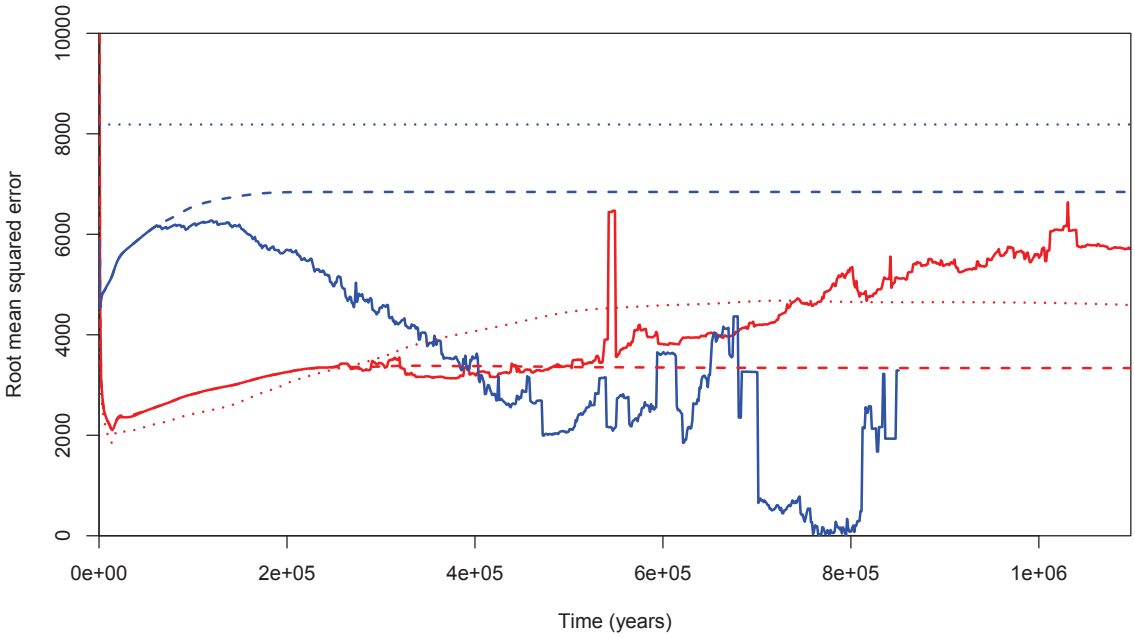


Figure 4.18: (a) The bias of the population size point estimator as a function of time and (b) the root mean squared error as a function of time. In both cases, the red solid line represents the full sample and the blue solid line the subsample for the original bias and root mean squared error. The long-dashed lines represent the bias and root mean squared error for population size estimates using Solution 1 and the short-dashed lines represent the bias and root mean squared error for population size estimates using Solution 2.

4.2.5 Bayesian Skyline Plot model output for simulated DNA sequences

This section will discuss the application of the Bayesian Skyline Plot model to the two sets of simulated DNA sequences. MCMC convergence was checked for each combination of demographic and mutation model in the same manner as was done for the Ingman data set (Section 4.1.3). For each demographic model there were 100 simulations, each with a sample size of 100. The results presented use Solution 1 to the problem of missing values in the population estimates. In each of these results, we are only concerned with the time period between the present day and 200,000 years ago.

Constant population size

Beginning with the constant population size demographic model, Figure 4.19 shows the inferred population size for the full *randomly* sampled set of DNA sequences and the sub *non-randomly* sampled set. The mean of the point estimate of each simulation, the posterior mean, $\hat{N}_i(t)$ for $i = 1, \ldots, N_s$ where $N_s = 100$ and is the number of simulations, is shown.

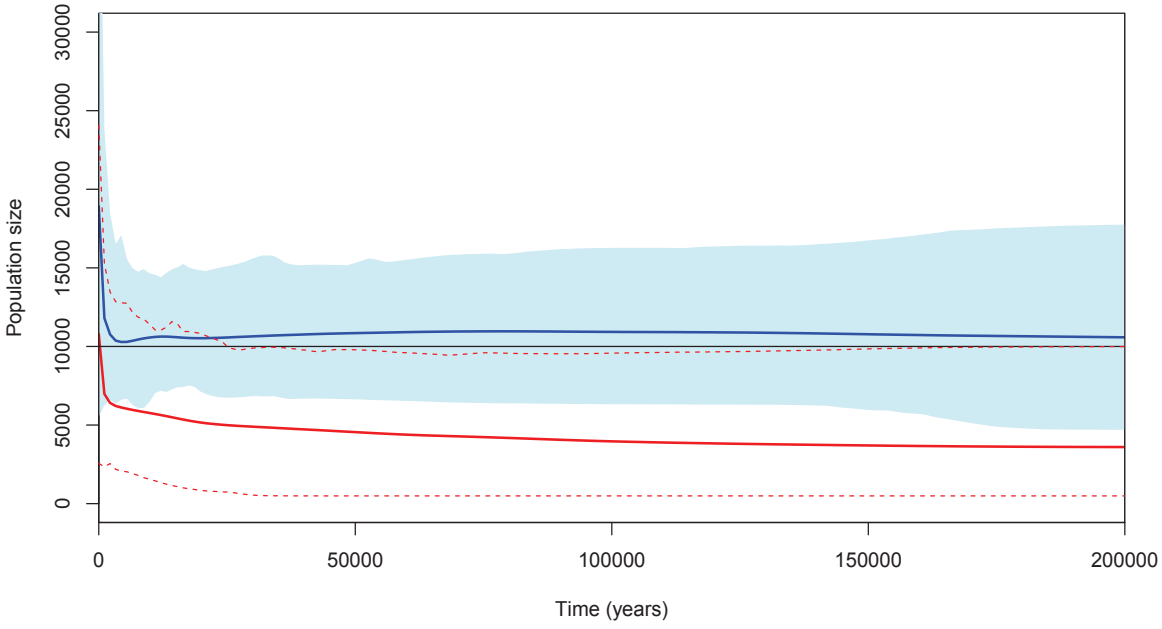


Figure 4.19: Bayesian Skyline Plot model output for DNA sequences simulated under a constant population size demographic model and TN93+Γ mutation model. The blue solid line represents the mean of 100 posterior mean population size estimates and the blue shaded area is the corresponding 2.5th and 97.5th percentiles of the posterior mean of the full sample. The red solid line represents the mean of 100 posterior mean population size estimates and the thin red dashed lines are the corresponding 2.5th and 97.5th percentiles of the posterior mean of the subsample. The true population size is shown by the black line.

The full samples of DNA sequences give a very good estimate of the population size, with just a slight overestimation. With the exception of the higher estimate for the initial population size (at time $t = 0$ in the present), the population size curve is almost a straight line showing that it captures the correct shape of this demographic model. The true population size lies comfortably within the 2.5^{th} and 97.5^{th} percentiles as well. Importantly, the subsample population curve also captures this pattern of a constant population size (with the exception of the higher estimate in the present) but this point estimate is consistently lower than in the full sample case. Apart from in the first few thousand years, the subsample point estimate lies outside the confidence bands for the full sample, and similarly the point estimate of the full sample lies outside the confidence bands for the subsample. Also worth noting is the fact that the true population size lies very much on the upper bound of the error bands for the subsample.

The consequences of the non-random sampling could simply be a scaling of the population size in this case. Also, given that in the case of the subsample, the Bayesian Skyline Plot model is drawing inferences from a much smaller number of DNA sequences, the fact that it so closely captures the correct shape of the demographic model is promising.

Step in population size

Next, consider a demographic model where population size abruptly grows in a 'step-wise' manner (Figure 4.20). As in the previous model, both samples capture the shape of the population size trajectory remarkably well. The estimate from the full sample follows the truth closely, with a little more discrepancy in the period from the present until 60,000 years ago. With the exception of the short period near the step, the inferred population confidence bands capture the true demographic model throughout.

On the other hand, the subsample population curve does not. The general shape is captured: a change in population size before and after 60,000 years ago is visible, but the population size estimate is too low in magnitude. During the first epoch, the population size point estimate for the subsample lies mostly below the confidence bands of the full sample. The confidence bands around the subsample are particularly wide however, and so do capture the point estimate population size of the full sample and, for the most part, the true demographic model. There is around 10,000 years (from 50,000 years ago) where the true population size lies above the subsample confidence bands. Then, in the second epoch, the estimates of population size become more accurate for both samples, with both confidence bands narrowing around the true population size of $N_1 = 10,000$. Still, the point estimate population size of the subsample is lower than that of the full sample.

For this demographic model, the Bayesian Skyline Plot model captures the underlying demographic model well, especially for the case of full randomly sampled sets of DNA sequences. Despite the non-random sample having smaller sample sizes, and therefore less information from genetic variation, the model still captures the shift in population size and for the most part

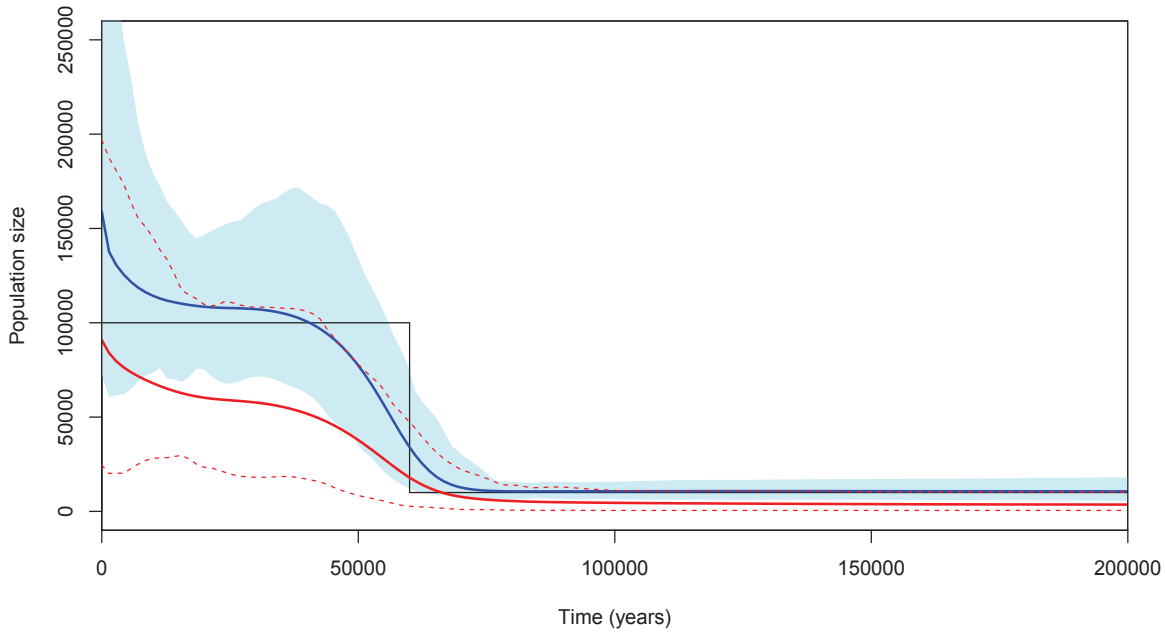


Figure 4.20: Bayesian Skyline Plot model output for DNA sequences simulated under a step population size demographic model and TN93+Γ mutation model. The blue solid line represents the mean of 100 posterior mean population size estimates and the blue shaded area is the corresponding 2.5th and 97.5th percentiles of the posterior mean of the full sample. The red solid line represents the mean of 100 posterior mean population size estimates and the thin red dashed lines are the corresponding 2.5th and 97.5th percentiles of the posterior mean of the subsample. The true population size is shown by the black line.

at the right point in time too. Again, like the constant population size model, it could be that a scale factor could be used to adjust the inferred population size from the non-random sample.

Population bottleneck

Next is the population bottleneck model. The Bayesian Skyline Plot model output for this demographic is shown in Figure 4.21 (population size now on a log scale).

Again the Bayesian Skyline Plot model captures this changing population size well. Focusing on the full sample, the initial population size is, as before, overestimated but the truth lies comfortably within the confidence bands. Then, at 23,000 years ago when there is a drop in population size (from 100,000 to only 1,000) the estimate does indeed capture this, although not quite the severity of the drop. The very small true population size lies outside the error bands, but as we enter the third epoch, the population size point estimate lies almost exactly on 10,000, the true population size.

Unfortunately, the subsample does not recover the bottleneck and really only provides evidence of a step. Although, as in the step model, the estimate does capture the initial change

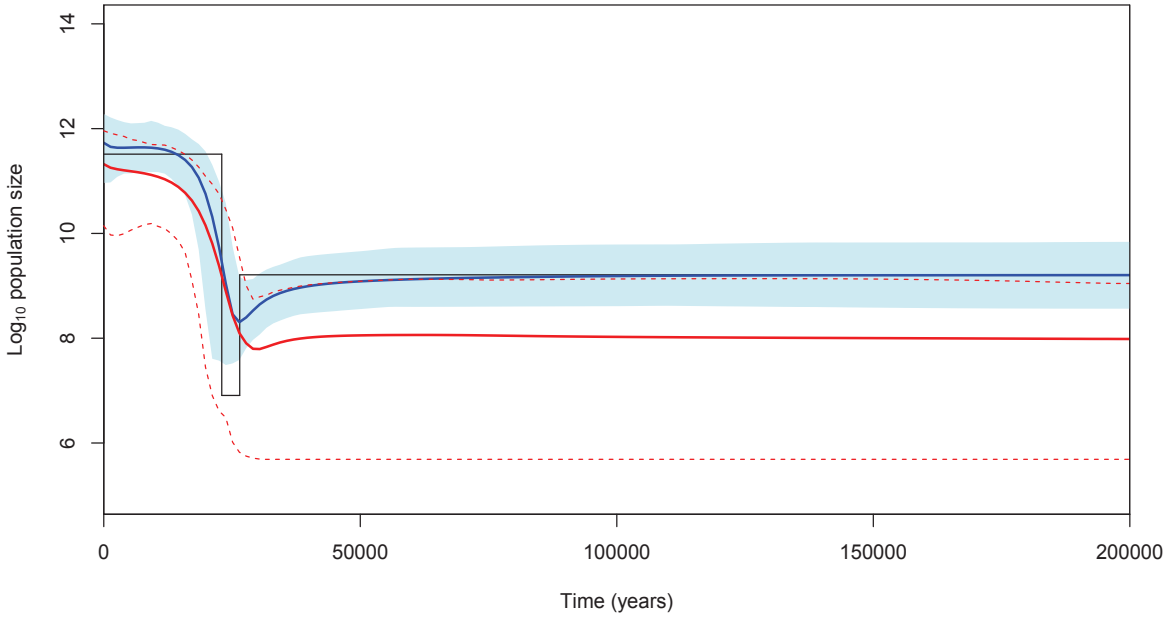


Figure 4.21: Bayesian Skyline Plot model output for DNA sequences simulated under a population bottleneck demographic model and TN93+Γ mutation model. The blue solid line represents the mean of 100 posterior mean population size estimates and the blue shaded area is the corresponding 2.5th and 97.5th percentiles of the posterior mean of the full sample. The red solid line represents the mean of 100 posterior mean population size estimates and the thin red dashed lines are the corresponding 2.5th and 97.5th percentiles of the posterior mean of the subsample. The true population size is shown by the black line.

in population size at the right point in time. Again, the population estimate is lower for the subsample than for the full sample. The population estimate for the subsample lies very much outside the confidence bands of the full sample during the third epoch, where the estimate is around 3,000, much lower than the true 10,000.

Given that this model is the most complicated of the four, it is perhaps not surprising that the Bayesian Skyline Plot model struggles to capture the truth particularly well for the subsamples. In addition, there is most likely not enough information from the non-random samples of DNA sequences to allow the Bayesian Skyline Plot model to infer these different population sizes.

Exponential growth in population size

Finally, the results of the exponential growth model are presented in Figure 4.22 (population size on a log scale).

Unlike for the previous three demographic models, the true population size is captured within the error bands of both samples throughout the entire time period between the present and 200,000 years ago. The population point estimate for the full sample tends to consistently

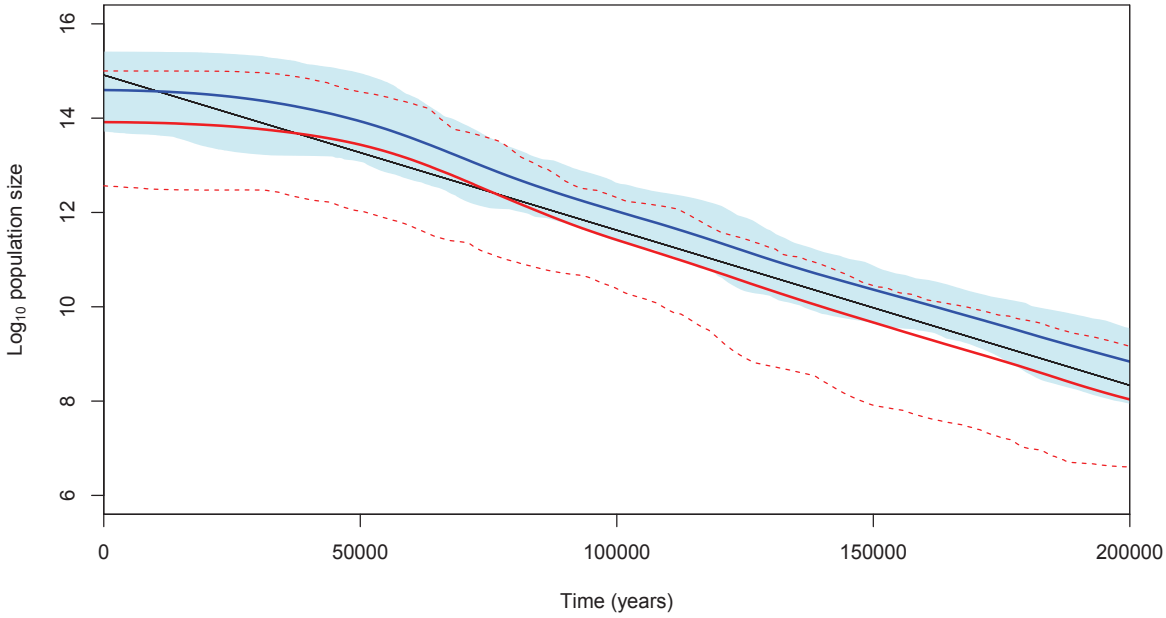


Figure 4.22: Bayesian Skyline Plot model output for DNA sequences simulated under an exponential demographic model and TN93+Γ mutation model. The blue solid line represents the mean of 100 posterior mean population size estimates and the blue shaded area is the corresponding 2.5^{th} and 97.5^{th} percentiles of the posterior mean of the full sample. The red solid line represents the mean of 100 posterior mean population size estimates and the thin red dashed lines are the corresponding 2.5^{th} and 97.5^{th} percentiles of the posterior mean of the subsample. The true population size is shown by the black line.

over estimate the truth, whilst the subsample does the opposite and underestimates the true population size. Furthermore, both point estimates of population size lie within the confidence bands of the other.

4.3 Discussion

The inferred population sizes from the Bayesian Skyline Plot model are promising. The model does remarkably well in capturing the true underlying population size of a set of DNA sequences that have been simulated under the Coalescent Process in particular, random sampling. From the results presented in Section 4.2.5 the estimates of population size from this model are manifestly reliable.

However, the inferred population sizes from the non-random samples are not as accurate and should be treated with caution. In all demographic models presented, the inferred population size from these samples was consistently underestimated by the model. Admittedly, the confidence bands around the point estimates are very wide, so that in most cases the truth or the inferred

population size from the full sample were captured within these.

When inferring the population size of a sample of DNA sequence that do not violate the underlying assumptions of the Coalescent Process, the Bayesian Skyline Plot produces very accurate estimates of the historic population size. However, once these assumptions are violated - as they are in any real analyses of DNA sequence data - the model fails to capture the true population size, most likely because of less genetic variation in the DNA sequences. Having more similar DNA sequences, as is the case for the simulated non-random samples in this study, leads to inferring shorter genealogies because the subsample will reach a common ancestor faster than the full sample would. As a result, the inferred population size is reduced.

Despite this, it is striking that in all models, the general shape of the demographic model was captured even with smaller *non-randomly* sampled sets of DNA sequences. It seems likely that there could be some way to account for the sampling scheme in practice, by exploring the relationship between the two population curves.

The results presented in this Chapter should not be over-interpreted, since the assumed demographic models are all very simplistic descriptions of historic human population patterns. A more elaborate demographic will be explored in Chapter 5.

The following chapter will now attempt to model the relationship between the population sizes inferred from the *randomly* and *non-randomly* sampled sets of DNA sequences. It would be useful if there was indeed only a scale factor between the two population size estimates that could then be applied in practice to results from non-randomly sampled data, when those samples are from a subtree of the full tree (such as samples from a mtDNA or Y chromosome haplogroup).

Chapter 5

Relating the Inferred Population Sizes from Random and Non-random Samples

This chapter will explore, in various ways, the relationship between the population sizes inferred from samples selected randomly and those selected non-randomly, when the inference is performed using the Bayesian Skyline Plot model, as described in Chapter 4. As discussed in Section 1.4, the purpose of this work is to explore the pitfalls of ignoring the sampling scheme and to give recommendations for the interpretation of population size analysis when the sample is from a subtree. As shown in Chapter 4, in general, the inferred population size is underestimated when a random sample is not analysed. The next step in this investigation is to quantify that underestimation which might then allow one to 'correct' results from non-random samples post hoc.

It would be useful to be able to provide a simple transformation that could be recommended to those inferring population sizes from samples taken from a subtree, perhaps as simple as a linear transformation, to provide a corrected inferred population size.

To be clear, a typical pipeline for such a study is the following.

1. A sample of DNA sequences that share a mutation is observed. Given that these sequences share a common mutation, the sample is not a random sample from the population. The mutation might be associated with a phenotype such as a disease, in which case the sample might consist of individuals having the disease. On the other hand, the mutation may have no known phenotypic association, such as most mutations defining mtDNA or Y chromosome haplogroups. For simplicity, assume the mutation occurs only once, so the sample is equivalently defined by an edge in the genealogy.

2. The population size as a function of time is inferred from this sample using the Bayesian Skyline Plot model.

3. Conclusions are then made based on this population size trajectory, ignoring the fact that the sample was non-randomly selected and thus the assumptions of the Coalescent Process

(the basis of the Bayesian Skyline Plot model) are violated.

Instead, this Chapter will propose a method that could account for the sampling scheme, resulting in a pipeline like the following.

1. A non-random sample of DNA sequences that share a mutation is observed.
2. The population size is inferred using the same methods as before.
3. Then, apply some transformation to the population size that could give results approximately as if from a random sample. The transformation might be as simple as

$$IPS_R = A \times IPS_{NR}, \tag{5.1}$$

   where IPS_{NR} are the inferred population sizes from the non-randomly sampled DNA sequences. The parameter A is a scalar multiplier and IPS_R is the corrected population size.
4. Interpretation proceeds from IPS_R, not IPS_{NR}.

So ultimately, the aim here is to *predict* the true population size, from IPS_{NR}, as accurately but as simply as possible. To do this, the inferred population sizes obtained from applying the Bayesian Skyline Plot model to the DNA sequences simulated under the conditions presented in Section 2.5 will be used, where we have inferred population sizes from 100 DNA sequences sampled randomly, corresponding inferred population sizes from sequences sampled non-randomly, as well as the true population size, under which the DNA sequences were simulated.

The response variable will be the inferred population size of the randomly sampled DNA sequences (or some variant of it) and the explanatory variable (or dependent variable) will be the inferred population size of the non-randomly sampled DNA sequences.

This modelling will be carried out in two sections. The first will explore the relationship between each of the population curves over 100 simulations for each demographic model. This will be done using functional data analysis and regression methods.

5.1 Exploring the relationship between population size estimates

To explore the relationship between the two sets of population curves, a simple regression setting is not adequate. For each simulation there is a posterior density of population sizes for each of the full and subsamples and corresponding posterior mean or median population sizes over a time range. To model these in the typical regression setting requires a summary over the 100 simulations at each time point, e.g., the mean of the posterior means. These summaries

were presented in Section 4.2.5 and are satisfactory for providing a general insight into the relationship between the inferred population sizes of the full samples, the inferred population sizes of the subsamples and the true demographic models. However, in terms of exploring the relationships between the curves, only focusing on the mean over simulations of the posterior mean population sizes removes lots of information and variation within the data.

Instead, a method that accounts for the variability in both the response and explanatory variables would be preferred. This exists in Functional Data Analysis (FDA) (Ramsay, 2004). FDA can account for variability around both the response and explanatory variables. It includes models of different types such as scalar-on-function, function-on-scalar and function-on-function models (Ramsay and Silverman, 2005). It is the latter model that we are most interested in here so that we can include population sizes as a function of time for both the full sample and the subsample. FDA is appropriate when one of the variables of interest can be considered as a smooth curve or function. Thus, FDA can be thought of as the statistical analysis of samples of curves - such as, in our case, inferred population curves over time for a randomly sampled set of DNA sequences (the response) and a non-randomly sampled set of DNA sequences (the explanatory variable).

FDA allows the response and explanatory variables to be defined on different time intervals, but in our case we have observations on a common discrete time axis (see the grid point discussion in Section 4.2.2) for each demographic model. In practice, we observe functional data at discrete time points with high frequency so that a functional observation $y_i(t)$ or $x_i(t)$ consists of m pairs (t_{ij}, y_{ij}) and (t_{ij}, x_{ij}), respectively, for $i = 1, \ldots, n_c$ curves and $j = 1, \ldots, m$ observations per curve and where y_{ij} and x_{ij} are the observed value of the ith curve at time t_{ij}. In this chapter we will focus on the range of t between the present ($t = 0$) and 200,000 years ago ($t = 200,000$).

The work in this section of the thesis uses the theory outlined by Ramsay and Silverman (2005) and by Kokoszka and Reimherr (2017). The model fitting was carried out in `R` (R Core Team, 2018), using Version 2.4.8 of the `fda` package (Ramsay et al., 2018). The code for model fitting was based around the code used by Ramsay, Hooker and Graves (2009, Ch.10).

Note that in this section of Chapter 5 we refer to the IPS_R and IPS_{NR} (as before). These refer to the inferred population sizes of randomly sampled DNA sequences and the inferred population sizes of non-randomly sampled DNA sequences. For each set of sampled DNA sequences we have 100 population sizes through time - let us call them population curves. Each of these 200 curves estimates the posterior mean population curve from 6 million iterations of the MCMC sampler from the posterior distribution of the Bayesian Skyline Plot model (see Chapter 3.3 for details) for the 100 randomly sampled and the corresponding 100 non-randomly sampled DNA sequences.

5.1.1 The functional linear model

One of the simplest functional linear models was first discussed under the name of the *varying coefficient model* (Hastie and Tibshirani, 1993) and is often referred to as the *concurrent* or *pointwise* model given that it relates the value of $y_i(t)$ to the value of $x_i(t)$ at the same time points t. The concurrent model is

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \varepsilon_i(t), \tag{5.2}$$

for $i = 1, \ldots n_c$, where n_c is the number of curves. Y_i is the i^{th} response curve consisting of the population size from the random sample as a function of time t, X_i is the i^{th} covariate curve namely, the population size from the non-random sample also as a function of time, with $\alpha(t)$ and $\beta(t)$ both functional parameters which represent the intercept function and regression coefficient function, respectively. The intercept function $\alpha(t)$ can be thought of as the coefficient of a scalar covariate whose value is always one and captures the variation in the response that does not depend on the covariate function. The error functions, $\varepsilon_i(t)$, are assumed to be independent and identically distributed Gaussian random functions over time. Ultimately, the difference between this model and the standard linear regression model is that we now have *functional* parameters (which are infinite dimensional objects) that must be estimated from a finite sample.

In the same way that we define scalar observations to be elements of $\mathbb{R}$ and vector observations to be elements of $\mathbb{R}^d$, where d is the dimension of the observed vector, we must define the space in which these objects exist. In FDA, functions observed at discrete time points are expressed as functional objects using basis expansions spanning the space of square integrable functions, denoted L^2. Each observed curve $y_i(t)$ and $x_i(t)$ belongs to the Hilbert space of square integrable functions defined on range T, denoted $L^2(T)$, with the inner product

$$\langle f, g \rangle = \int_T f(t)g(t)\mathrm{d}t \quad \forall f, g, \in L^2(T).$$

Horváth and Kokoszka (2012, Ch.2) give a detailed explanation of Hilbert space theory.

Before fitting the concurrent model, the first step in FDA is to construct the sets of smooth functional curves from their discrete observations. The functional data observed with error are y_{ij} and x_{ij} for the j^{th} observation of the i^{th} curve for each sample, where

$$y_{ij} = y_i(t_{ij}) + \varepsilon_{ij}, \text{ and } x_{ij} = x_i(t_{ij}) + \delta_{ij}, \quad \text{for } i = 1, \ldots, n_c, \text{ and } j = 1, \ldots, m, \tag{5.3}$$

where ε_{ij} and δ_{ij} denote the error in y_{ij} and x_{ij}, respectively. Depending on what demographic model we are analysing, the value of m changes, but is the same for both the random and non-random samples. For each demographic model m is fixed across simulations and the time grid is the same, so $t_{ij} = t_j$, say.

As stated previously, each curve belongs to a continuous time domain, but in practice we

observe the curve at discrete time points $t_1, \ldots, t_m$ so that, say, observations x_{ij} are a discrete representation of the infinite dimensional object $X_i(t)$. Incidentally, this is a benefit of FDA methods since it is a dimensionality reduction technique and improves computational efficiency. To move between the finite space and the infinite space, we convert the observed discrete samples into functional samples, i.e., curves, using smoothing techniques. Depending on the type of data under analysis, or the researchers' preference, a number of smoothing techniques can be used, such as regression splines, smoothing splines, P-splines or local polynomial smoothing (Zhang, 2013). Here B-splines were chosen since they are computationally efficient and numerically stable (Fahrmeir et al., 2013).

The curves $Y_i(t)$ and $X_i(t)$ can be considered to belong to a finite-dimensional space generated by two sets of basis functions $\{\phi_1(t), \ldots, \phi_p(t)\}$ and $\{\varphi_1(t), \ldots, \varphi_q(t)\}$ so that

$$Y_i(t) = \sum_{k=1}^{p} a_{ik}\phi_k(t) = \boldsymbol{\phi}(t)^T \mathbf{a}_i \quad \text{and} \quad X_i(t) = \sum_{k=1}^{q} b_{ik}\varphi_k(t) = \boldsymbol{\varphi}(t)^T \mathbf{b}_i, \tag{5.4}$$

where $\mathbf{a}_i$ and $\mathbf{b}_i$ are vectors of basis coefficients $(a_{i1}, \ldots, a_{ip})^\intercal$ and $(b_{i1}, \ldots, b_{iq})^\intercal$ respectively to be estimated for the ith curve and where $\boldsymbol{\phi}(t)$ and $\boldsymbol{\varphi}(t)$ are vectors of basis functions $(\phi_1(t), \ldots, \phi_p(t))^\intercal$ and $(\varphi_1(t), \ldots, \varphi_q(t))^\intercal$, respectively.

Spline smoothing methods provide an estimate of a smooth curve by fitting piecewise polynomials. A kth order spline is a piecewise polynomial function of order k and the points at which these segments join are referred to as *knots*. The order k is the degree of the polynomial plus one and there are constraints imposed on the piecewise polynomial function that ensure continuity of the piecewise curves (Bowman and Azzalini, 1997). The most commonly used splines are cubic splines (order four), which are used here. The coefficients $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ of the curves $Y_i(t)$ and $X_i(t)$, respectively, can be estimated by minimising the residual sum of squares:

$$(\mathbf{y}_i - \phi_i \mathbf{a}_i)^\intercal (\mathbf{y}_i - \phi_i \mathbf{a}_i) \quad \text{and} \quad (\mathbf{x}_i - \varphi_i \mathbf{b}_i)^\intercal (\mathbf{x}_i - \varphi_i \mathbf{b}_i),$$

where the matrices $\phi_i = \big(\phi_k(t_{ij})\big)_{m_i \times p}$ and $\varphi_i = \big(\varphi_k(t_{ij})\big)_{m_i \times q}$ for each i, giving the least squares estimates of $\mathbf{a}_i$ and $\mathbf{b}_i$ as

$$\hat{\mathbf{a}}_i = (\phi_i^\intercal \phi_i)^{-1} \phi_i^\intercal \mathbf{y}_i \quad \text{and} \quad \hat{\mathbf{b}}_i = (\varphi_i^\intercal \varphi_i)^{-1} \varphi_i^\intercal \mathbf{x}_i,$$

respectively, and the resulting curves $\boldsymbol{\phi}(t)^\intercal \hat{\mathbf{a}}_i$ and $\boldsymbol{\varphi}(t)^\intercal \hat{\mathbf{b}}_i$ are known as regression splines. However, since in FDA it is common to place a basis function at every observation (i.e., saturate the model), then to prevent an over-fitted curve, the smoothness of the curve can be controlled using a *penalised regression spline*. In penalised regression splines, there are roughness penalty terms $\lambda_1 \mathbf{a}_i^\intercal \mathrm{D} \mathbf{a}_i$ and $\lambda_2 \mathbf{b}_i^\intercal \mathrm{D} \mathbf{b}_i$ for each curve which are added on to each of the corresponding sum of squares. Note that we can allow for both the basis functions and the roughness penalty to vary from one curve to the next, since each curve might require a different level of smoothing. The

coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ are then estimated as

$$(\phi_i^\intercal \phi_i + \lambda_1 D_1)^{-1} \phi_i^\intercal \mathbf{y}_i \quad \text{and} \quad (\varphi_i^\intercal \varphi_i + \lambda_2 D_2)^{-1} \varphi_i^\intercal \mathbf{x}_i,$$

where λ_1 and λ_2 represents the smoothing parameter in each case and D_1 and D_2 the roughness penalty. The smoothing parameters are non-negative and their purpose is not surprisingly to influence the overall smoothness of the functions. When they equal zero there is no penalty attached to the respective function and so each function will pass through each data point. As they approach infinity, the function becomes increasingly smooth. This parameter can be selected using methods such as GCV, AIC, BIC, etc., and the method selected in this case will be explained in more detail in the following section. The roughness penalties, D_1 and D_2, are the integrated square of the second order derivative of the function since the second derivative measures curvature. Matrices D_1 and D_2 are square non-negative definite matrices with order equal to the number of basis functions chosen, where the ij^{th} entry of D_2 is $\int \varphi_i''(t)\varphi_j''(t)\mathrm{d}t$ and analogously for D_1.

Now that we have obtained our sample functional curves, we can continue with our functional linear model (Equation 5.2). To obtain estimates of $\alpha(t)$ and $\beta(t)$ another discrete expansion is used. The intercept function, $\alpha(t)$, is represented as an expansion of F basis functions $\{B_f(t) : f = 1,\ldots,F\}$ by

$$\alpha(t) = \sum_{f=1}^{F} \alpha_f B_f(t),$$

where F is large. Similarly, the regression coefficient function, $\beta(t)$, is represented by

$$\beta(t) = \sum_{g=1}^{G} \beta_g B_g(t),$$

for large G. Again, cubic splines are used for these basis functions. The unknown parameters are now the scalars α_f $(f = 1,\ldots,F)$ and β_g $g = 1,\ldots,G$, which are organised into (high-dimensional) vectors,

$$\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_F]^\intercal \text{ and } \boldsymbol{\beta} = [\beta_1,\ldots,\beta_G]^\intercal.$$

We can now specify a penalised least squares criterion, starting by defining the sum of squared residuals as

$$S(\alpha,\beta) = \sum_{i=1}^{n_c}\sum_{j=1}^{m} \Big(Y_i(t_j) - \mu_i(t_j;\alpha,\beta)\Big)^2, \tag{5.5}$$

where

$$\mu_i(t;\alpha,\beta) = \sum_{f=1}^{F} \alpha_f B_f(t) + \sum_{g=1}^{G} \beta_g B_g(t) X_i(t).$$

Then, penalty terms are applied to each of the intercept and regression coefficient functions again using the integral of the second derivative. These steps are described by Kokoszka and

Reimherr (Ch.5, 2017). Ultimately, the penalised least squares estimates are obtained by minimising

$$S(\alpha,\beta)+\lambda_3\boldsymbol{\alpha}^T D_3\boldsymbol{\alpha}+\lambda_4\boldsymbol{\beta}^T D_4\boldsymbol{\beta}, \tag{5.6}$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Care should be taken when choosing the smoothing parameter and there are various methods for doing so (Kokoszka and Reimherr, 2017), but here generalised cross validation (GCV) (Craven and Wahba, 1979) was used as it most commonly used in FDA. This was implemented in `R` using the `fda` package (Ramsay et al., 2009, 2018). In general, we expect that as we increase the value of the smoothing parameter, our smooth curve becomes more flat and the bias increases and, as we decrease the value of the smoothing parameter, the smooth curve becomes less smooth and the variance increases (Bowman and Azzalini, 1997).

Then, as in the case of simple linear regression modelling, to test the significance of a parameter in functional linear modelling we can derive a functional version of the univariate F-test (Snedecor and Cochran, 1980). This tests whether any of the independent variables are significant in a multiple linear regression model with an alternative hypothesis that at least one coefficient is equal to zero. The result from the `fda` package in `R` is a plot of the observed F-statistic along with the pointwise and maximal F-statistics and their corresponding permutation-based critical values (Ramsay et al., 2009).

5.1.2 Concurrent model results

The four demographic models were analysed separately and, to avoid repetition, only the constant population size model will be presented in this chapter. The results of the step, bottleneck and exponential models can be found in Appendix D. However, a full discussion of the results will be presented in Section 5.1.3.

To begin, look at the 100 population curves for the random samples of DNA sequences and the paired 100 population curves for the non-random samples (Figure 5.1). Note that these are the estimated posterior mean population size over time for each simulation. Contrast this to the results presented in Chapter 4 which summarised the 100 curves by taking the mean over simulations. The curves of Figure 5.1 will be referred to as ‘raw’ population curves.

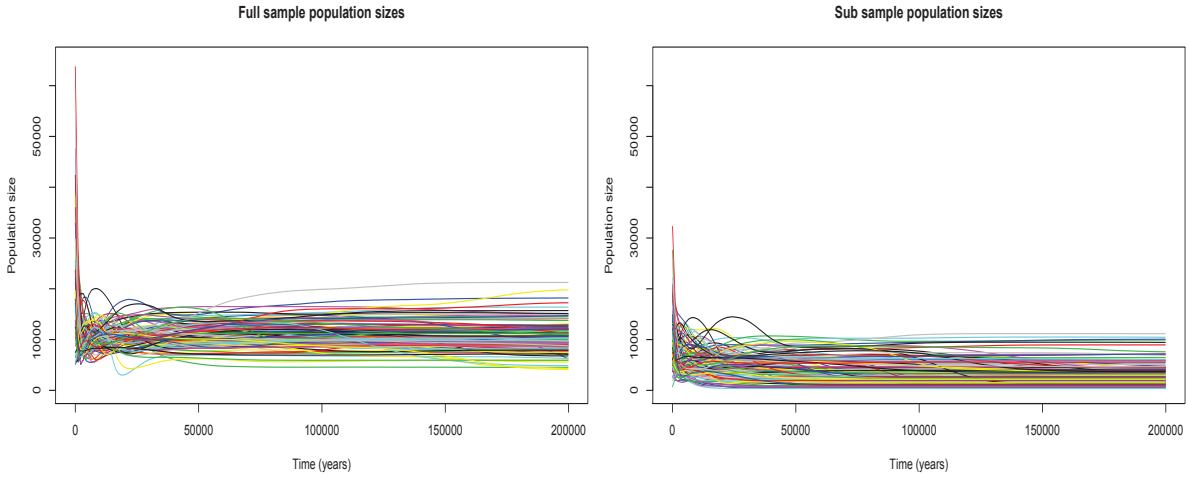


Figure 5.1: Population curves for the 100 randomly and non-randomly sampled sets of DNA sequences simulated under a constant population size demographic model.

For both the full sample and the subsample, the population sizes at the present (time $t = 0$) are quite variable for a short time. The initial estimates of the population size for the full sample range to very high values, whereas the subsample stays much lower. Both sets of curves capture the true underlying demographic model of a constant population size generally rather well, although there is a clear underestimate in the subsample case. This is consistent with the summaries presented in Section 4.2.5.

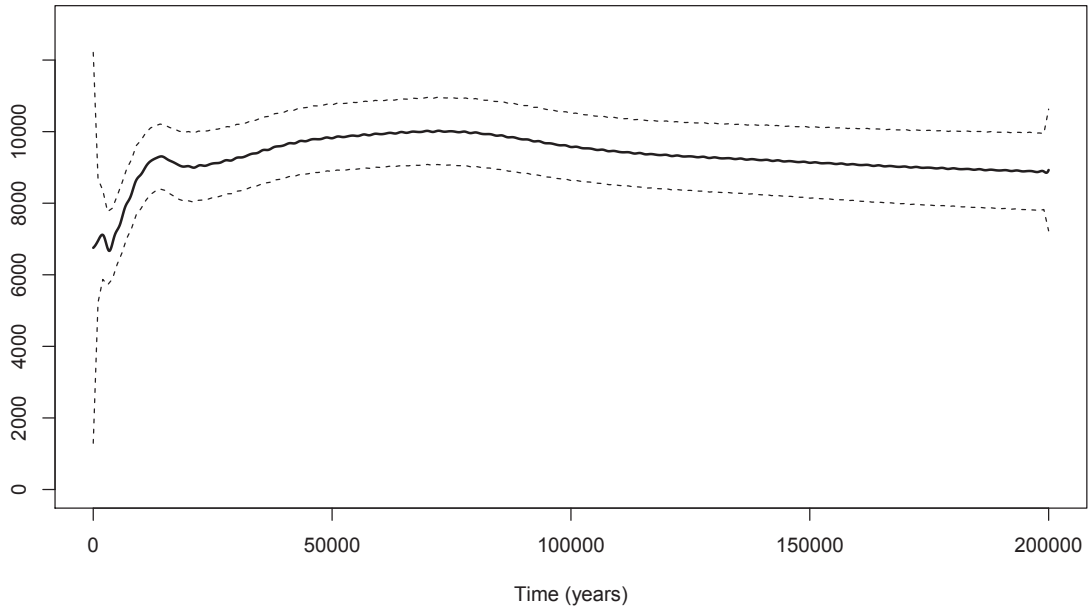
Given that there are different levels of variability in each set of curves, smoothing over individual curves is carried out separately across the two sets. There were 183 basis functions, one for each observation in the time range between the present and 200,000 years ago, and the roughness penalty was allowed to differ for each sample (but was kept constant across the 100 curves for computational ease). The GCV statistic was used to select the best fitting smooth function for the raw population curves, giving $\lambda_1 = 55000$ (full sample) and $\lambda_2 = 83000$ (subsample). With the smoothed curves the concurrent model was fitted and the intercept and regression coefficient function estimates are shown in Figure 5.2.

The intercept curve rises rapidly to about 10,000 and the 95% confidence intervals simultaneously shrink. The coefficient curve starts higher at the present and then drops time increases. Zero is never contained in the 95% pointwise confidence intervals and therefore the regressor is significant in the model. The regression coefficient then stays relatively constant through time, with the 95% confidence intervals widening slightly further back in time.

We can formally test the significance of the explanatory variable in the model using the functional F-statistic which tests the hypothesis

$$\beta(t) = 0 \quad \text{vs} \quad \beta(t) \neq 0.$$

(a)



(b)

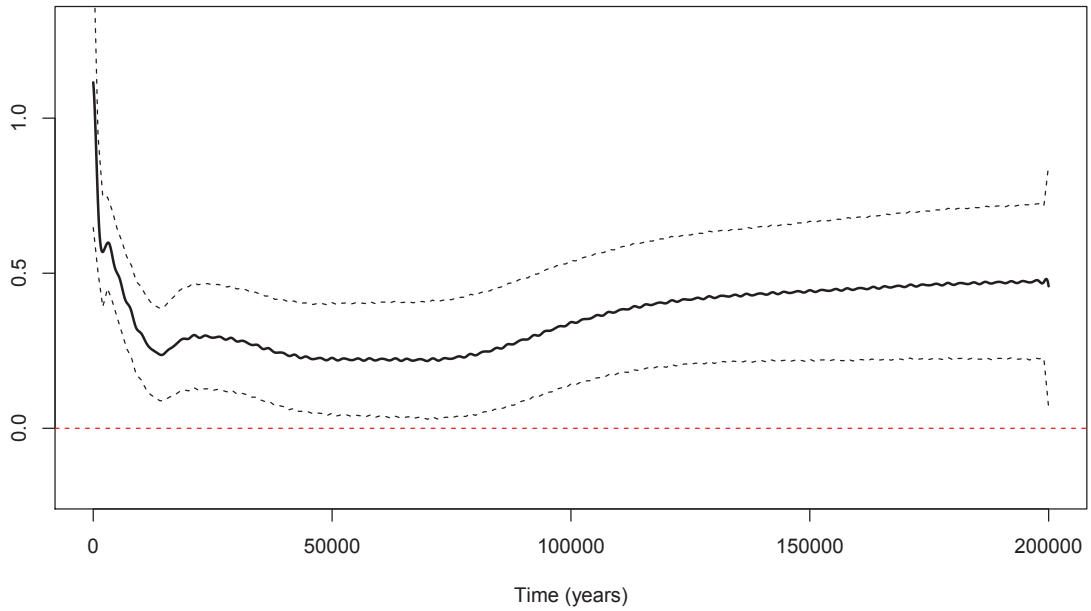


Figure 5.2: Constant demographic model. Intercept (a) and regression coefficient (b) functions for the concurrent model relating the non-random sample estimates to the random sample estimates, with 95% pointwise confidence intervals. The red dotted line sits at 0.

That is to say that we are testing whether or not the coefficient of the subsample's population size is significant in the model explaining the full sample's population size. The result of the functional F-test is shown in Figure 5.3.

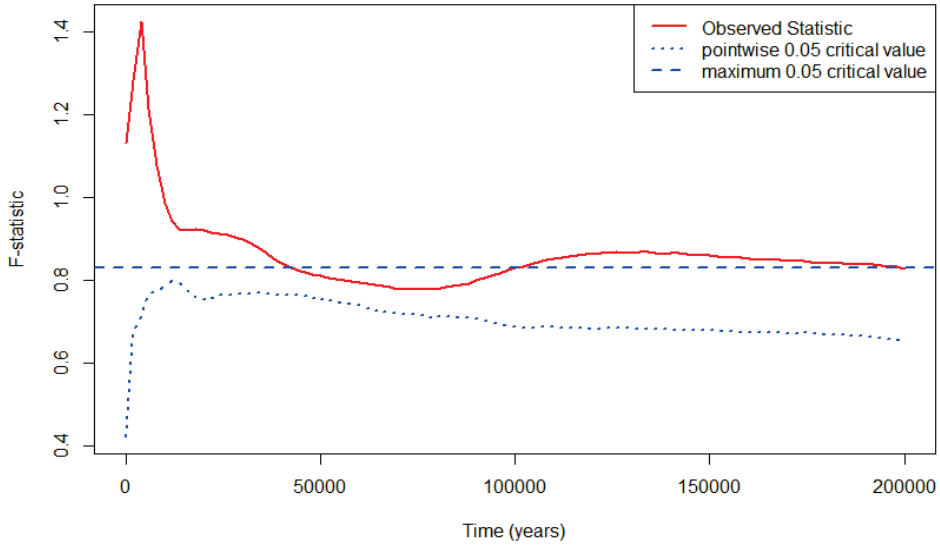


Figure 5.3: The result of the permutation F-test (testing for no effect) for the concurrent model for a constant population demographic model, at the 5% significance level. The red line represents the observed F-statistic and the horizontal blue dashed line is the maximum critical value at 5% significance.

The observed statistic is greater than the maximum critical value (at the 5% significance level) at most time points and so we can reject the null hypothesis and conclude that the coefficient of the population size of the non-random sample is not equal to zero. The variable is significant in the model describing the relationship between the population size of the randomly and non-randomly sampled sets of DNA sequences.

Secondly, it is possible to look at each time point individually and assess the significance of the covariate at that time point. If the observed statistic is greater than the pointwise critical value at 5% significance, then we reject the null hypothesis for that time point and conclude that the coefficient for that time point is not equal to zero. In Figure 5.3, the pointwise critical value is always less than the observed test statistic and so the population size of the non-randomly sampled set of DNA sequence is significant in the model explaining the population size of the randomly sampled DNA sequences at every time point.

Given that the purpose of this work is to investigate whether or not analysing a non-random sample of DNA sequences gives biased results compared to the population size estimate from a sample of random DNA sequences, another check would be to look at the inferred population size from the random sample and compare that to the predicted population size using the population size from the non-random sample as a predictor in the concurrent model. A random selection from four simulations is shown in Figure 5.4.

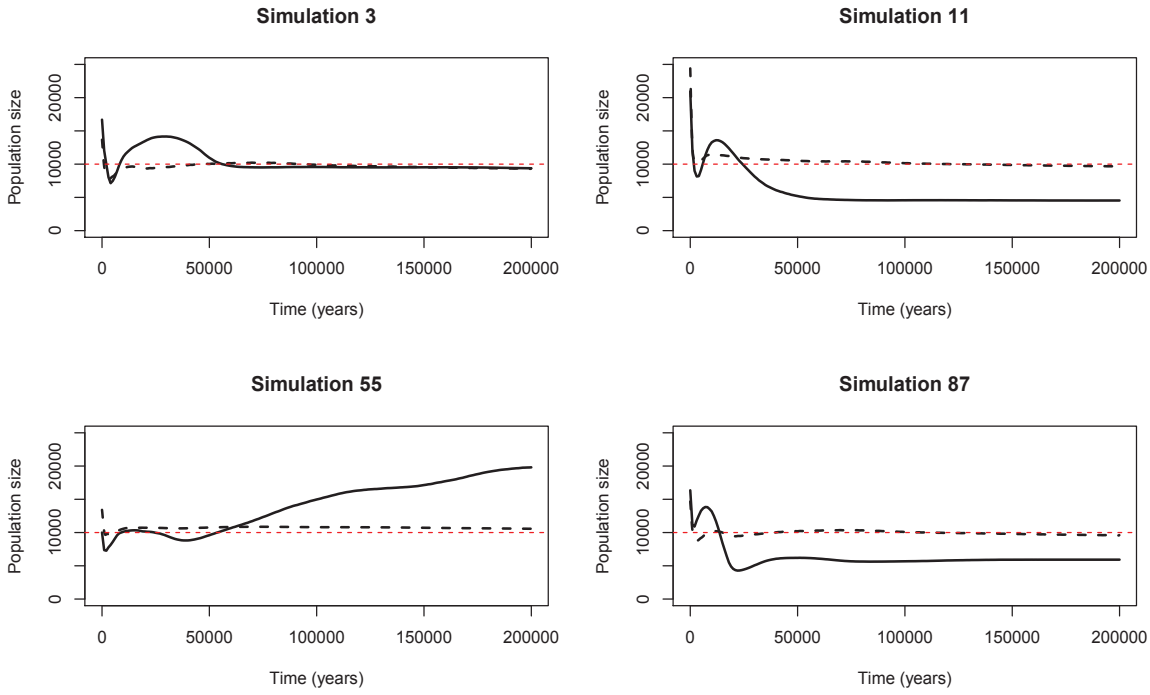


Figure 5.4: A selection of inferred and predicted population sizes for the random sample of DNA sequences under a constant population size demographic model. The inferred population size is the black solid line, the black dotted line is the predicted population size and the red dotted line is the true population size, $N_0 = 10000$.

In all simulations (not only those shown here) the model does extraordinarily well at predicting the true underlying population size from the inferred subsample population size. This is presumably since the fitted intercept function is on its own a very good predictor of the population size. Since, in reality the response variable would be entirely unknown, it would be useful to test the possibility that we could use only information from IPS_{NR} to model IPS_R. The next step, therefore, is to remove this intercept term from the model and fit the following model,

$$Y_i(t) = \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100),$$

where the only functional parameter to estimate now is the regression coefficient $\beta(t)$. The significance of the population size of non-random samples variable in the model is explored in the functional F-test of Figure 5.5. The term is significant in the model, since the observed statistic is greater than the maximum critical value at at least one point in time. In fact, the variable is significant at every time point since the pointwise critical value is always below the observed test statistic.

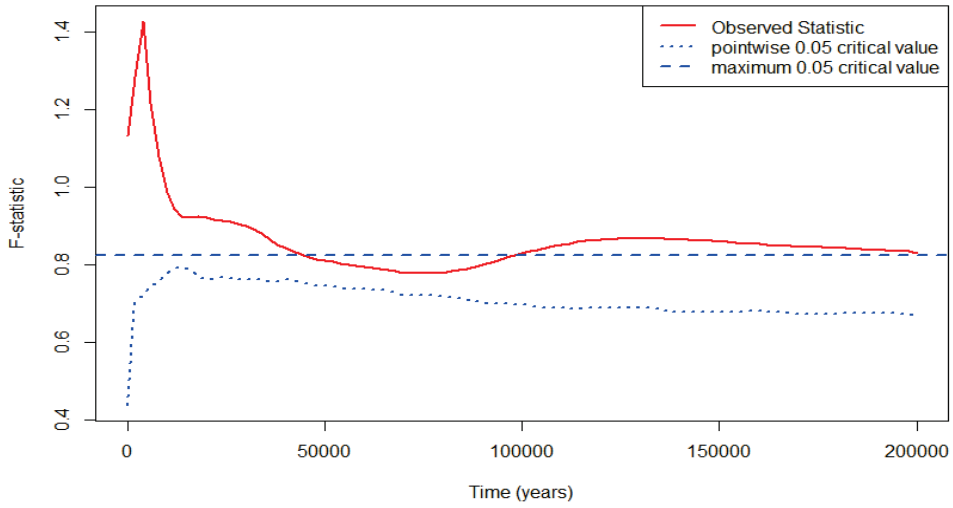


Figure 5.5: The permutation F-test (testing for no effect) for the no-intercept concurrent model under a constant population demographic model, at the 5% significance level.

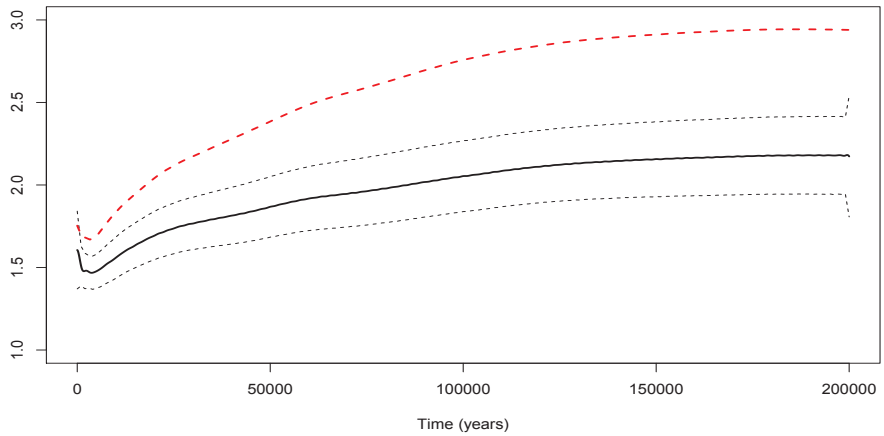


Figure 5.6: Regression coefficient function (black solid line) for the no-intercept concurrent model for DNA sequences simulated under the TN93 mutation model and a constant population demographic model, with 95% confidence intervals (black dashed lines). The true ratio is shown by the red dashed line.

Fitting this model to the data simulated under the constant demographic model gives the $\beta(t)$ shown in Figure 5.6. The regression coefficient varies smoothly with time and increases from 1.5 to around 2 as t increases. Also plotted is the ratio of the inferred population sizes from the random sample compared to the non-random sample (an empirical $\beta(t)$ if you will). It is encouraging that the no-intercept model model captures the shape of the variation in time of this ratio between the two samples. However, the magnitude of the empirical ratio is considerably larger. In fact, it does not lie in the error bands of the coefficient function. What is interesting in this plot is that the coefficient function is almost constant over time. There is some curvature in the point estimate of the function, but because the error bands are so wide a constant function would largely remain within the bands. This may indicate that the relationship between the two sets of curves is not far from just a scale factor. To see how the no-intercept model manages in terms of prediction, look at the same set of randomly selected simulations (from Figure 5.4) in Figure 5.7.

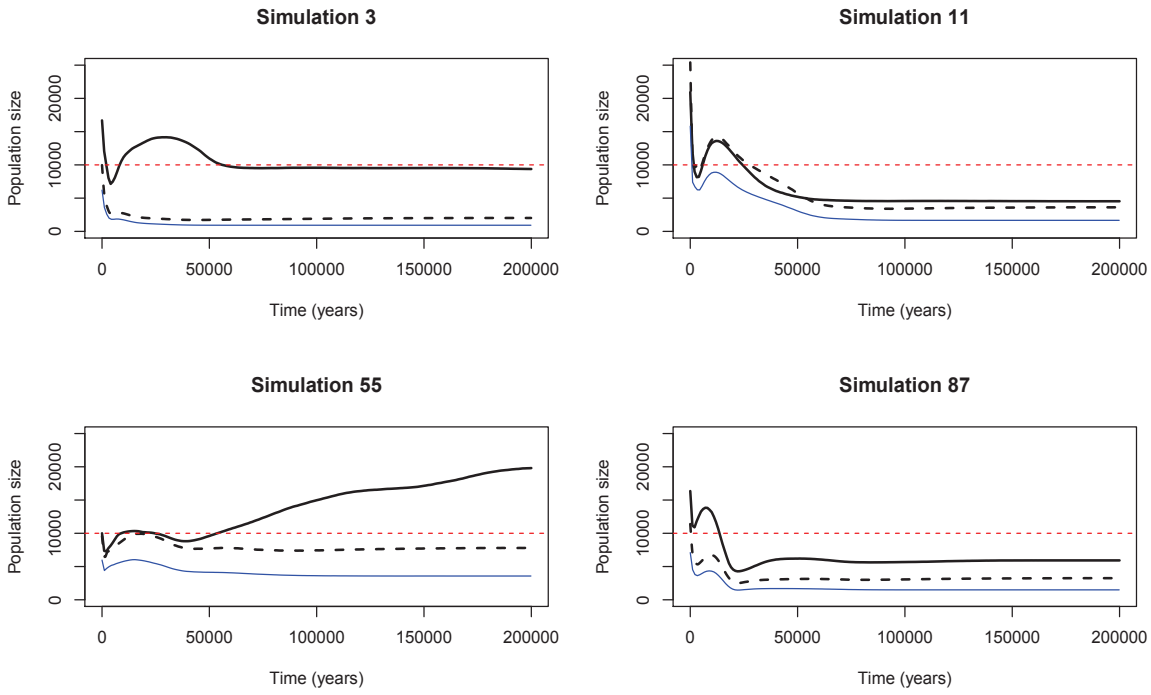


Figure 5.7: The same selection of simulations giving inferred and predicted population sizes as in Figure 5.4. As before, the inferred population size from the full sample is the black solid line, the predicted population size the black dashed line, the true population size the red dashed line and the blue line is the corresponding inferred population size from the subsample.

Note that the predicted population size does not capture either the true population size or the inferred population size from the random sample. In taking away the intercept term, there is little information around the true population size left in the model and so the predicted population size is completely driven by the subsample population curves, which we know from Chapter 4 are biased and on average underestimate the true population size over time.

Since removing the intercept term from the model and leaving only the covariate in the model had such a detrimental effect on prediction, look next at the concurrent model with *only* an intercept term, for completeness, where

$$Y_i(t) = \alpha(t) + \varepsilon_i(t).$$

Figure 5.8 shows the inferred intercept function, which represents the mean population size across curves. As shown in Chapter 4, under the Bayesian Skyline Plot model the inferred population size from a randomly sampled set of DNA sequences is approximately unbiased. This is what the concurrent functional intercept-only model is capturing. The predicted population size under the intercept-only model does as would be expected at predicting the true population size. This is useful in that it is further evidence that the Bayesian Skyline Plot model is unbiased.

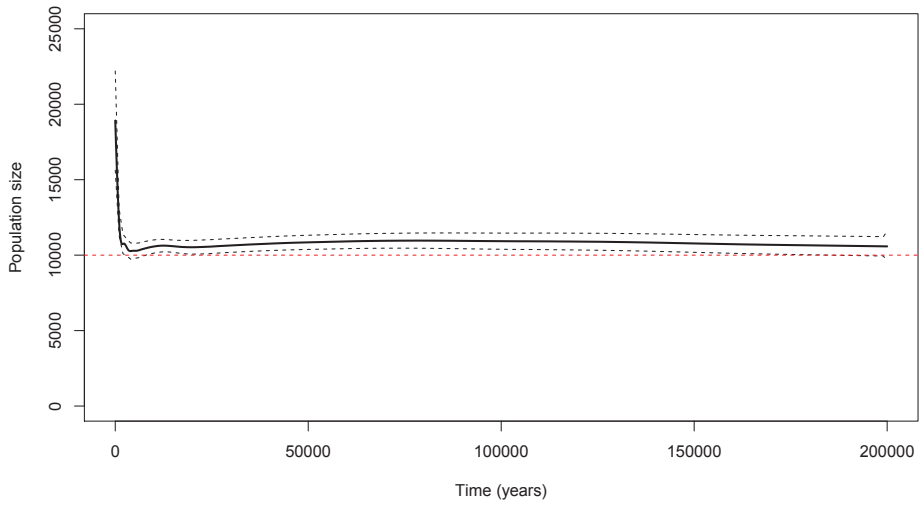


Figure 5.8: Intercept function for the intercept-only concurrent model under a constant population demographic model, with 95% pointwise confidence intervals. The red dotted line represents the true demographic model.

5.1.3 Discussion of functional data analysis and conclusions

The results of modelling the population curves from each sample under a functional data analysis framework are instructive. There is a significant relationship between the inferred population

size for a randomly sampled set of DNA sequences and a corresponding non-randomly sampled set, and the results of this model allow us to see how this relationship changes over time. Having both parameter functions $\alpha(t)$ and $\beta(t)$ in the model provides a very accurate prediction of the inferred population size that would be inferred from a random sample. As a consequence, this estimate is close to the true population size, since the inferred population size of the random sample case was shown to be unbiased in Chapter 4.

For the other three demographic models, the story is very similar (Appendix D). In the case of the step model (Section D.1), Figure D.2 shows the true population size is captured by the intercept function over time, with error bands being quite wide to start with, while the regression coefficient is more variable with no real pattern displayed. Zero is not contained within the 95% confidence intervals for the regression coefficient and so it is significant in the model. Removing the intercept function and comparing the estimated coefficient function with the empirical β over time (Figure D.6), the model captures the correct pattern with an increasing coefficient function over time with a dip at the change in population size but again this coefficient is underestimated. An interesting feature is that in each of the epochs the coefficient function is almost a straight line, indicating that the difference between the two sets of population curves may only be a scale factor. As in the constant case, the intercept-only model simply highlights that the Bayesian Skyline Plot model results are unbiased for the random sample (Figure D.8). In both the full model and the no-intercept model, the coefficient function is significant.

With the bottleneck population model (Section D.2), there is a similar situation. When modelling the two sets of population curves with the concurrent functional model with both an intercept and a coefficient function, the intercept function has more variability around the estimate in the present, and as we move back in time the error bands become narrower (Figure D.10). Meanwhile, the regression coefficient function changes through time, with a dip during the low population size of the bottleneck model. In fact, during this dip zero only just lies within the 95% confidence intervals for this parameter. Also, the further back in time we go, the larger the error bands around this estimate become. Removing the intercept function from the model (Figure D.14) leaves the coefficient function to capture the correct bottleneck shape, but as we have seen before in the two previous demographic models, this underestimates the true empirical β. However, as we move further back in the past, the regression coefficient increases to almost two and stays fairly constant there for the remainder of the third epoch in the bottleneck model. As we seen in the constant and step models, this regression coefficient function stays fairly linear in each epoch. This could again indicate that a simple scaling might be effective. Again, in both models the coefficient of the inferred population size of the non-randomly sampled DNA sequences is generally significant, apart from during the dip in population size.

Lastly, look at the exponential growth in population size demographic model (Section D.3). When fitting the concurrent model to this data, the intercept function, as before, captures the true demographic model fairly well (Figure D.18). The coefficient function never contains zero in the

95% confidence intervals and stays relatively constant through time. The error bands also stay fairly constant, with some slight widening at the start and end time points. In the intercept-only model for this data, as in all previous cases the coefficient function mirrors the shape of the true empirical β from the data, but the estimates do not overlap (Figure D.22). With a slight dip at around 50,000 years ago, the regression coefficient function lies at 1.5 for the exponential case, indicating that there could be a factor of 1.5 between the two sets of inferred population sizes. Again, both the full and no-intercept models show that the coefficient of the inferred population size of the non-randomly sampled DNA sequences is significant in both cases.

So, overall the concurrent functional model captures the true population size very well across all four demographic models. This modelling approach is very interesting since we can explore how the relationship between the inferred population size from the random sample and the inferred population size from the non-random sample changes over time. In most cases, there was more variability in the present, and this is reflected in the raw data curves, where for all four models we see quite high population estimates that then settle down to either the true population size or lower. In each of the four models the smoothing penalties λ_1 and λ_2 were allowed to differ, meaning that each functional model could be as flexible as required to capture the shapes in the curves.

However, the Achilles heel to this approach in this particular work is that it does not provide a usable estimate of the relationship between the inferred population sizes of the two samples. In other words, there is no way to simply *correct* the subsample population size using an equation like 5.1. In addition, the method is complicated and, given that the aim of this work to provide a general recommendation for inferring population sizes in reality from observed DNA sequence data, it seems impossible to translate these functions into a method one could use for interpretation of corrected population sizes. Furthermore, the parameters in this model are functions of time and would not generally be known so now, instead of considering the population curves, we must summarise them in some way to continue with our aim to correct the inferred population sizes.

5.2 Quantifying the relationship

Given that it has been established that there is a strong and significant relationship between the two sets of inferred population sizes, and understood how this relationship changes over time, the next step in this process is to quantify the relationship in an applicable manner. In doing so, one aims to come up with a simple transformation that would allow researchers to 'correct' their inferred population sizes of a set of non-randomly sampled DNA sequences to those which would be expected under data that did not violate the assumptions of the Coalescent Process. Given that the latter perform well (are approximately unbiased), this corrected population size trajectory would be also a good estimate of the population size.

This section will present the results of modelling the population sizes in this way, and then discuss ways in which the inferred population size from a non-random set of DNA sequences could be used to better learn about the true population size.

5.2.1 Regression of the mean inferred population sizes

A natural starting point is simple linear regression. To do this, summarise the inferred curves by taking the mean of the 100 posterior mean population sizes at each time point, represented by the vectors $MIPS_R$ and $MIPS_{NR}$, the mean inferred population sizes from the random and non-random samples. Figure 5.9 shows the relationship between these two variables under each of the four demographic models.

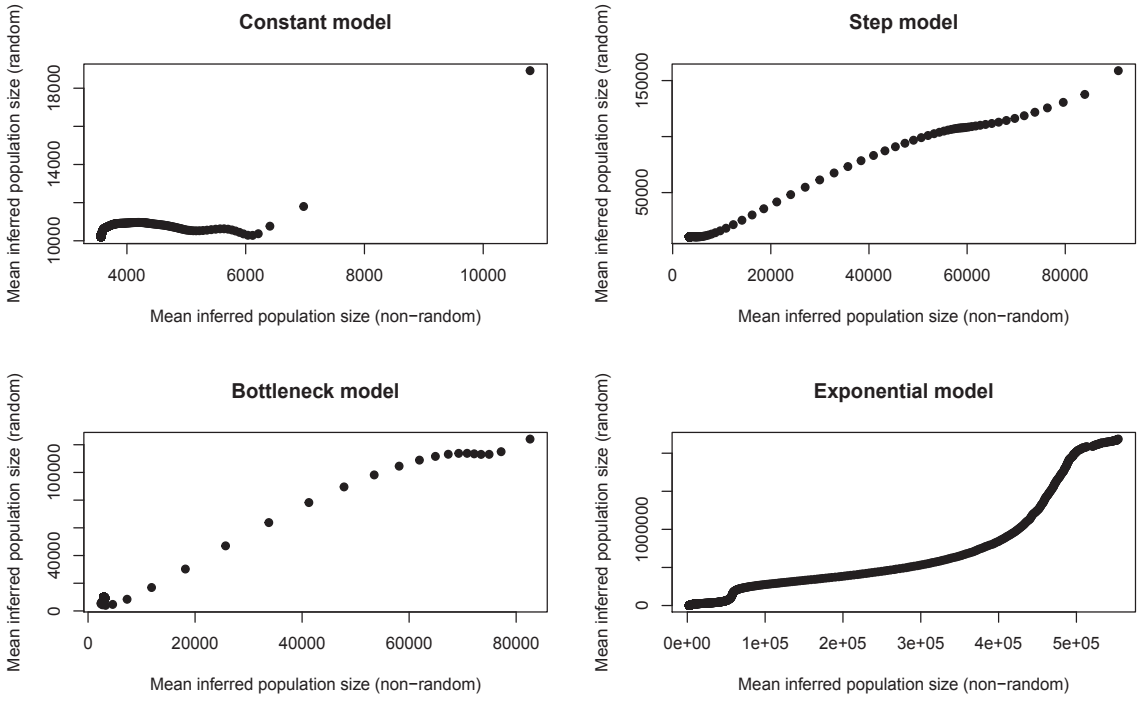


Figure 5.9: Scatterplots of $MIPS_R$ versus $MIPS_{NR}$ for each demographic model.

For the step and bottleneck models there is a positive and potentially linear trend between the two variables. In the case of the constant and exponential models, linearity is dubious, but in the exponential case there is evidence of a positive relationship between the variables. Despite the doubt around the linearity of the relationships, to allow a simple correction equation to be formed one could consider simple linear regression, and in this instance seemed a natural first step. Simple linear regression models the expected value of one random variable given the value of another as a linear function. A model such as this could describe the relationship

$$MIPS_R = A \times MIPS_{NR} \tag{5.7}$$

if a model with no intercept term was used. However, with the data involved this simple linear regression model would be an unsuitable choice because $MIPS_{NR}$ itself is not error free. Hence, Deming regression is more appropriate since it permits errors in both variables. This regression model differs from simple linear regression by measuring the error perpendicular to the line of best fit instead measuring the error only in one variable (the so-called response variable) (Deming, 1943). These errors are measured as the sum of the squared differences between the observed value and the mean scaled by the number of replicates, separately for each variable.

In Deming regression, each pair of 'measured' observations (x_i, y_i) are observations of the 'true' values (x_i^*, y_i^*) with independent errors ε_i and η_i

$$\begin{aligned} x_i &= x_i^* + \varepsilon_i \\ y_i &= y_i^* + \eta_i, \end{aligned}$$

for $i = 1, \ldots, n$ where n is the number of observations (which varies over demographic model) and x_i and y_i represent the i^{th} component of $MIPS_{NR}$ and $MIPS_R$, respectively. The model is

$$y^* = Ax^*, \tag{5.8}$$

with no intercept. We assume that the error variance ratio, defined by

$$\delta = \frac{Var(\varepsilon_i)}{Var(\eta_i)} = \frac{\sigma_\varepsilon^2}{\sigma_\eta^2},$$

is constant. This error ratio can either be pre-specified or estimated from the data by multiple measurements of subjects. In this work, this ratio is put equal to the ratio of variances of each variable. To estimate the slope parameter from the model, Deming regression uses the standard least-squares approach to minimise the weighted residual sum of squares,

$$SSR = \sum_{i=1}^{n} \left(\frac{\varepsilon_i^2}{\sigma_\varepsilon^2} + \frac{\eta_i^2}{\sigma_\eta^2} \right). \tag{5.9}$$

This gives the least-squares estimate of $\hat{A}$ to be

$$\hat{A} = \frac{s_{yy} - \delta s_{xx} + \sqrt{(s_{yy} - \delta s_{xx})^2 + 4\delta s_{xy}^2}}{2s_{xy}}, \tag{5.10}$$

where s_{xx} and s_{yy} are the corrected sums of squares for x and y, respectively, and s_{xy} is the corrected sum of squares of the product of x and y. Modelling the mean of the posterior means of the two sets of inferred population sizes for each of the four different demographic models gives the model output shown in Table 5.1.

In all four demographic models, the slope parameter is significantly different from zero. In

Table 5.1: No-intercept Deming regression for each demographic model

	$\hat{A}$	s.e.($\hat{A}$)	95/% confidence interval
Constant	2.561	0.030	$(2.503, 2.619)$
Step	2.208	0.039	$(2.132, 2.284)$
Bottleneck	2.842	0.067	$(2.710, 2.974)$
Exponential	1.901	0.006	$(1.889, 1.914)$

addition, all slope parameters and their confidence intervals are positive, indicating a positive relationship between the two sets of inferred population sizes. It should be noted here that these parameter estimates of the slope are very similar to those estimated from a standard simple linear regression model (results not shown). The estimates of the slope for the different demographic models are all rather similar to each other, although they do not overlap in terms of their confidence intervals.

Such a simple transformation is useful, but it is tempting to wonder whether we have access to any further information about the sample that might allow a more illuminated, but still practical, correction procedure.

5.2.2 A practical recommendation

For each of the 100 simulations we have

- the randomly sampled sets of DNA sequences each of size $n = 100$,
- the corresponding non-randomly sampled sets of DNA sequences each simulation of size n_s ($s = 1, \ldots, 100$),
- the inferred population sizes from the random samples (IPS_R),
- the inferred population sizes from the non-random samples (IPS_{NR}) and
- the true demographic model.

Compare this to what we would have if we were inferring population sizes from real sequence data, when we would have only

- a set of non-randomly sampled DNA sequences of some size and
- the inferred population size of that sample.

It seems at least plausible that the size of the subsample might affect the relationship between the two inferred population sizes. Table 5.2 shows the distribution of the subsample sizes across

Table 5.2: Subsample population sizes by demographic model.

Model	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
Constant	34.00	40.75	55.50	57.67	71.25	99.00
Step	34.00	42.00	50.00	56.21	69.25	99.00
Bottleneck	34.00	42.00	52.00	57.23	70.00	97.00
Exponential	34.00	40.00	49.00	55.10	65.00	97.00

demographic models. Over each of the four demographic models there is a range of subsample sizes, from 34 up to 99.

Given that the number of DNA sequences sampled will be a known quantity in any experiment, it would be interesting to investigate the relationship between the inferred population size and the subsample size. At the very least, one would expect that a small subsample would provide a much less accurate estimate of the population size compared to the full sample (of size 100). On the other hand, a subsample of around 90 or above would provide, one would expect, a very similar population estimate to the full sample case.

So, rather than model the mean of the posterior mean inferred population sizes, each simulation was modelled separately so that for each simulation there exists an individual slope parameter, $\mathbf{A} = [A_1, A_2, \ldots, A_{100}]$ and a corresponding subsample size $\mathbf{n_s} = [n_{s_1}, n_{s_2}, \ldots, n_{s_{100}}]$ of the 100 simulations. To be clear, A_1 denotes the slope parameter from the relationship between the posterior mean population size of the first randomly sampled set of DNA sequences and the posterior mean population size of the first non-randomly sampled set of DNA sequences, modelled by Deming regression and n_{s_1} the subsample size of the first simulation. Each demographic model was treated separately.

Let us suppose, in addition to the sampled sequences, that we have access to the prevalence of the subtree in the population (from another source, if necessary). Figure 5.10 explores the relationship of the slope $\hat{A}$ to the prevalence of the subtree in the simulations, i.e., n_s/n.

There is a clear non-linear relationship between the prevalence and the slope parameter in all four demographic models, specifically a decreasing quasi-exponential behaviour. In the step, bottleneck and exponential demographic models there is a clear indication that as the prevalence increases from around 0.3 the slope parameter decreases. In the constant model this is still the case, but with much more variability around the lower values of the prevalence. Therefore, to model this relationship, a non-linear regression was performed with the response variable being the slope parameter (estimated from Deming regression on the population sizes) and the prevalence as the explanatory. The model proposed is

$$A_i = \alpha\left(e^{\beta x_i} - e^{\beta}\right) + 1 + \varepsilon_i, \quad (i = 1, \ldots, 100), \tag{5.11}$$

where x_i is the prevalence in the i^{th} simulation and ε_i the error. This model is proposed as it can

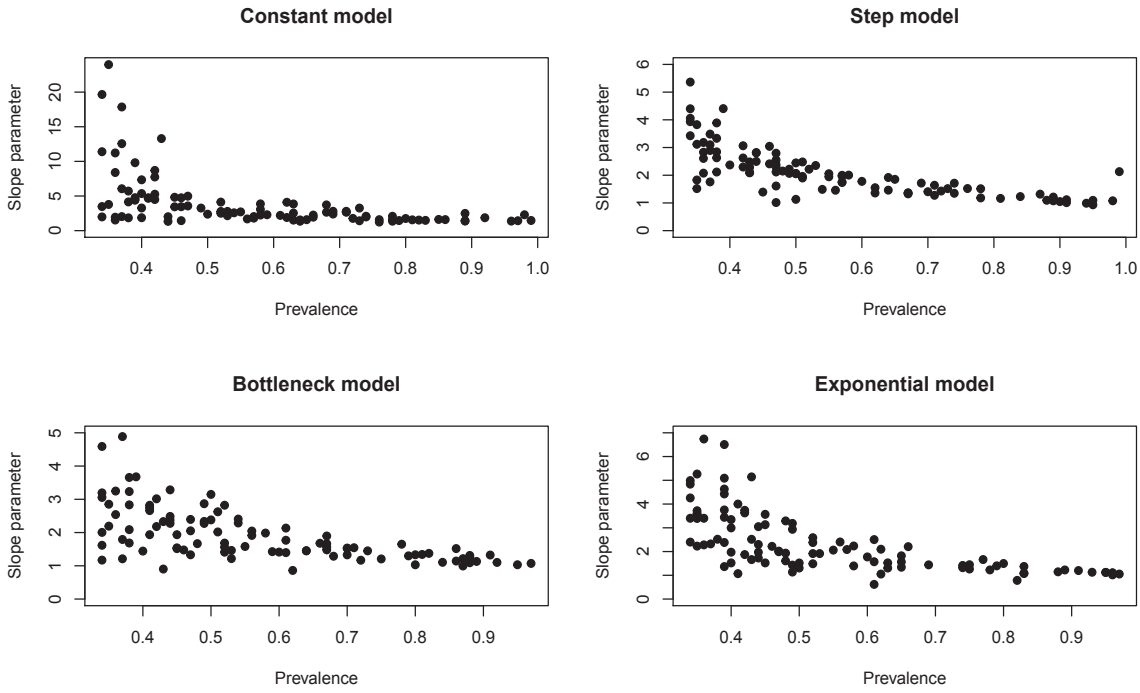


Figure 5.10: Scatterplots showing the relationship between the prevalence and the slope parameter across demographic models.

Table 5.3: Parameter estimates for the non-linear model describing the relationship between slope parameters and prevalences.

	α	s.e.(α)	β	s.e.(β)
Constant	101.979	69.663	-7.462	1.788
Step	10.938	2.757	-4.370	0.698
Bottleneck	5.013	1.150	-2.464	0.912
Exponential	19.787	7.798	-5.500	1.105

capture exponential decay (seen in Figure 5.10) for negative β and it forces the response to be 1 when the prevalence is 1 (as this is the case when we would be modelling a full sample). This resulted in the following model output shown in Table 5.3.

The parameter estimates are found by non-linear least squares using a Gauss-Newton algorithm, where the starting values for α and β are found by fitting a simple linear model between the logarithm of the slope parameter and the prevalence. Figure 5.11 shows the model fit.

These four estimates of α and β (Table 5.3) were then combined using a weighted mean (Woodward, 2013) to give an overall,

$$\hat{\alpha} = 15.193 \quad \text{and} \quad \hat{\beta} = -4.730, \tag{5.12}$$

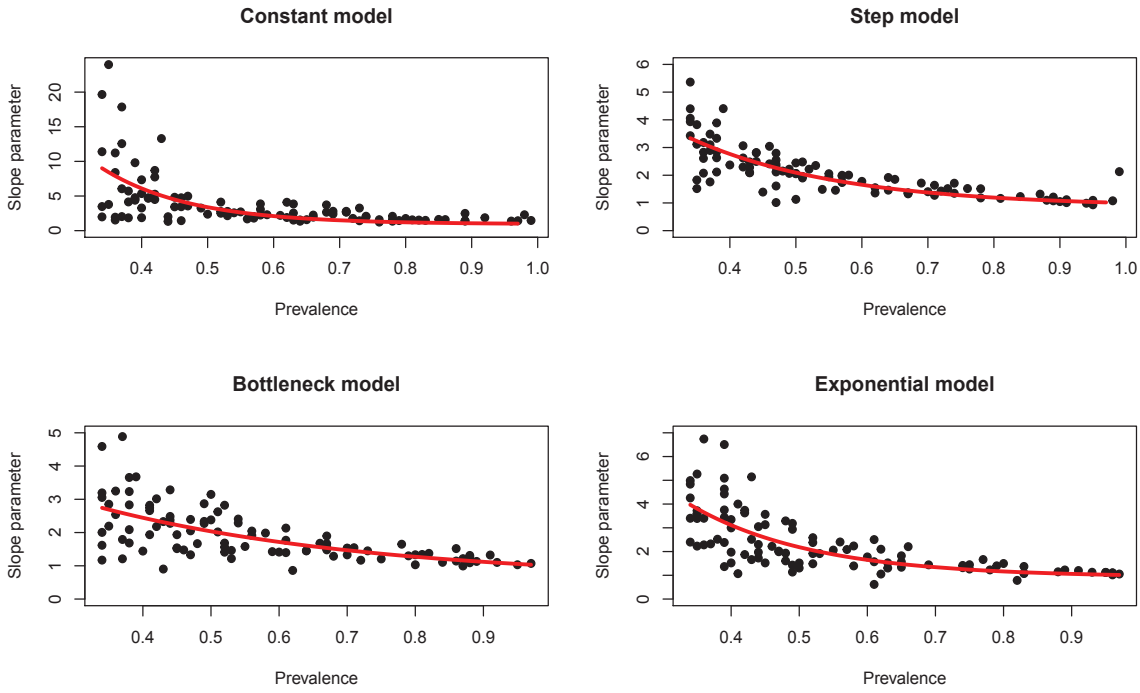


Figure 5.11: Non-linear models fitted to the slope parameter and prevalence across each demographic model.

with associated standard errors of s.e.$(\hat{\alpha}) = 0.705$ and s.e.$(\hat{\beta}) = 0.502$.

Thus, the final recommendation for correcting the inferred population size from a sample belonging to a subtree is the simple relationship,

$$IPS_R = A \times IPS_{NR}, \tag{5.13}$$

where the scalar multiplier A is given by,

$$A = \left[15.193\left(e^{-4.73x} - e^{-4.73}\right)\right] + 1, \tag{5.14}$$

when the subtree has prevalence x. Given that this correction equation was formed from results of full samples consistently of 100, of quite simplistic representations of demographic models and from sets of non-randomly sampled sequences all of size 34 or above, this recommendation should be taken with great caution. However, the approach could be extended and tested in more general scenarios. The following section applies this correction methodology to the Ingman data and to a more elaborate simulation.

5.3 Application to real data

Using this recommendation based on simulated data, the approach is illustrated on two more data sets. The first is the Ingman data set that has been used previously in this thesis to provide parameter values and the second is a more realistic simulated data set than those that have been used previously.

5.3.1 Ingman data set

Recall that the Ingman data set (Ingman et al., 2000) contains complete mtDNA sequences from 53 humans of diverse origins. This data set represents a *diverse* sample of humans, but by no means a random sample. We can use the reanalysis of this data set (Bandelt et al., 2006) to select a subsample based on a mutation at site 4,104, defining haplogroup L3. This gives a group of 38 sequences, excluding 15 African sequences.

Two analysis were carried out, the first was the same as that done in Chapter 4 for obtaining the parameter estimates for simulation by using the full Ingman data set of 53 sequences, and the second was carried out on this subsample of 38 sequences. In each sample, the prior distributions on the parameters were as described in Section 3.3 and the length of the MCMC chain was set to be 10 million in each case, thinned to every 1,000$^{\text{th}}$ iteration. This ensured convergence and good mixing of chains. The assumed mutation model was again chosen to be TN93. Figure 5.12 shows the resulting estimated population curves.

In Figure 5.12, the blue line represents the posterior mean inferred population size over time of the full data set (of 53 DNA sequences) and the corresponding 95% confidence intervals. The black line is the posterior mean inferred population size over time of the subdata set (of 38 DNA sequences) and the corresponding 95% confidence intervals. As we can see, the two population curves are quite similar in this case, with really the only difference being that the subsample population curve decreases to around 10,000 much faster than the full sample. However, both point estimates lie within the credible bands of one another and so it is hard to distinguish any real difference between the two curves. Applying the correction, with a prevalence of $38/53 = 0.717$, the red curve in Figure 5.12 is obtained.

The intervals around the corrected population size are approximated using the delta method (Rice, 2006). These intervals are very wide, so are able to capture very high initial population sizes closer to the present. The corrected population size inflates the initial population size but also prevents the population size from dipping too early, as happens with the non-random sample. This certainly looks a promising result.

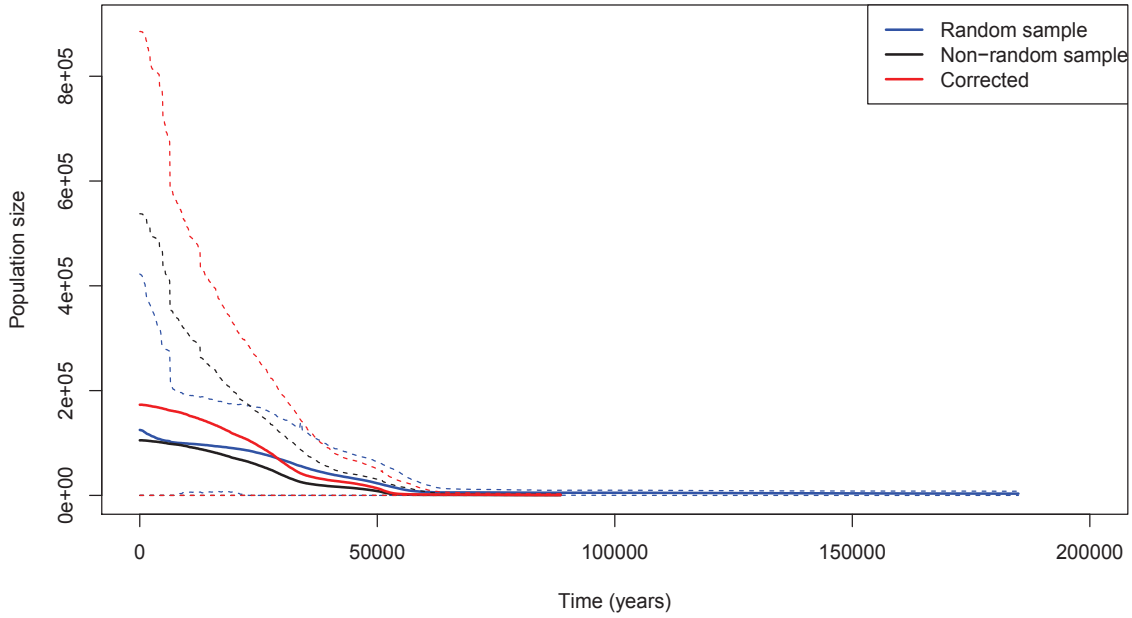


Figure 5.12: Inferred population sizes from the BSP model of DNA sequences from the full sample of the Ingman data set and the subsample belonging to haplogroup L3. The blue solid line is the posterior mean inferred population size of the full set (of 53 DNA sequences) and the corresponding 95% credible intervals are shown by the blue dashed lines. The black solid line is the posterior mean inferred population size of the subsample (of 38 sequences) and the corresponding 95% credible intervals shown by the black dashed lines. The red solid line and red dashed lines represent the corrected population size and credible intervals.

5.3.2 Variable population size

Now, a highly variable population size trajectory is simulated motivated by events through history and prehistory that could have impacted the population size. The aim here is to capture a more complex demographic model than the simplistic constant, step, bottleneck and exponential models that have been explored throughout this thesis. Figure 5.13 shows a cartoon of the proposed true population size (not to scale). This demographic model was selected to represent key historic events that would have influenced any significant change in population size, and the corresponding population sizes were selected to attempt to capture possible population expansions or contractions.

Using this demographic as the true population model, 100 trees were simulated using the Coalescent Process (Section 2.2.1) and DNA sequences evolved through the tree according to the TN93 mutation model. The population sizes were then estimated using the Bayesian Skyline Plot model. Then, this process was repeated another 100 times, this time selecting a subtree by a mutation of interest, thereby identifying a non-random sample of DNA sequences from which inference was made. The difference between this process of simulation and the simulations used

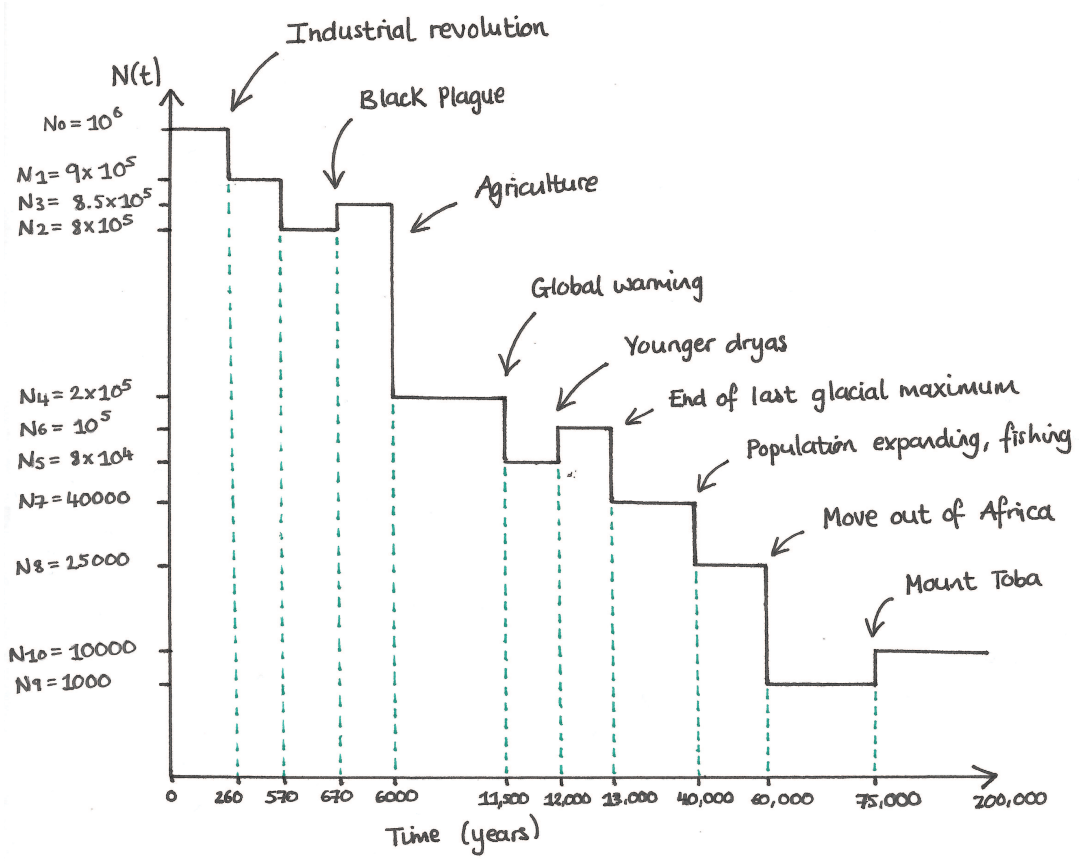


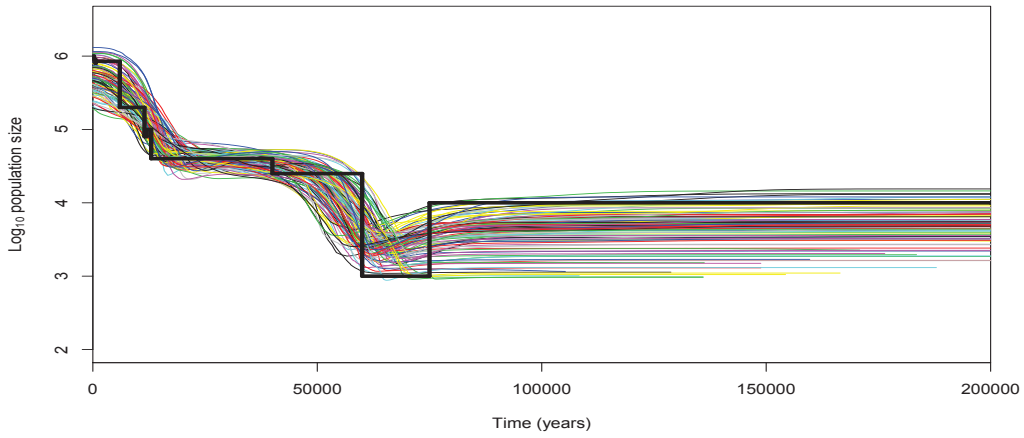
Figure 5.13: Cartoon of variable population size demographic model.

for modelling in previous sections is that the two sets of 100 simulations are not paired. The posterior mean population sizes from the random sample are presented first in Figure 5.14(a).

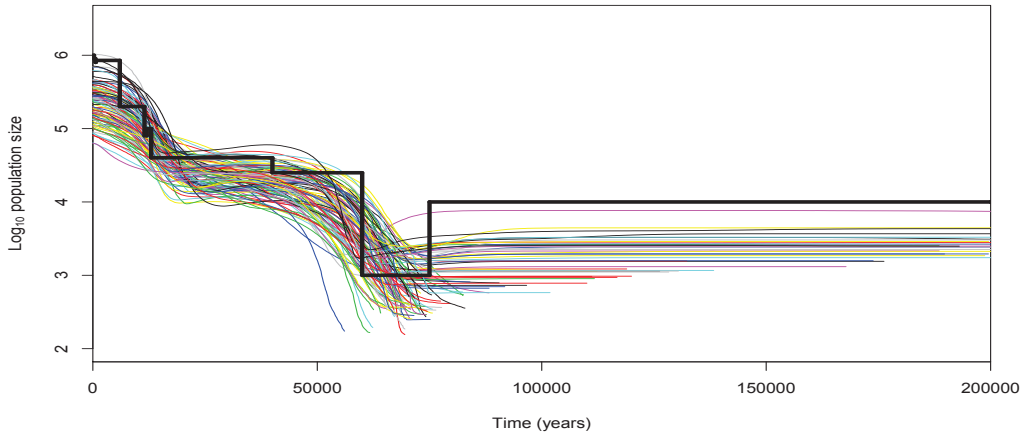
Again, the Bayesian Skyline Plot model very accurately inferred the true population demographic model from a sample of DNA sequences that do not violate the assumptions of the model. The initial population size is, on average, captured by the model, as is the initial drop in population size to around 6,000 years ago. The true population size is then relatively constant, as reflected by the inferred curves, before dipping to create a bottleneck 60,000 years ago. Some of the curves do not capture the bottleneck and stay constant from this point resulting in an underestimate of the population size in this final epoch, presumably because the simulated trees have reached their MRCA before this point. Figure 5.14(b) presents the inferred population sizes of 100 samples of DNA sequences that have been deliberately sampled non-randomly, as is often done in reality. The underlying true demographic model was the same for these samples.

The population curves in this figure tell a different story. As in previous examples, the BSP model underestimates the true population size. The initial population size is not captured, although the model has inferred the largest two steps in population size although not quite as accurately as the randomly sampled case. Many of the inferred population size trajectories do not reach beyond 100,000 years ago, a feature commonly seen in the simulation study of non-

(a)



(b)



(c)

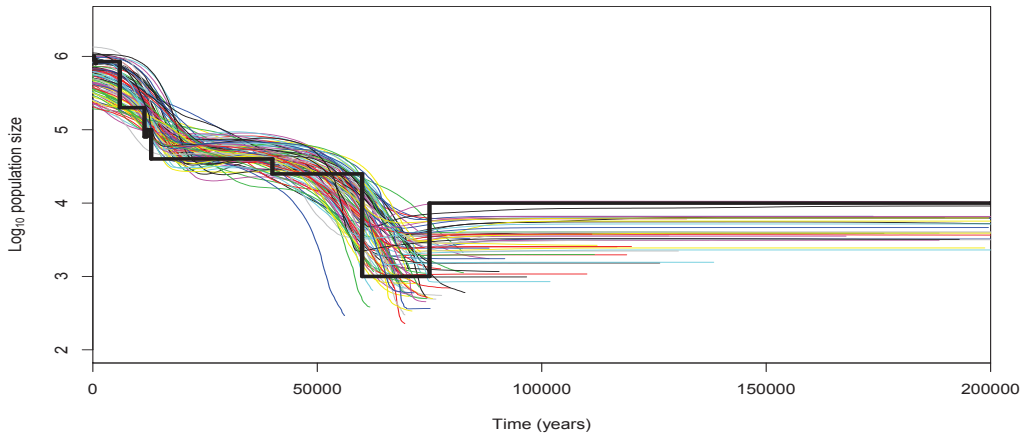


Figure 5.14: Inferred population sizes for 100 completely randomly sampled (a) and 100 non-randomly sampled (b) sets of DNA sequences simulated from a variable population size demographic model. (c) shows the corrected inferred population sizes using 5.13 and 5.14. In these figures we are looking at the logarithm of 100 posterior mean inferred population sizes over time represented by the coloured lines, and the true demographic model by the thick black line.

random samples of DNA sequences. This will be because the subsample will reach its MRCA before the full sample on average. In addition, the inferred population sizes in Figure 5.14(b) consistently underestimate the population size in the final epoch, more so than those in Figure 5.14(b). Overall, there is a clear difference between the inferred population sizes from the set of randomly sampled and the set of non-randomly sampled DNA sequences.

Using Equations 5.13 and 5.14, the inferred population sizes of the non-random sample were corrected and the results are presented in Figure 5.14(c). In general, these results prove to be very encouraging. Using the correction, and therefore the prevalence, to determine the scalar multiplier means that the corrected population sizes are much more similar to those of a randomly sampled set of DNA sequences. The initial population size is now captured by these curves, albeit in the tail of the distribution of these 100 simulations, and the true population size (represented by the black line again) lies comfortably within the distribution of the curves throughout the time period between the present and 60,000 years ago. As the bottleneck approaches many population size trajectories peter out (a feature discussed with the non-random sample) but the population size at this time is much more accurate. The population size in the last epoch is again underestimated but not as much as in Figure 5.14(b).

No measure of error is provided in this application simply because it would be impossible to display in one figure. However, as in the application to the Ingman data set (Section 5.3.1), in each of the 100 simulations the variance around the corrected population size ($\widehat{IPS}_R$) was calculated using the Delta method and an expression for this is:

$$Var(\widehat{IPS}_R) = \widehat{IPS}_{NR}^2 \, Var(\hat{A}) + \hat{A}^2 \, Var(\widehat{IPS}_{NR}), \tag{5.15}$$

where $\widehat{IPS}_{NR}$ is the inferred population size of the sample from the Bayesian Skyline Plot model (usually the posterior mean) and $Var(\widehat{IPS}_{NR})$ is calculated from the standard deviation across the iterations of the BSP. Then $\hat{A}$ is calculated from Equation 5.14 and the variance of $\hat{A}$ is found using the Delta method giving

$$Var(\hat{A}) = 0.497\left(e^{\beta x} - e^{\beta}\right) - 0.39\left(e^{\beta x} - e^{\beta}\right)\left(\alpha\, e^{\beta x} - \alpha e^{\beta}\right) + 0.252\left(\alpha x e^{\beta x} - \alpha e^{\beta}\right), \tag{5.16}$$

The error bands for each of these 100 simulations (not shown) were very wide, capturing both the true population size and almost all the inferred population sizes from the random samples (from Figure 5.14(a)).

5.4 Discussion

The aim of this chapter was to explore the relationship between the inferred population sizes from the randomly sampled set of DNA sequences (IPS_R) and the inferred population sizes from the non-randomly sampled set of DNA sequences (IPS_{NR}) and to model that relationship

in a way that could be used to correct population sizes when samples were taken from a subtree. This can happen if the samples share a mutation which is correlated with a phenotype on the basis of which the samples are selected. But it also occurs when a haplogroup is analysed out of the context of the rest of the sequences in the population.

Using the Functional Data Analysis methods presented in Section 5.1, the relationship between the two sets of population curves was shown to be statistically significant, and that relationship was not highly variable over time across all four demographic models. In the constant and exponential models, when removing the intercept from the concurrent functional model the relationship between IPS_R and IPS_{NR} stayed almost consistently linear over time. This implied that the relationship between the two variables could well be explained with a scalar multiplier. In the case of the step and the bottleneck models, this relationship was not as clear because of the sudden changes in population sizes, but within each epoch the relationship was shown to be somewhat linear.

Treating the inferred population sizes in a regression setting, as presented in Section 5.2.1, provided a simple linear model between the mean of the inferred population size of the non-random sample ($MIPS_{NR}$) and the mean of the inferred population size of the random sample ($MIPS_R$) for each of the four demographic models, with the constant case being the least convincing.

A method for estimating the slope of the linear relationship was then derived. This included using the prevalence of the mutation of interest to learn about the slope in a regression model of the inferred population sizes. The relationship between the prevalence and the slope parameter was approximately an exponential decay; such a model was a good fit.

One set of models that could have been appropriate for the task are the class of flexible regression models. These models, similar to the FDA models, provide flexibility in the relationship being modelled. In simple terms, a flexible regression model fits a function to the data without a strict parametric assumption and as a result there is no fixed number of unknown parameters to estimate. In flexible regression one usually uses smoothing splines to control the smoothness of the relationship if the data is not sufficiently linear. This type of modelling, although beneficial because it assumes no predetermined relationship between the variables, is not suitable for the purpose of modelling in this work. This is due to the fact that we are seeking a relatively simple and general relationship between two known variables that allows correction of population curves through time in a wide variety of contexts. Having to explain the relationship in terms of many basis functions takes away from this.

My proposal provides a way for researchers to acknowledge the violations of the assumptions of the models, but more importantly to obtain a more convincing estimate of population size. This goes some way to avoiding misinterpretation of population sizes inferred from the Bayesian Skyline Plot model. Using this correction could then mean that any interpretations of population size could be made a little more definitively than is currently possible.

Chapter 6

General Discussion

The work in this thesis aimed to investigate the accuracy of the inferred population size produced from the Bayesian Skyline Plot model and to determine whether or not any bias existed in the population size estimates when the DNA sequences formed a non-random sample, specifically from a subtree, violating the model assumptions. As discussed in Chapter 1, models for inferring population sizes from DNA sequence data are all based on some unrealistic assumptions that any real data set will undoubtedly violate due to the nature of data collection. In particular, the Bayesian Skyline Plot model is a popular model for inferring such population sizes and is used widely in the field of population genetics, for humans and other organisms. However, to the best of our knowledge, there has been little research exploring the impact of violated model assumptions (such as the sampling scheme) on its output and what this means for the inferred population sizes that result.

With a simulation approach to compare a random and non-random sample formed from a subtree, this effect of non-random sampling could be explored and was quantified. A recommendation is made to researchers for interpreting the inferred population size (as a function of time) from non-random samples of DNA sequences. This work makes a start at addressing the problem in this sampling context by suggesting a post-hoc correction that could be applied to the inferred population sizes if the frequency of the subpopulation sampled in the subsample is known.

6.1 Conclusions

Carrying out this work on simulated DNA sequence data permitted us to control the analysis in ways that are impossible with real data. Based on real DNA sequences we aim to estimate population sizes far back in time beyond where we have historic records, using a mathematical description of how population size changes affect the ancestry of a sample. As discussed in Chapters 1 and 2, the Coalescent Process makes assumptions about the sample of data being analysed and the population from which those samples were drawn. These assumptions are

unrealistic for human populations and so no sample of real data can satisfy them completely. As such, using simulated DNA sequences allows us to ensure that all assumptions are met and thus all inferences drawn from the models can reflect them. More importantly, using simulated data allows us to compare the inferred population sizes from the model to the true population size. As shown in Chapter 4, these simulated DNA sequences have similar properties to real sequences, for example in respect of measures of genetic variation.

Furthermore, the work presented in Chapter 4 showed that with random samples of DNA sequences, the historic population size estimated from the Bayesian Skyline Plot model were unbiased and on average very accurate in capturing the true demographic model, whether a constant population size, a step in population size, a population bottleneck or an exponential growth in population size. The results here are consistent with another study that investigated the bias of the Bayesian Skyline Plot model using simulated data but where the focus was on migration rather than the sampling scheme (Heller et al., 2013).

To reiterate, in order to assess the consequences of analysing a non-randomly sampled set of DNA sequences again, simulation was used so that the true underlying population size was known. The inferred population size estimated from the Bayesian Skyline Plot model was compared to this and to the inferred population size of the randomly sampled set. The non-random sample was a subsample of the random DNA sequences, chosen by certain criteria outlined in Chapter 2, whereby the resulting subsample was expected to contain a shared mutation, reflecting samples analysed in reality from a certain haplogroup.

As discussed in Chapters 4 and 5, the resulting population curves for the subsample were consistently underestimated by the Bayesian Skyline Plot model. However, the model did capture the correct shape (i.e., a constant, step or curve). In the step and bottleneck models the magnitude of the population size was misinterpreted but the changes observed in population sizes corresponded to the times of changes in the true demographic model. Modelling the two sets of population curves together provided some important insights. Since the inferences were essentially functions, Functional Data Analysis was used to explore the relationship between the two sets of curves. A significant relationship was found between the two inferred population sizes. However, the parameters estimated from this model are functions of time. These would not in general be known in the case of treating a single actual data set. So, although the analysis demonstrated the two samples have signals in common, it did not provide a practically useful method of correcting the inference from the non-random sample.

As such, summarising the curves by the mean of 100 posterior mean inferred population sizes (for each sample) gave us point estimates of inferred population sizes and modelling these provided the starting point in creating some form of post-hoc correction. To learn about the relationship between the two curves, the prevalence of the subtree in the population was used, which is usually known (Section 5.2.2). It shed some light on how accurate the inferred population size of the non-random sample will be. The posterior mean inferred population size

from a subsample was much closer to the inferred population size of the corresponding random sample the higher the prevalence. The next section presents the proposed recommendation to researchers inferring population sizes based on these results.

6.2 A proposed correction

The zeroth-order recommendation from this work is that caution should be taken when inferring historic population sizes from a non-randomly sampled set of DNA sequences. It has been shown that less information is gained with such a subsample of DNA sequences, which is not surprising as this is the case in any data analysis, not limited to genetic data.

By its nature, a subsample formed from a subtree will provide a less informative population size estimate (in terms of how far back time we can estimate this population size) compared to the corresponding coalescent tree. This is because the subtree will reach its most recent common ancestor sooner than the full tree will. However, we are not interested so much in how far back in time we can infer population sizes, but rather the shape of the inferred population sizes over time, i.e., did the subsample simulated under the bottleneck model capture the bottleneck? The results in Chapters 4 and 5 showed that these curves did and, by modelling the sets of inferred population sizes from the random and non-random samples together, it was clear that there exists a strong relationship between the two.

Throughout the analysis of population sizes it emerged that just a scalar multiplier could potentially fix misinterpreted population sizes from the non-random sample. This was clear from modelling the mean of the posterior mean population sizes in Chapter 4 and from modelling the data using Functional Data Analysis (Section 5.1). Therefore, a suggested correction for this error generated by the sampling scheme is of the form

$$IPS_R = A \times IPS_{NR}, \tag{6.1}$$

where IPS_R and IPS_{NR} are the inferred population sizes of the randomly sampled and non-randomly sampled sets of DNA sequences, respectively. The slope, A, can be found using the prevalence in the population, denoted x, by

$$A = \alpha\left(e^{\beta x} - e^{\beta}\right) + 1, \tag{6.2}$$

for $\alpha = 15.193$ (s.e. 0.705) and $\beta = -4.730$ (s.e. 0.502) and $0 \leq x \leq 1$. The values for α and β were estimated from a non-linear model describing the relationship between the prevalence x and the slope parameter A between the two sets of posterior mean population sizes for each simulation. The results from each of the four demographic models were combined to generate the values of α and β that would not be unduly influenced by one underlying demographic model.

It has been shown that Equation 6.1 does well when used to correct population size from the results of a real data application and some simulated data from a population with a highly variable population size (motivated by events in human prehistory). However, it is based on a restricted set of simulations and should be used with caution. More important perhaps than the correction equation (6.1) itself, this work proves that, in the instance of a non-random sample, a biased population estimate will occur but it usually captures the shape of the population size function accurately, just not its magnitude. This is most likely why it was possible to determine a simple mathematical relationship between the sets of curves, because a scalar multiplier was largely sufficient.

The work of Rito et al. (2013) and Soares et al. (2011) was presented at the start of this thesis as examples of non-random samples analysed in practice and one wonders what would happen to their results if they applied either this correction or something similar. Given that both acknowledge a signal from one haplogroup and no signal from a subhaplogroup, I would predict that applying a correction to the inferred population size would provide a fuller picture of the population size. However, even then it would be difficult to judge that picture since we have no known truth to compare to, unlike in the simulation study.

6.3 Limitations and prospects

This work explored how our inference of population size functions from the Bayesian Skyline Plot changed depending on the true demographic model or the mutation model from which DNA sequences were simulated. It would be interesting to see whether these results, in particular the correction (6.1), changed if we were looking at, say, Y-chromosomal data, which has a lower rate of mutation (Helgason et al., 2015).

In addition, the sample size of the random sample was kept constant at $n = 100$ over all simulations. There is not much evidence (Jobling et al., 2014) that increasing the sample size would affect properties of the trees (e.g., T_{MRCA}) or genetic diversity (the mean pairwise difference), but making the sample size smaller would be interesting. In this work, it had to be kept large enough to ensure an adequate subsample size. As well, the scheme for selecting the subtree was determined by a branch containing a mutation in expectation (Section 2.5.2). Rather than select the subtree this way, it could be selected in the same manner as in Section 5.3.1 where a subtree of the Ingman data set was identified using a specific observed mutation.

The work done by Heller et al. (2013) in investigating the inferred population size from the Bayesian Skyline Plot model under the violation of assumptions of panmixia could be extended in the same style as this work, i.e., to explore whether or not a correction could be applied to said population sizes. The work carried out in that paper is similar in approach to that presented here in Chapter 4, and the Bayesian Skyline Plot model misinterprets the population size when analysing data that violates assumptions of the sampling scheme or the population structure.

Throughout this thesis I have suggested that we would prefer a random sample of DNA sequences and shown that indeed a random sample of DNA sequences provides more accurate results of the true population size function. However, as has been discussed before, a random sample will never be feasible for any study. Non-random sampling in the form of selecting a subtree is commonly done and we understand this form of non-randomness. So, in place of applying a post-hoc correction, which would be better principled, another option would be to put a mechanism that describes the non-random sampling into the model itself. In the case of a non-random subtree this would require the distribution of a subtree of a coalescent tree. In principle, this could be found by marginalising out the rest of the (full) tree, but this is challenging. Even if theoretical results are not easily obtained, simulation-based approximations might be possible. Limited work has been done in this general area (Slatkin and Rannala, 1997). However, to do so requires further model assumptions to be made and threatens the danger of creating a bigger problem than we already have, if those further assumptions are unrealistic.

To conclude, the Bayesian Skyline Plot model does remarkably well at inferring historic population sizes from mtDNA, with a sample that does not violate model assumptions. Researchers tend to ignore the fact that their inferences may be misleading due to their non-random sample violating these model assumptions (Macaulay et al., 2005; Rito et al., 2013; Soares et al., 2011). As this thesis shows, caution should most definitely be taken when making any strong claims about population sizes from a non-random sample, but, in general, if *changes* in population size rather than the *magnitude* of population size is the aspect on which interpretation will be built, the Bayesian Skyline Plot can be a very powerful tool.

Appendix A

List of Main Symbols

α	Mutation rate parameter of transitions.
α_1	Transition rate parameter for purines.
α_2	Transition rate parameter for pyrimidines.
α_G	Shape parameter for the Gamma distribution of Gamma distributed mutation rates across sites in the DNA sequence. This is discussed in Section 2.4.1.
β	Mutation rate parameter of transversions.
β_G	Scale parameter for the Gamma distribution of Gamma distributed mutation rates across sites in the DNA sequence. This is discussed in Section 2.4.1.
C	Number of categories for approximating the continuous Gamma distribution by a discrete Gamma distribution for the mutation rates across sites. This is discussed in Section 4.1.2.
ε	Minimum length for a branch on which the mutation of interest could lie. This is discussed in Section 2.5.2.
η	Proportion of invariant sites in the DNA sequences, as described in Section 2.5.3.
g	Number of years per generation. In this thesis, this value is 30.
IPS_{NR}	Inferred population size of a non-random sample using the Bayesian Skyline Plot model.
IPS_R	Inferred population size of a random sample using the Bayesian Skyline Plot model.
k	Number of lineages in the genealogy before a coalescent event (from n down to 2).
κ	Transition-transversion ratio. This is defined in Sections 2.3.2 and 2.3.3.
κ_1	Pyrimidine transition-transversion ratio in the TN93 mutation model. This is defined in Section 2.3.3.
κ_2	Purine transition-transversion ratio in the TN93 mutation model. This is defined in Section 2.3.3.
l	Length of the DNA sequence. In this thesis, this value is taken to be 15,446 sites.

L	Length of the tree.
λ_k	Rate of coalescence when there are k lineages.
$MIPS_{NR}$	Mean inferred population size from the Bayesian Skyline Plot over 100 simulations of non-random samples.
$MIPS_R$	Mean inferred population size from the Bayesian Skyline Plot over 100 simulations of random samples.
μ	Average nucleotide mutation rate per site. This value is taken as 1.26×10^{-8} per year.
μ_s	Mutation rate of site s.
N	Haploid population size.
N_0	Present day population size.
n	Sample size of DNA sequences that have been randomly sampled (the full sample).
n_s	Subsample size of DNA sequences that have been non-randomly sampled.
$\boldsymbol{\pi}$	Vector containing the four nucleotide equilibrium frequencies $(\pi_T, \pi_C, \pi_A, \pi_G)$ for nucleotides T, C, A and G, respectively.
π	Mean pairwise difference of a sample of DNA sequences. This is discussed in Section 4.2.1.
t	Time (in years, unless specified otherwise).
T_{MRCA}	Time until the tree reaches the most recent common ancestor, or the depth of the tree.
W_k	Random waiting times between coalescent events (when there are k lineages).

Appendix B

Glossary of Some Genetic Terminology

Adenine	A compound which is one of the four constituent bases of nucleic acids. A purine derivative, it is paired with thymine in double-stranded DNA.
Allele	An alteration in DNA sequence so that it is a variant form of a given gene.
Cell	A tiny compartment contained by membrane, they are the smallest things that can reproduce themselves. Every living thing is made up of a cell or cells.
Chromosome	A threadlike structure of nucleic acids and protein found in the nucleus of most living cells, carrying genetic information in the form of genes.
Cytosine	A compound which is one of the four constituent bases of nucleic acids. It is a pyrimidine derivative and is paired with guanine in double-stranded DNA.
Deletion	A type of mutation and happens when part of the genome is lost during the replication process.
DNA	Deoxyribonucleic acid, double long chains of nucleotides that includes the ciphertext for protein manufacture.
Gene	A substring of the genome with a particular function, often to code for one (or part of one) protein.
Genealogy	A tree of descent traced from an ancestor (ancestral tree).
Genome	The complete set of genetic material, contained in a cell.
Guanine	A compound which is one of the four constituent bases of nucleic acids. It is a pyrimidine derivative and is paired with cytosine in double-stranded DNA.
Haplogroup	A group of descendants that are distinguished by a mutation so create a subtree.
Insertion	A type of mutation consisting of the addition of one or more nucleotide base pairs into the DNA sequence during the replication process.
Meiosis	One cell divides twice creating four haploid daughter cells containing half of the original genetic information. These are sex cells (gametes) and are eggs in females and sperm in males.

Mitochondrion	The powerhouse of the cell, providing a source of energy to the cell. Only passed on through the female line.
Mutation	An alteration of the DNA sequence.
Nucleotide	A type of molecule and another term to describe the bases A, G, C and T.
Purine	Nitrogenous bases that make up the one of the two different kinds of nucleotide bases in DNA. Adenine and guanine are purines.
Pyrimidine	Nitrogenous bases that make up the one of the two different kinds of nucleotide bases in DNA. Thymine and cytosine are pyrimidines.
Recombination	The rearrangement of genetic material, by crossing over between part of chromosomes during meiosis.
Substitution	A type of mutation in which one nucleotide is replaced by one other.
Thymine	A compound which is one of the four constituent bases of nucleic acids. A purine derivative, it is paired with adenine in double-stranded DNA.
Transition	Interchanges of purines (A & G) or of pyrimidines (C & T).
Transversion	Interchanges between purine and pyrimidine bases.

Appendix C

Properties of the Quasi-Random Sample

Continuing on from Section 4.2.3, this Appendix compares the properties of trees simulated under truly random conditions to those simulated quasi-randomly. Genetic variation in the two samples are compared as well as the inferred population sizes from the Bayesian Skyline Plot model. Results from the step, bottleneck and exponential demographic models are presented. In this Appendix the term *population curve* refers to the mean of 100 posterior mean population sizes for the truly randomly sampled and the quasi-randomly sampled DNA sequences estimated from the Bayesian Skyline Plot model.

C.1 Step demographic model

Beginning with the step demographic model, Figure C.1 shows boxplots comparing tree properties between the truly random and quasi-random samples and the mean pairwise differences between the samples to compare the genetic variation. From Figure C.1 there is only a very small difference between the two samples. The quasi-random sample has a slightly smaller tree depth, tree length and mean pairwise difference than what we see with the truly random sample. However, none of these difference are significant with a two-sample t test at 5% significance level.

Figure C.2 shows the inferred population size from the truly random and quasi-random samples. The inferred population sizes from each of the samples are very similar. Both sets of error bands overlap almost always and the mean inferred population sizes are consistent with one another through time. The quasi-random sample begins with a slight overestimation, then around 30,000 years ago shows an over estimation, but not drastically so. Overall, there is no statistically significant difference between the two samples and so the quasi-random sample for data simulated under the step demographic model will be considered random.

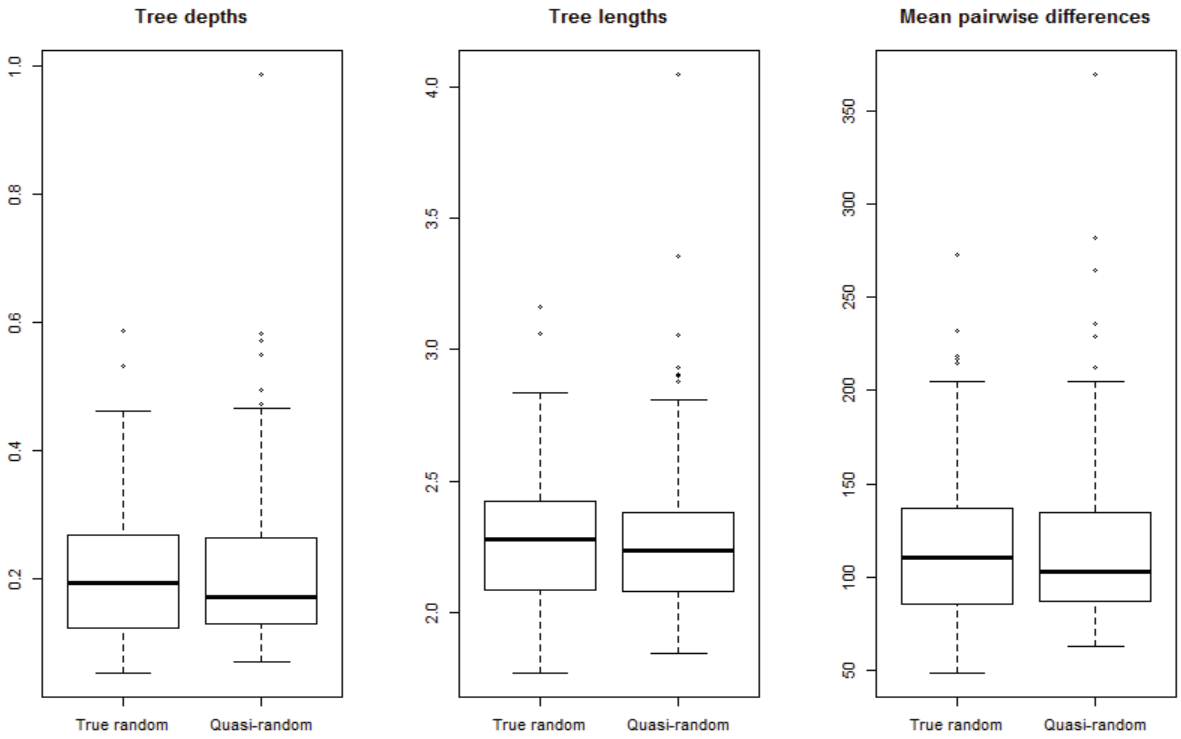


Figure C.1: Boxplots comparing the tree properties of the truly random and quasi-random simulated samples. Each of these were simulated under a step demographic model. Time is measured in coalescent units.

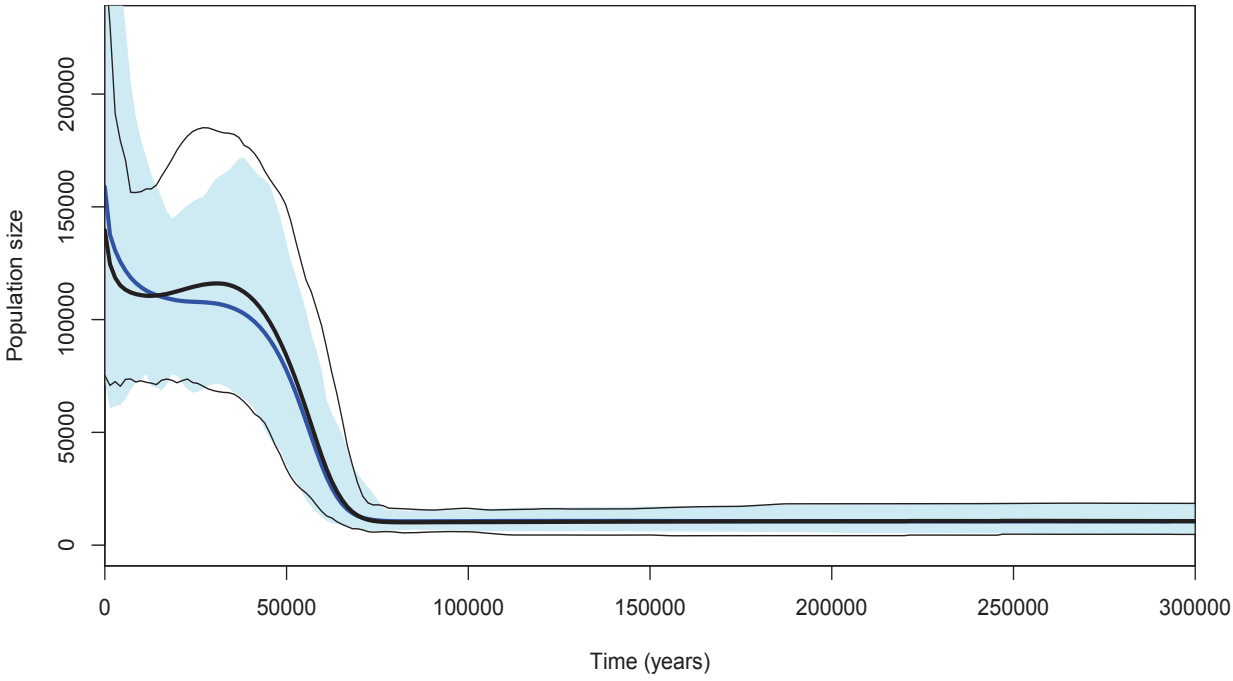


Figure C.2: Population curves from the Bayesian Skyline Plot model for DNA simulated under the step demographic model. The plot shows the truly random sample's population curve (black solid line) and the quasi-random sample's population curve (solid blue line). The pale blue shaded area represents the 2.5^{th} and 97.5^{th} percentiles of the posterior means for the quasi-random samples, and the thin black lines the same but for the truly random samples.

C.2 Bottleneck demographic model

Next, we turn attention to the bottleneck demographic model. Figure C.3 shows three boxplots comparing the tree properties and the genetic variation between the truly random and quasi-random samples simulated in this case.

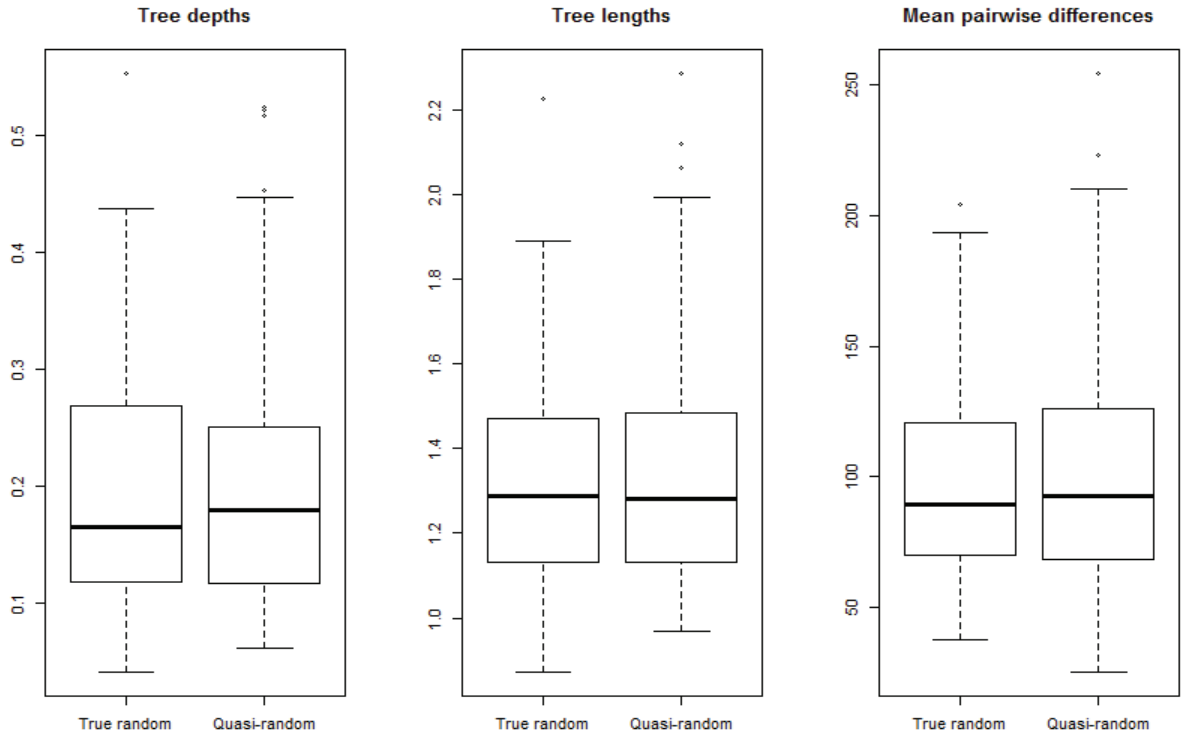


Figure C.3: Boxplots comparing the tree properties of the truly random and quasi-random simulated samples. Each of these were simulated under a bottleneck demographic model. Time is measured in coalescent units.

From Figure C.3 there is only a very small difference between the two samples. The quasi-random sample has a slightly bigger tree depth but the tree lengths and mean pairwise differences are almost equal. None of these difference are significant with a two-sample t test at 5% significance level. Figure C.4 shows the inferred population size from the truly random and quasi-random samples. As before, the two sample's error bands overlap consistently and, again, the mean inferred population size for each sample is almost the same with the quasi-random sample estimating it slightly lower at around 30,000 years ago. Overall, as before, there is no statistically significant difference between the two samples.

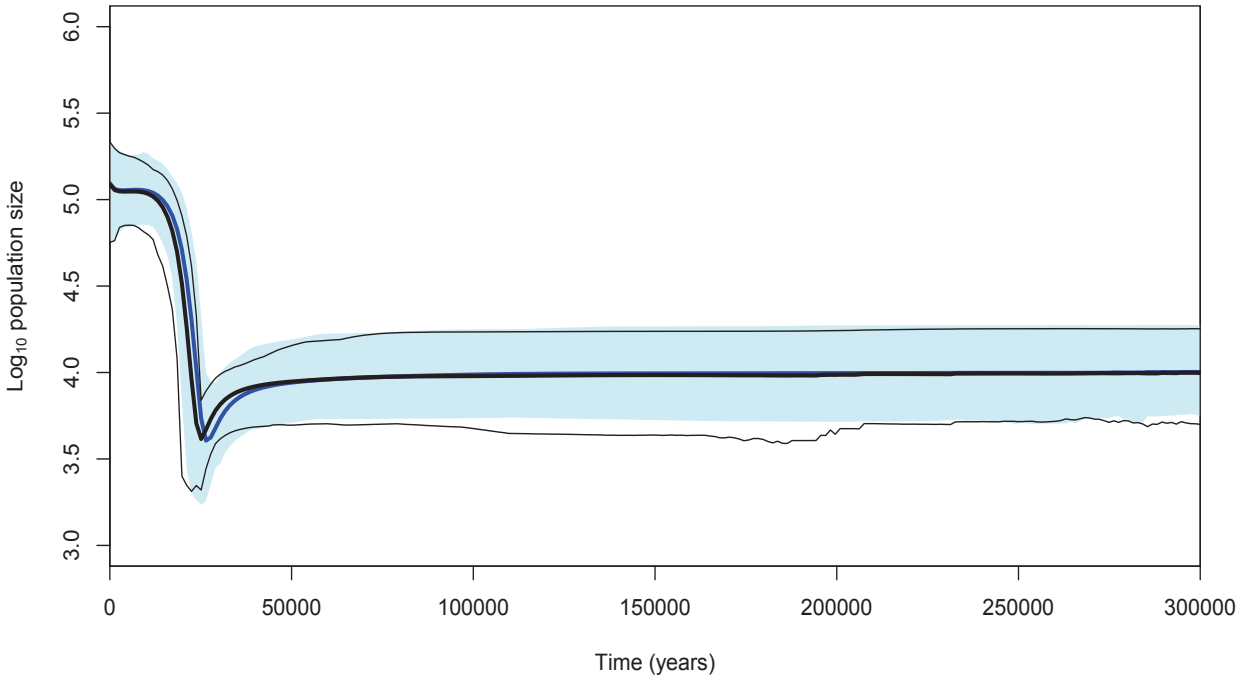


Figure C.4: Population curves from the Bayesian Skyline Plot model for DNA simulated under the bottleneck demographic model. The plot shows the truly random sample's population curve (black solid line) and the quasi-random sample's population curve (solid blue line). The pale blue shaded area represents the 2.5^{th} and 97.5^{th} percentiles of the posterior means for the quasi-random samples, and the thin black lines the same but for the truly random samples. The population sizes here are shown on the log scale.

C.3 Exponential demographic model

Lastly, we consider the exponential population growth demographic model. Figure C.5 shows, as before, the three boxplots comparing the tree properties and genetic variation between the truly random and quasi-random samples. From Figure C.5 there is only a very small difference between the two samples. The quasi-random sample has a slightly larger tree depth, larger tree length and smaller mean pairwise difference than what we see with the truly random sample. However, none of these difference are significant at the two-sample t test at 5% significance level. Figure C.6 shows the inferred population size from the truly random and quasi-random samples. The pattern for the inferred population size from the exponential model follows the same pattern as before, where there is a slight difference between the two population size estimates. But the mean of the 100 posterior means for both samples are similar and the error bands for each overlap throughout time. Again, the quasi-random sample was considered a random sample going forward.

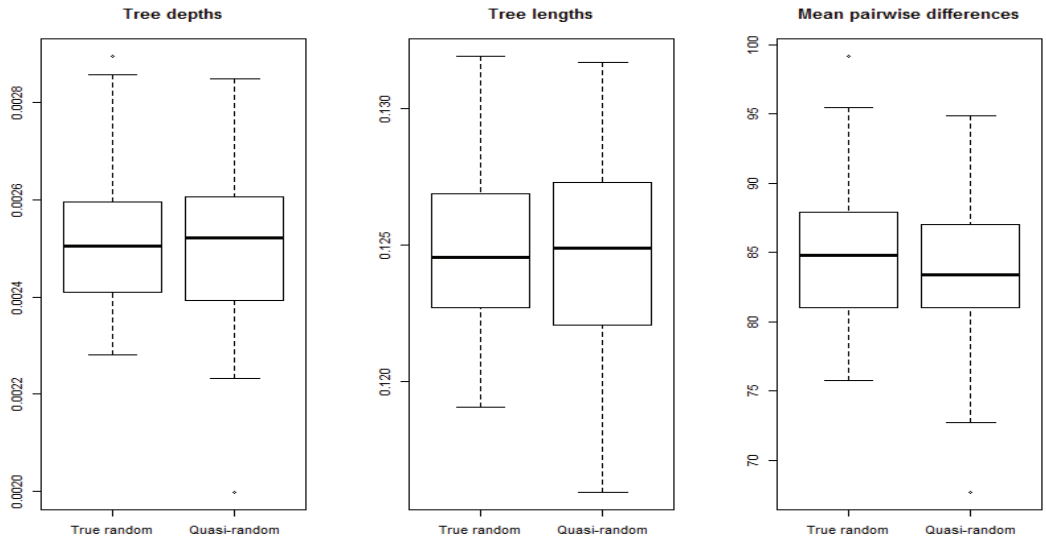


Figure C.5: Boxplots comparing the tree properties of the truly random and quasi-random simulated samples. Each of these were simulated under an exponential model. Time is measured in coalescent units.

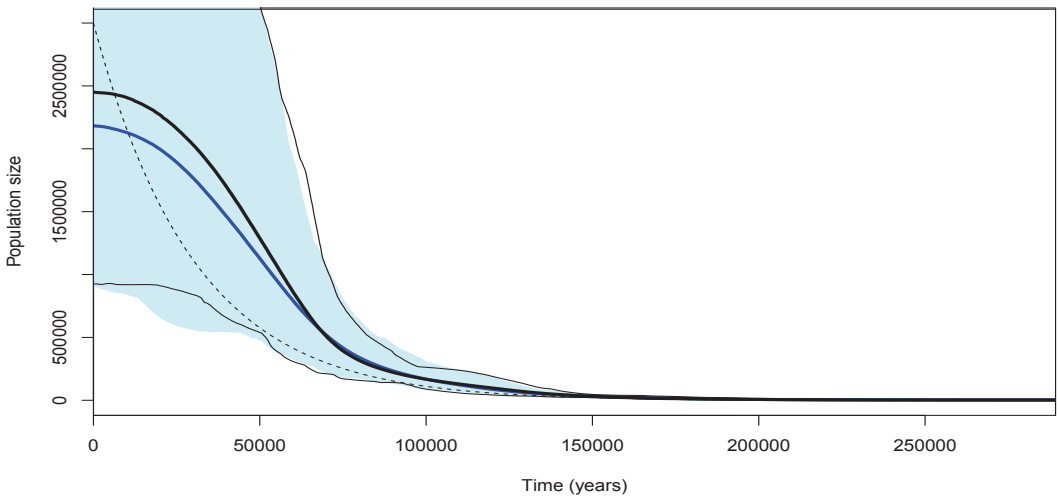


Figure C.6: Population curves from the Bayesian Skyline Plot model for DNA simulated under the exponential model. The plot shows the truly random sample's population curve (black solid line), the quasi-random sample's population curve (solid blue line) and the true population size (thin dotted line). The pale blue shaded area represents the 2.5^{th} and 97.5^{th} percentiles of the posterior means for the quasi-random samples, and the thin black lines the same but for the truly random samples.

Appendix D

FDA for Other Demographic Models

The work presented in this section continues from the analyses carried out on the population size curves inferred from data simulated under the constant population size demographic model in Section 5.1. Each of the three other demographic models will be considered here.

D.1 Step in population size

As before for the constant demographic model, we begin with the 'raw' population curves, the inferred population sizes from the Bayesian Skyline Plot model. Figure D.1 shows these population curves inferred from DNA sequences simulated under a step in population size.

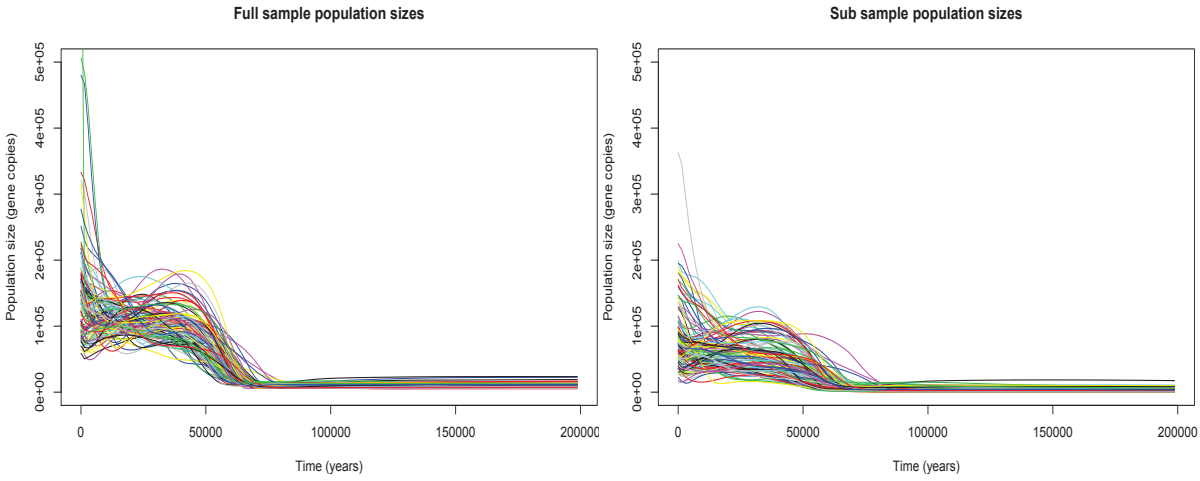
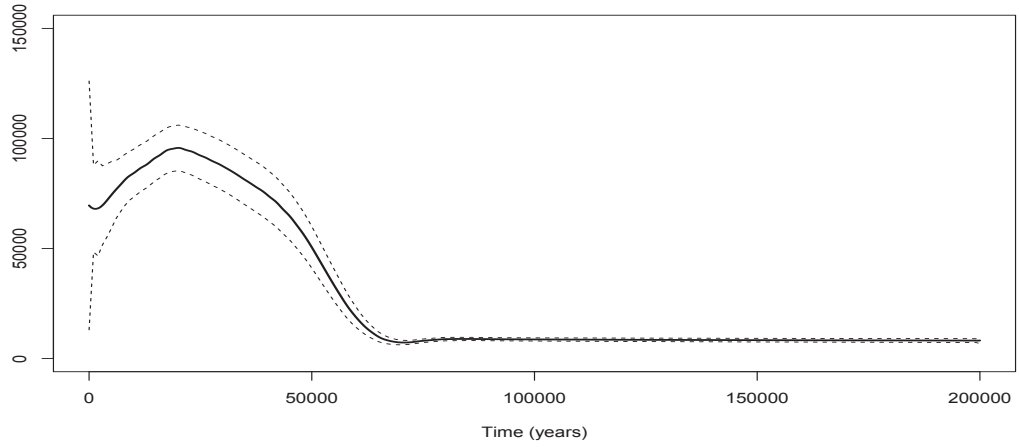


Figure D.1: Population curves for the 100 randomly and non-randomly sampled sets of DNA sequences simulated under a step in population size demographic model.

Similar to the results presented in Section 4.2.5, where the mean across simulations of the posterior means was considered, each of the individual 100 posterior mean curves recover the true population size quite well, some with more variability than others. In general, however, the

(a)



(b)

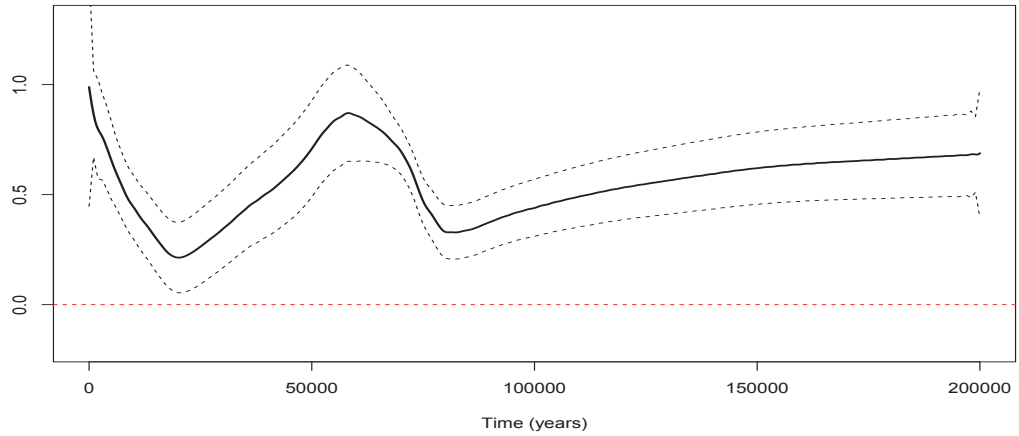


Figure D.2: Step demographic model. Intercept (a) and regression coefficient (b) functions for the concurrent model relating the non-random sample estimates to the random sample estimates, with 95% pointwise confidence intervals. The red dotted line sits at 0.

real step is rounded off in both samples. There is high variability in the initial population size, as we have seen before. We fit the concurrent model

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100) \tag{D.1}$$

to the inferred population sizes. GCV gave the values of the smoothing parameters to be $\lambda_1 = 281,000$ and $\lambda_2 = 4,620,000$. The inferred $\alpha(t)$ and $\beta(t)$ are shown in Figure D.2.

The parameter functions in Figure D.2 are consistent with those from the constant demographic model in that the intercept function captures the true population size and the regression coefficient function does not contain zero at any time point. Figure D.3 shows the result of the permutation F-test.

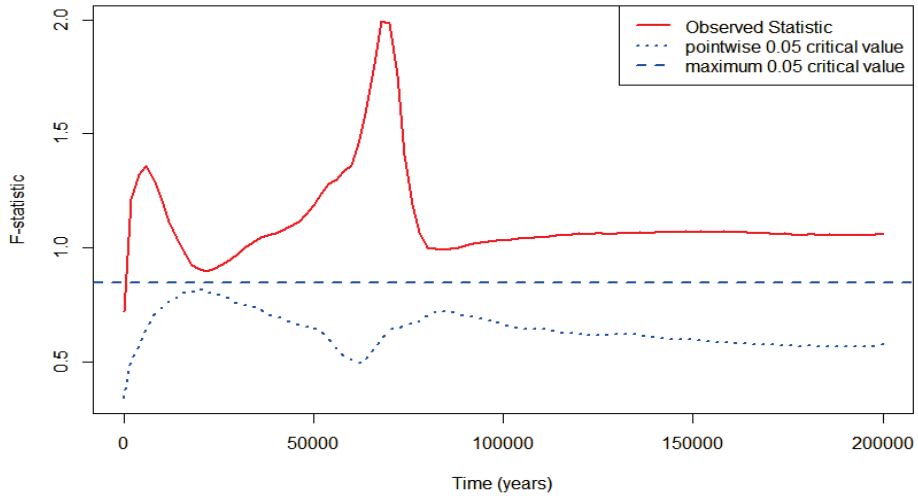


Figure D.3: The result of the permutation F-test (testing for no effect) for the concurrent model for a step demographic model, at the 5% significance level. The red line represents the observed F-statistic and the horizontal blue dashed line is the maximum critical value at 5% significance.

Again, the covariate term is significant in the model, since the observed statistic is greater than the maximum critical value at at least one point in time. In fact, the variable is significant at every time point since the pointwise critical value is always below the observed test statistic. Figure D.4 shows observed and predicted population sizes from the concurrent model for a selection of four simulations.

Like in the constant case, in all simulations (not only those shown here) the model does extraordinarily well at predicting the true underlying population size from the inferred subsample population size. As before, we remove this intercept term from the model and fit the following model,

$$Y_i(t) = \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100).$$

Figure D.5 shows the permutation F-test and Figure D.6 shows the regression coefficient function for this no-intercept model.

Again, the covariate is significant in the model (Figure D.5) at every time point. In Figure D.6 we see that the model captures the correct pattern with an increasing coefficient function over time with a dip at the change in population size but again this coefficient is underestimated. An interesting feature is that in each of the epochs the coefficient function is almost a straight line, indicating that the difference between the two sets of population curves may only be a scale factor. Figure D.7 shows the same set of four simulations and their predicted population sizes from the model.

As in the constant case, the predicted population size does not capture either the true population size or the inferred population size from the random sample. Lastly, Figure D.8 shows the

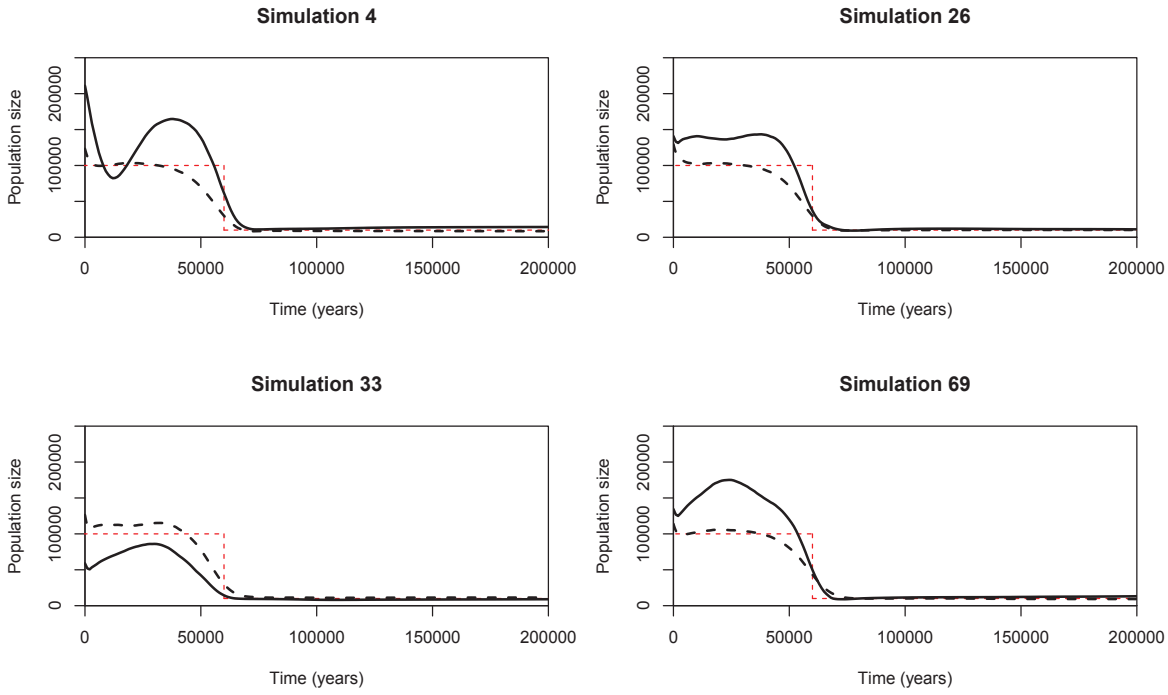


Figure D.4: A selection of inferred and predicted population sizes for the random sample of DNA sequences under a step demographic model. The inferred population size is the black solid line, the black dotted line is the predicted population size and the red dotted line is the true population size.

intercept-only concurrent model. As in the constant case, the intercept-only model simply highlights that the Bayesian Skyline Plot model results are unbiased for the random sample (Figure D.8). In both the full model and the no-intercept model, the coefficient function is significant.

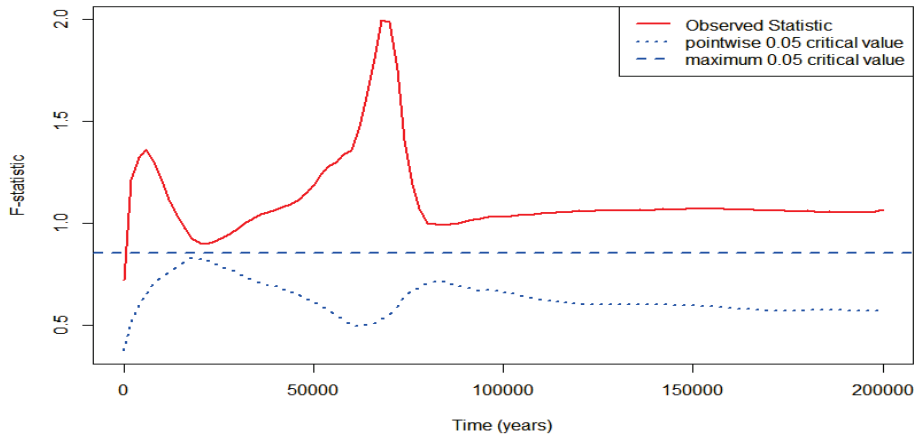


Figure D.5: The permutation F-test (testing for no effect) for the no-intercept concurrent model under a step demographic model, at the 5% significance level.

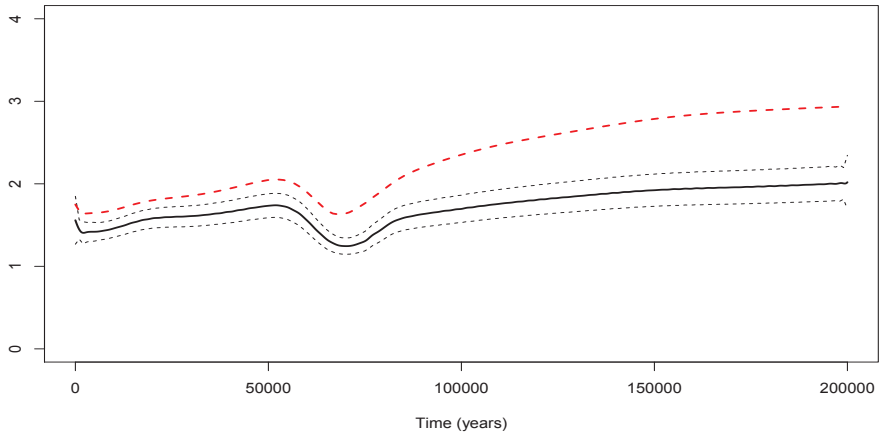


Figure D.6: Regression coefficient function (black solid line) for the no-intercept concurrent model for DNA sequences simulated under the TN93 mutation model and a step demographic model, with 95% confidence intervals (black dashed lines). The true ratio is shown by the red dashed line.

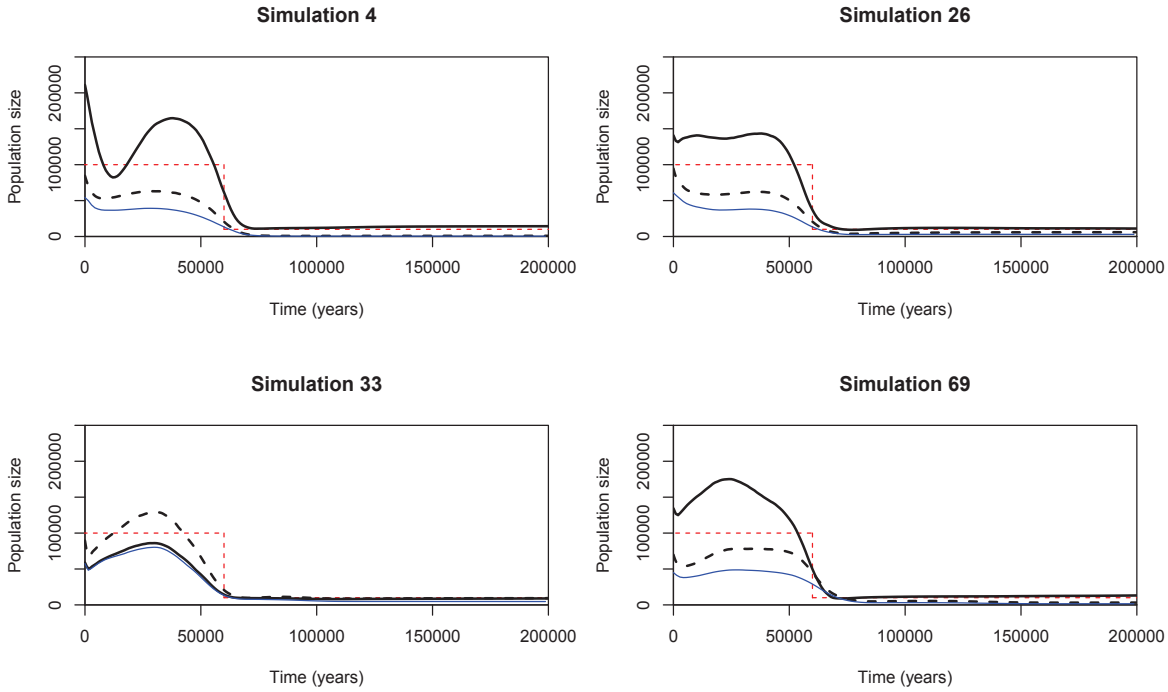


Figure D.7: The same selection of simulations from the step model giving inferred and predicted population sizes as in Figure D.4. As before, the inferred population size from the full sample is the black solid line, the predicted population size the black dashed line, the true population size the red dashed line and the blue line is the corresponding inferred population size from the subsample.

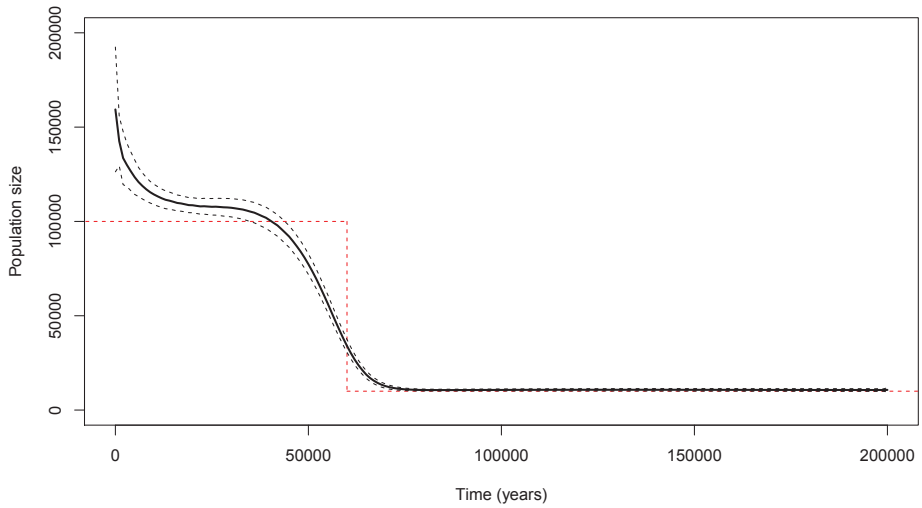


Figure D.8: Intercept function for the intercept-only concurrent model under a step demographic model, with 95% pointwise confidence intervals. The red dotted line represents the true demographic model.

D.2 Bottleneck model

Now with the bottleneck model we begin with the 'raw' population curves again, the inferred population sizes from the Bayesian Skyline Plot model. Figure D.9 shows these population curves inferred from DNA sequences simulated under a bottleneck population model.

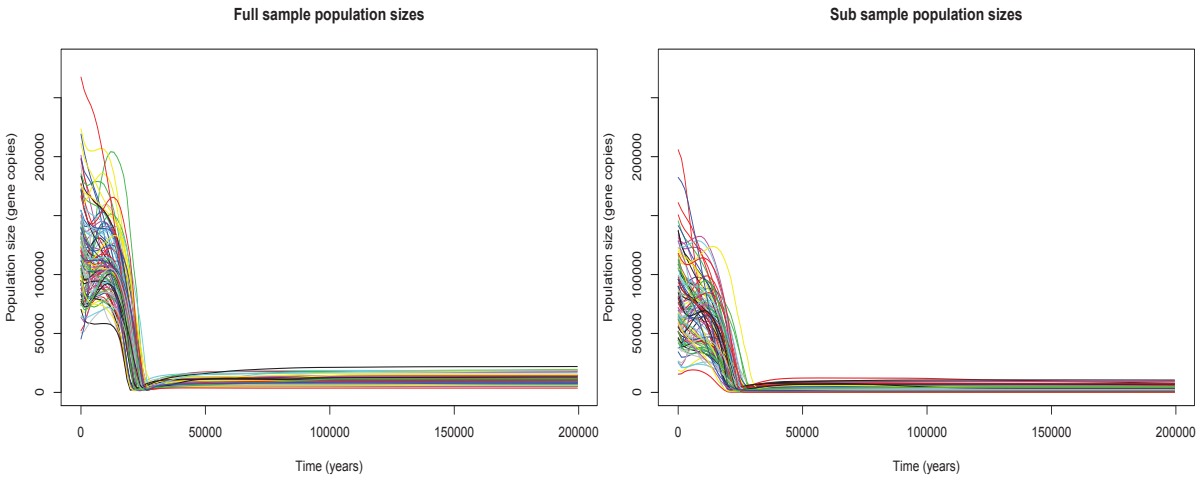


Figure D.9: Population curves for the 100 randomly and non-randomly sampled sets of DNA sequences simulated under a bottleneck population demographic model.

Again, there is high variability in these raw population curves but the model does capture the correct time of the population changes. The bottleneck (the increasing at 26,500 years ago) is only very slightly captured by the full sample and not really by the subsample. We fit the concurrent model as before

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100) \tag{D.2}$$

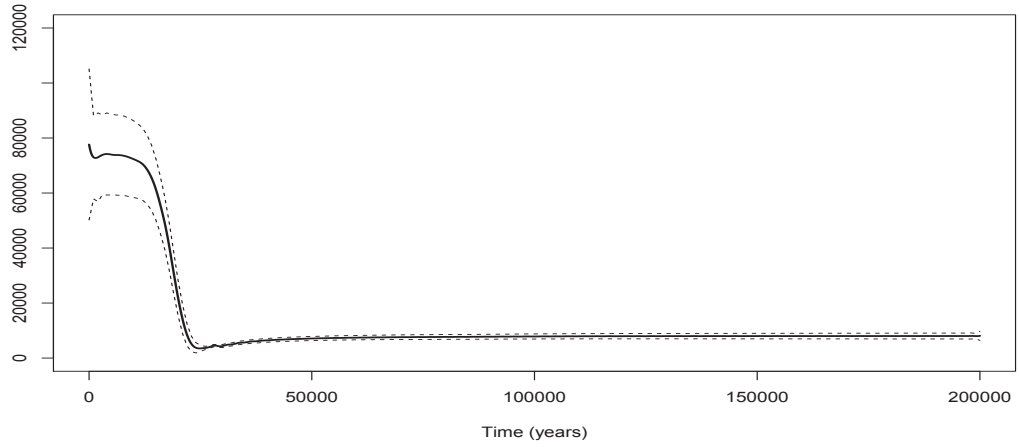
to the inferred population sizes. GCV gave the values of the smoothing parameters to be $\lambda_1 = 142000$ and $\lambda_2 = 214000$. The inferred $\alpha(t)$ and $\beta(t)$ are shown in Figure D.10.

The intercept coefficient function (Figure D.10(a)) captures the true population size in general and the regression coefficient function only just contains zero at one time point, which happens to be during the bottleneck. To test the significance of the coefficient of the subsample population size in the model we carry out a permutation F-test as before (Figure D.11).

The test highlights what we seen in the regression coefficient function in Figure D.10(b), in that at one time point during the bottleneck, the term is not significant in the model. However, the observed statistic is only just below the pointwise critical value. Next we assess how well the model predicts the full sample population size. Figure D.12 shows a selection of four simulations.

Like in the constant case, in all simulations (not only those shown here) the model does extraordinarily well at predicting the true underlying population size from the inferred subsample

(a)



(b)

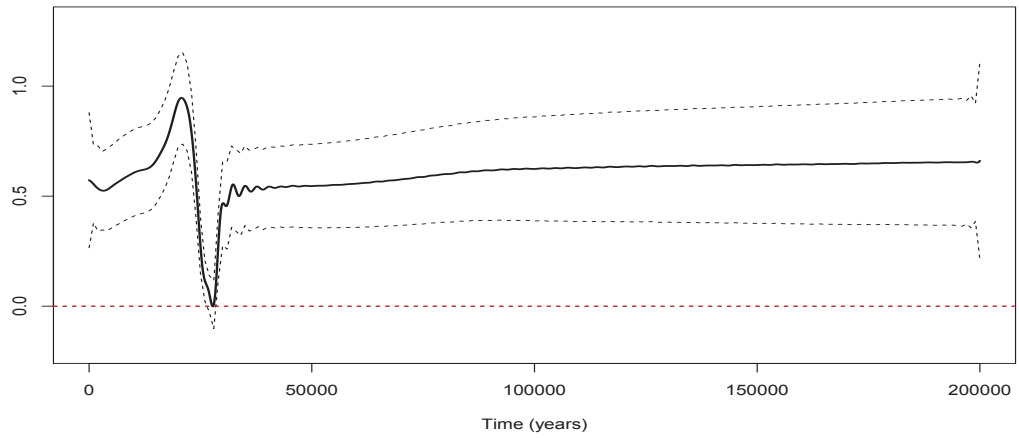


Figure D.10: Bottleneck demographic model. Intercept (a) and regression coefficient (b) functions for the concurrent model relating the non-random sample estimates to the random sample estimates, with 95% pointwise confidence intervals. The red dotted line sits at 0.

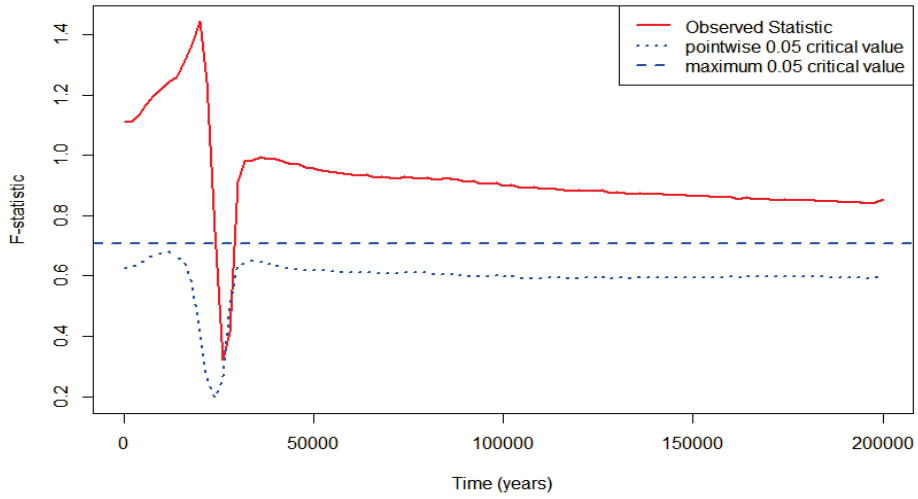


Figure D.11: The result of the permutation F-test (testing for no effect) for the concurrent model for a bottleneck demographic model, at the 5% significance level. The red line represents the observed F-statistic and the horizontal blue dashed line is the maximum critical value at 5% significance.

population size. As before, we remove this intercept term from the model and fit the following model

$$Y_i(t) = \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100).$$

Figure D.13 shows the permutation F-test and Figure D.14 shows the regression coefficient function for this no-intercept model. Again, the covariate term is significant in the model (Figure D.13) at every time point apart from during the bottleneck, but again the lines of the observed statistic and pointwise critical value only just cross. In Figure D.14 we see that the model captures the correct pattern with an increasing coefficient function over time with a dip at the change in population size but again this coefficient is underestimated. As before, in each of the epochs the coefficient function is almost a straight line, indicating that the difference between the two sets of population curves may only be a scale factor. Figure D.15 shows the same set of four simulations and their predicted population sizes from the model. Like in the constant and step cases, the predicted population size does not capture either the true population size or the inferred population size from the random sample. Lastly, Figure D.16 shows the intercept-only concurrent model. As in the previous cases, the intercept-only model simply highlights that the Bayesian Skyline Plot model results are unbiased for the random sample (Figure D.16). In both the full model and the no-intercept model, the coefficient function is significant.

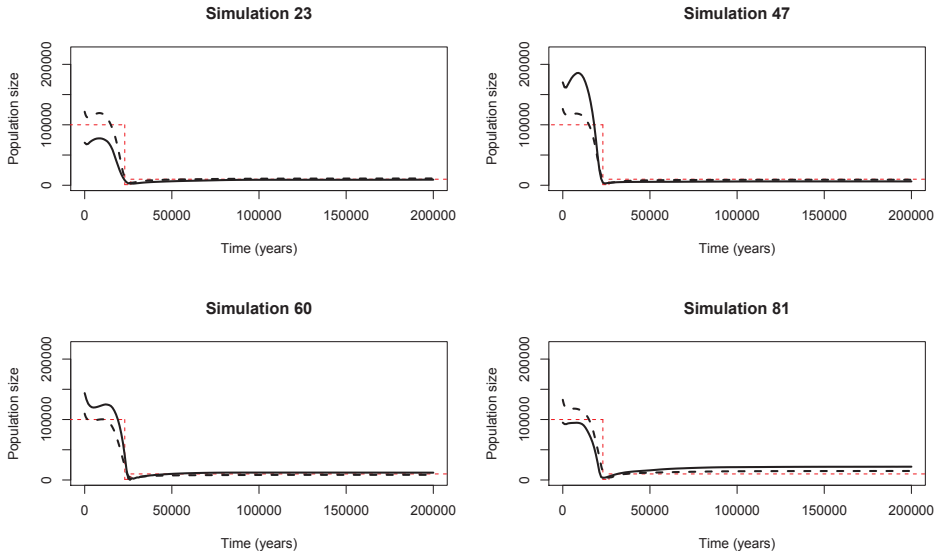


Figure D.12: A selection of inferred and predicted population sizes for the random sample of DNA sequences under a bottleneck demographic model. The inferred population size is the black solid line, the black dotted line is the predicted population size and the red dotted line is the true population size.

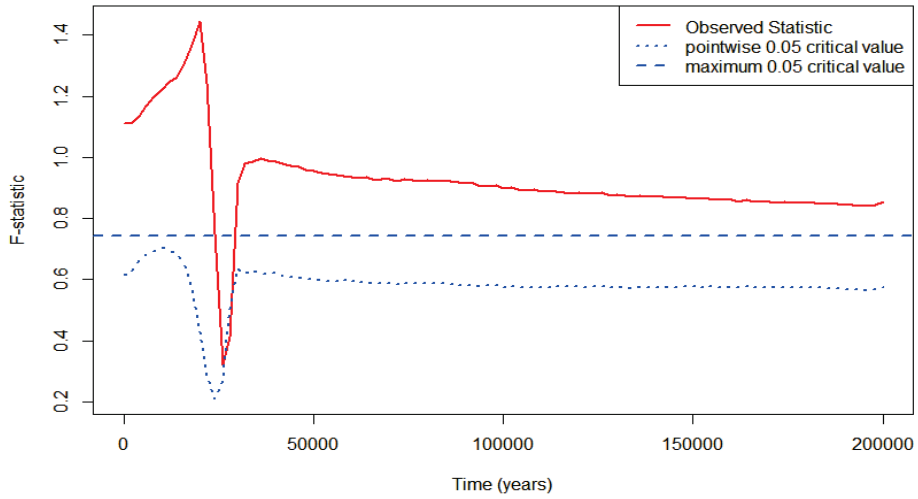


Figure D.13: The permutation F-test (testing for no effect) for the no-intercept concurrent model under a bottleneck demographic model, at the 5% significance level.

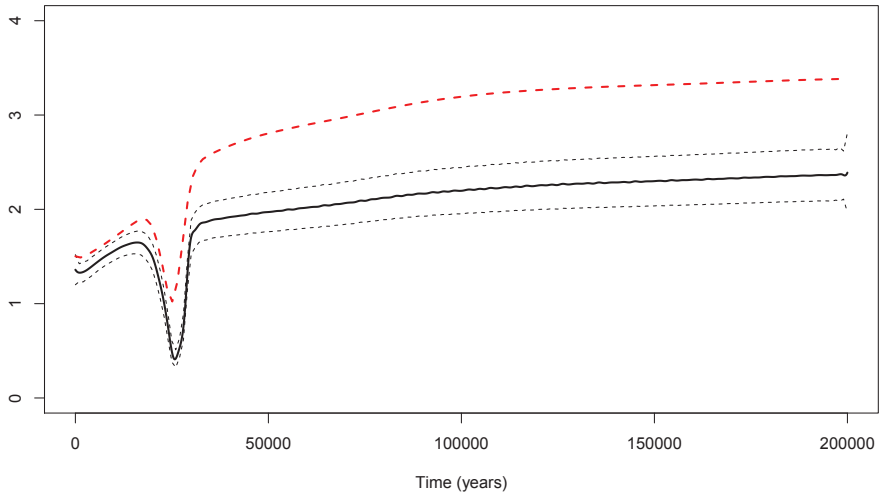


Figure D.14: Regression coefficient function (black solid line) for the no-intercept concurrent model for DNA sequences simulated under the TN93 mutation model and a bottleneck demographic model, with 95% confidence intervals (black dashed lines). The true ratio is shown by the red dashed line.

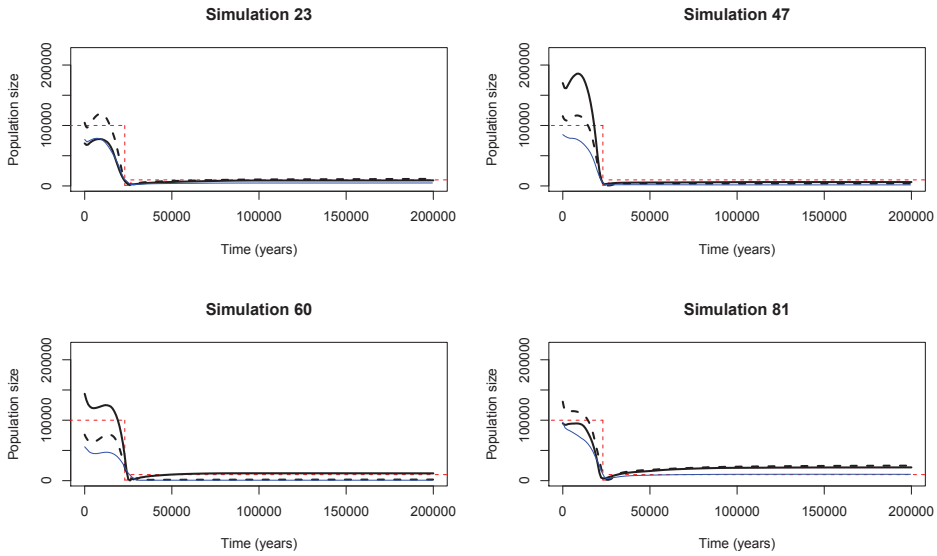


Figure D.15: The same selection of simulations from the bottleneck model giving inferred and predicted population sizes as in Figure D.12. As before, the inferred population size from the full sample is the black solid line, the predicted population size the black dashed line, the true population size the red dashed line and the blue line is the corresponding inferred population size from the subsample.

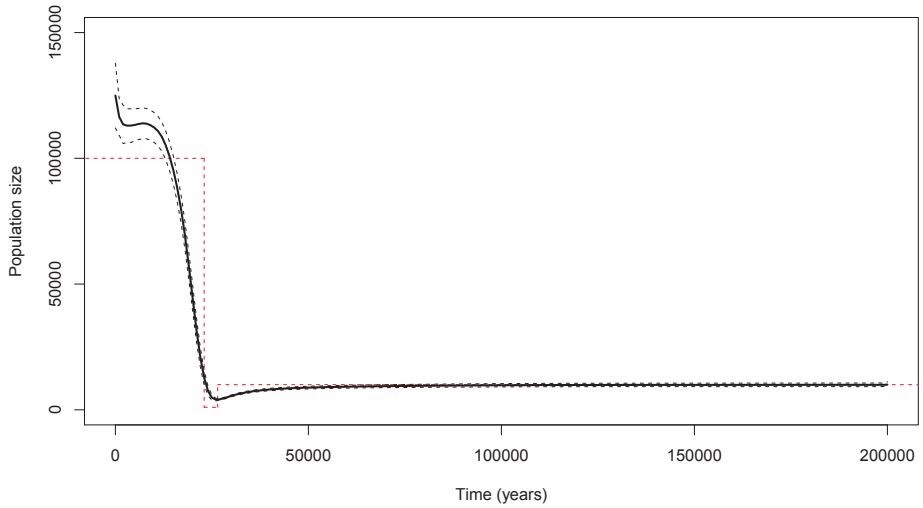


Figure D.16: Intercept function for the intercept-only concurrent model under a bottleneck demographic model, with 95% pointwise confidence intervals. The red dotted line represents the true demographic model.

D.3 Exponential growth in population size

Lastly, we move on to the exponential demographic model. Figure D.17 shows the raw data curves for the full and subsamples.

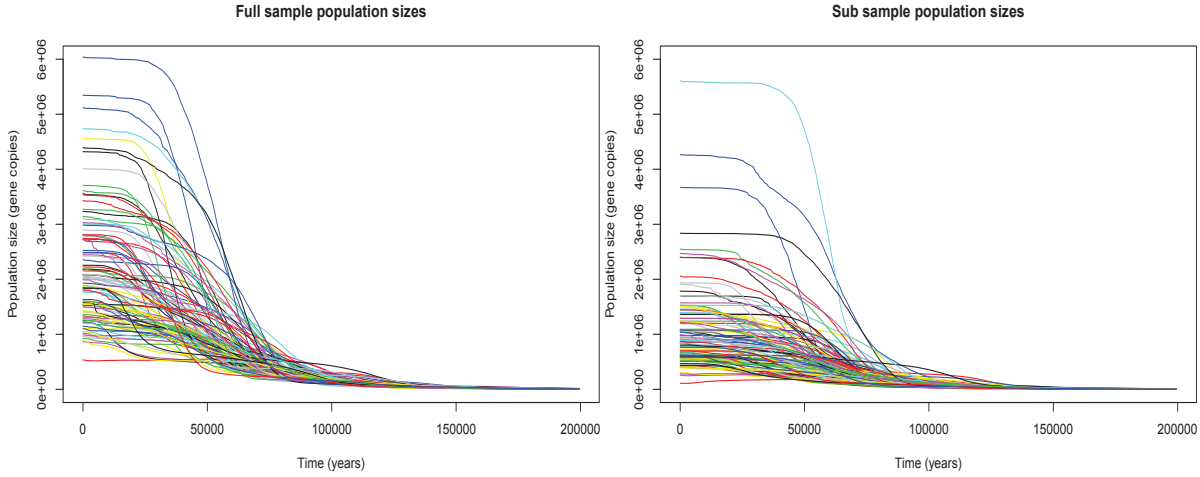


Figure D.17: Population curves for the 100 randomly and non-randomly sampled sets of DNA sequences simulated under an exponential growth population demographic model.

Again, we see lots of variation in the data, although the model does capture the overall exponentially decaying shape of the true population size. Next, we fit the concurrent model

$$Y_i(t) = \alpha(t) + \beta(t)X_i(t) + \varepsilon_i(t) \tag{D.3}$$

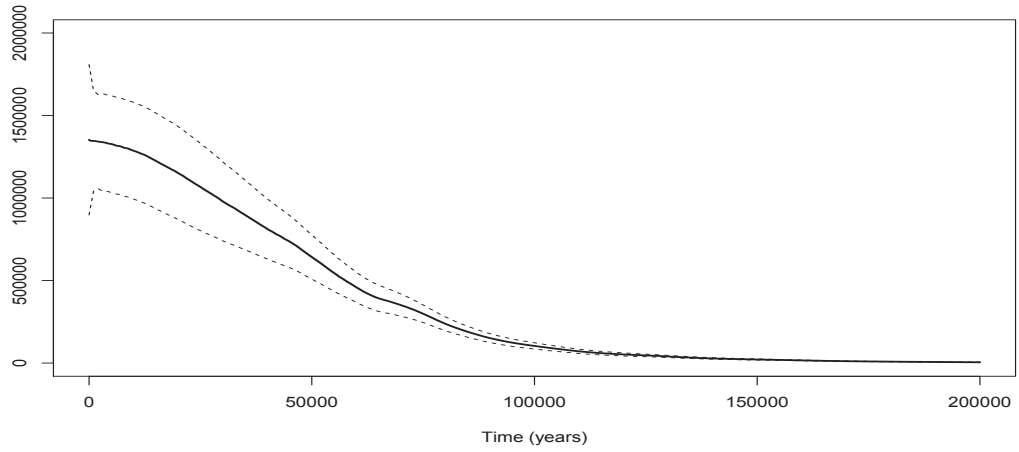
to the inferred population sizes. GCV gave the values of the smoothing parameters to be $\lambda_1 = 1.7 \times 10^8$ and $\lambda_2 = 1.2 \times 10^8$. The inferred $\alpha(t)$ and $\beta(t)$ are shown in Figure D.18.

The parameter functions in Figure D.2 are consistent with those from the constant demographic model in that the intercept function captures the true population size and the regression coefficient function does not contain zero at any time point. Figure D.3 shows the result of the permutation F-test. In this case, the coefficient is significant apart from a time around 70,000 years ago. We can assess the model in terms of its ability to predict the full sample population size shown in Figure D.20. As in the previous three cases, the model does particularly well at predicting the true population size. Next, we look at the no-intercept model

$$Y_i(t) = \beta(t)X_i(t) + \varepsilon_i(t), \quad (i = 1, \ldots, 100).$$

Figure D.21 shows the permutation F-test and Figure D.22 shows the regression coefficient function for this no-intercept model. As in all previous cases, the coefficient is significant in the model at all time points. Figure D.22 shows the coefficient curve. With a slight dip at around 50,000 years ago, the regression coefficient function lies near 1.5 for the exponential

(a)



(b)

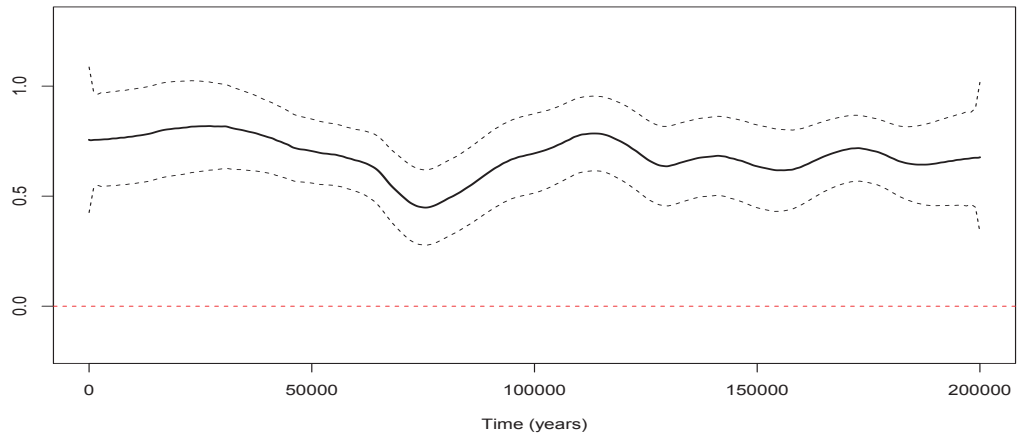


Figure D.18: Exponential demographic model. Intercept (a) and regression coefficient (b) functions for the concurrent model relating the non-random sample estimates to the random sample estimates, with 95% pointwise confidence intervals. The red dotted line sits at 0.

case, indicating that there could be a factor of 1.5 between the two sets of inferred population sizes. Figure D.23 show the same selection of simulations and their predicted values. As in all cases, this model does not do as well at capturing the true population size. Lastly, Figure D.24 shows the intercept-only model, which does particularly well and further highlights how unbiased the Bayesian Skyline Plot is with a full sample. Again, both the full and no-intercept models show that the coefficient of the inferred population size of the non-randomly sampled DNA sequences is significant in both cases.

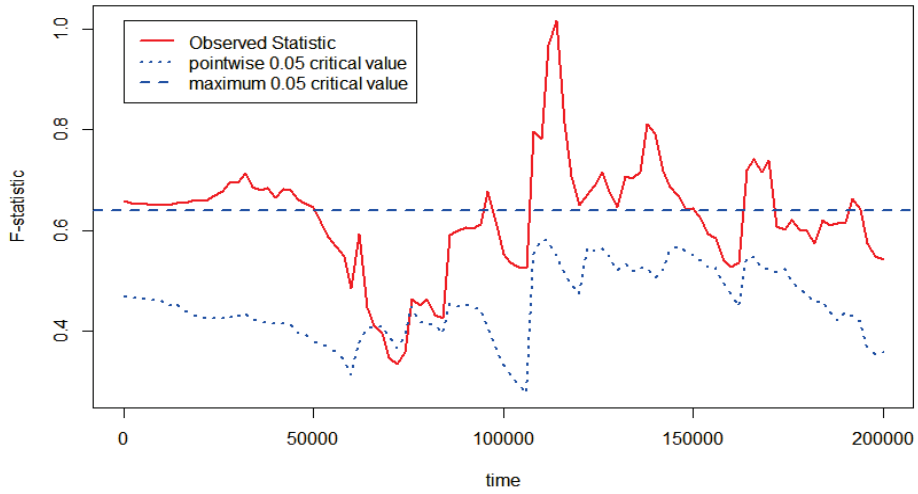


Figure D.19: The result of the permutation F-test (testing for no effect) for the concurrent model for an exponential model, at the 5% significance level. The red line represents the observed F-statistic and the horizontal blue dashed line is the maximum critical value at 5% significance.

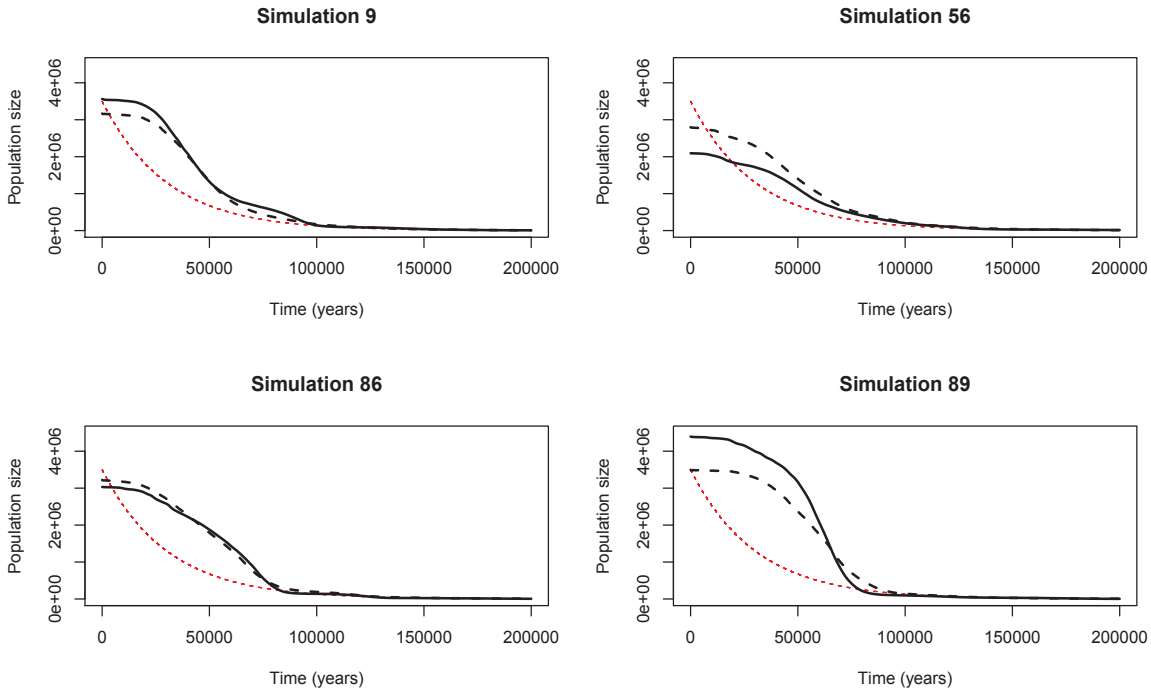


Figure D.20: A selection of inferred and predicted population sizes for the random sample of DNA sequences under an exponential model. The inferred population size is the black solid line, the black dotted line is the predicted population size and the red dotted line is the true population size.

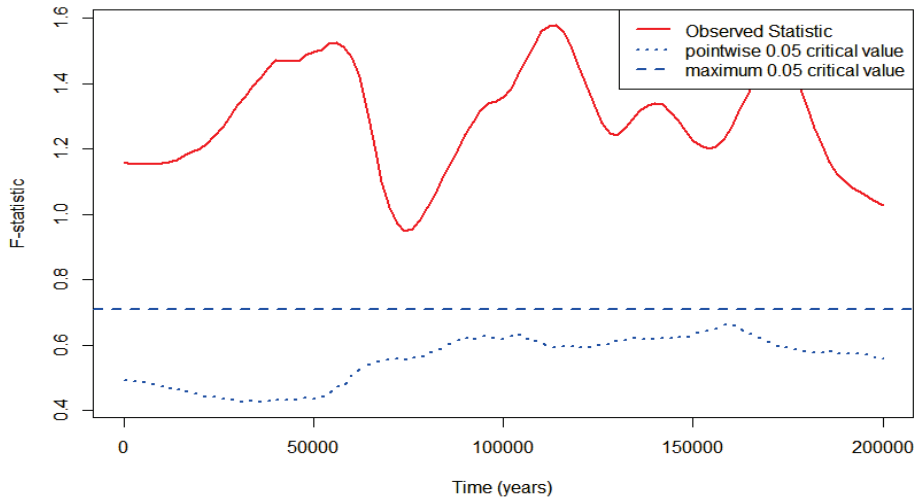


Figure D.21: The permutation F-test (testing for no effect) for the no-intercept concurrent model under an exponential model, at the 5% significance level.

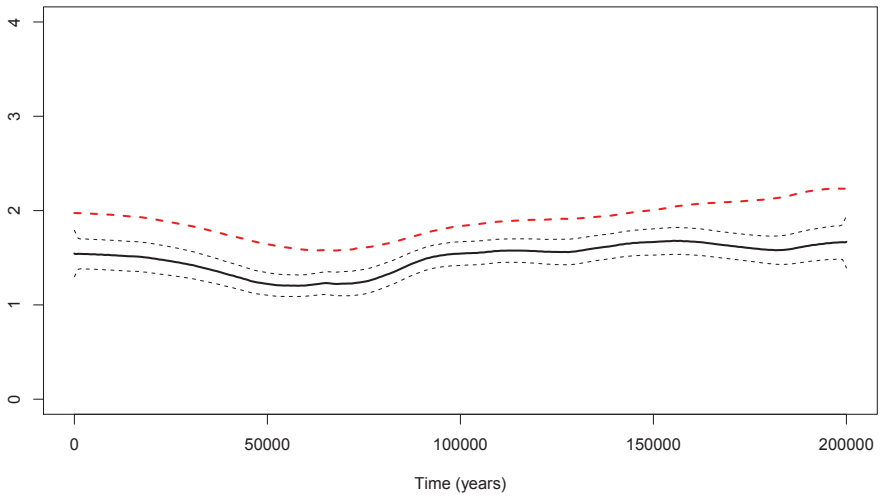


Figure D.22: Regression coefficient function (black solid line) for the no-intercept concurrent model for DNA sequences simulated under the TN93 mutation model and an exponential model, with 95% confidence intervals (black dashed lines). The true ratio is shown by the red dashed line.

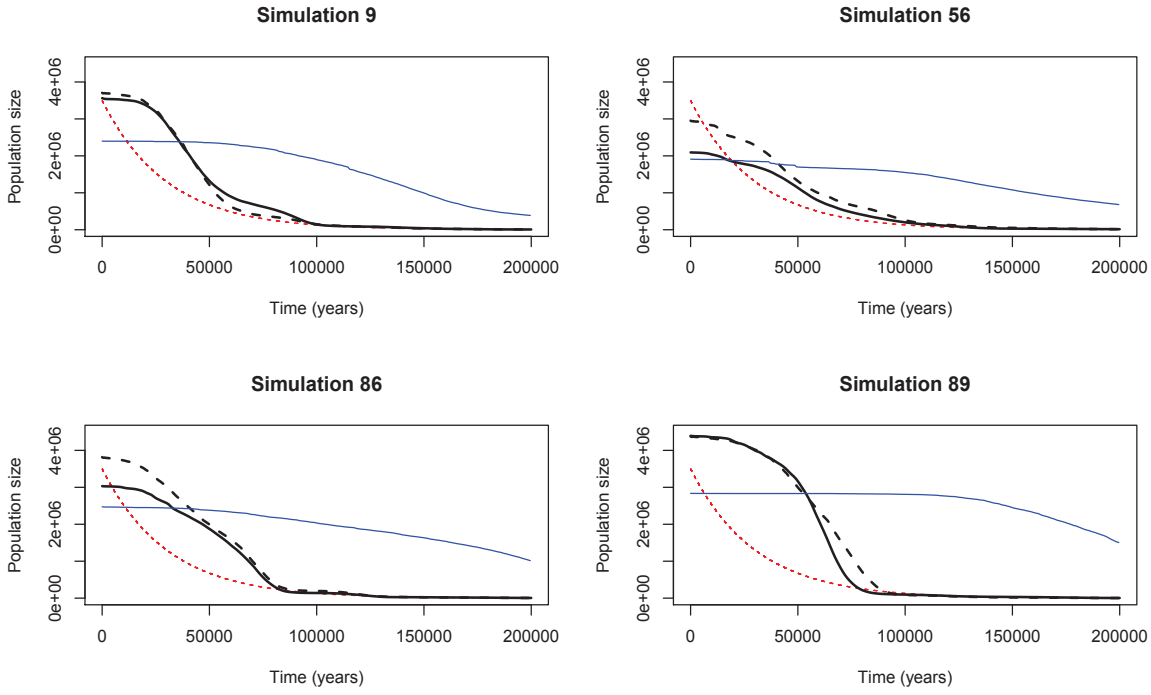


Figure D.23: The same selection from the exponential model of simulations giving inferred and predicted population sizes as in Figure D.20. As before, the inferred population size from the full sample is the black solid line, the predicted population size the black dashed line, the true population size the red dashed line and the blue line is the corresponding inferred population size from the subsample.

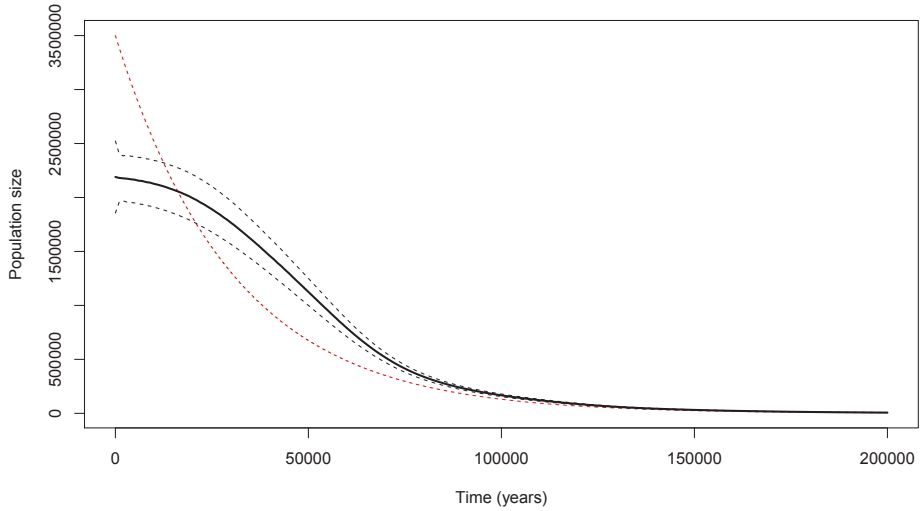


Figure D.24: Intercept function for the intercept-only concurrent model under an exponential model, with 95% pointwise confidence intervals. The red dotted line represents the true demographic model.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

S. Anderson, A.T. Bankier, B.G. Barrell, M.H.L. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, et al. Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–465, 1981.

R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, and N. Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23:147, 1999.

Q.D. Atkinson, R.D. Gray, and A.J. Drummond. mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Molecular Biology and Evolution*, 25:468–474, 2007.

Q.D. Atkinson, R.D. Gray, and A.J. Drummond. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proceedings of the Royal Society B: Biological Sciences*, 276(1655):367–373, 2008.

H.-J. Bandelt, V. Macaulay, and M. (eds.) Richards. *Human mitochondrial DNA and the evolution of Homo sapiens*. Springer, Berlin, 2006.

M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

D.M. Behar, R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, et al. The dawn of human matrilineal diversity. *The American Journal of Human Genetics*, 82(5):1130–1140, 2008.

S. Behjati and P.S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98:236–238, 2013.

A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368: 455–457, 1994.

A.W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Oxford University Press, Oxford, 1997.

B. Brinkmann, M. Klintschar, F. Neuhuber, J. Hühne, and B. Rolf. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, 62:1408–1415, 1998.

K. P. Burnham and D. R. Anderson. *Model selection and Multimodel inference: a practical information-theoretic approach*. Springer, New York, 1998.

R. L. Cann and A. C. Wilson. Length mutations in human mitochondrial DNA. *Genetics*, 104: 699–711, 1983.

R.L. Cann, M. Stoneking, and A.C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31, 1987.

P.U. Clark, A.S. Dyke, J.D. Shakun, A.E. Carlson, J. Clark, B. Wohlfarth, J.X. Mitrovica, S.W. Hostetler, and A.M. McCabe. The last glacial maximum. *Science*, 325:710–714, 2009.

P. Craven and G. Wahba. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

K. Csilléry, M.G.B. Blum, O.E. Gaggiotti, and O. François. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410–418, 2010.

C. Darwin. *On the origin of species by Means of Natural Selection*. John Murray, London, 1859.

W. E. Deming. *Statistical adjustment of data.* Wiley, New York, 1943.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22: 1185–1192, 2005.

A.J. Drummond, G.K. Nicholls, A.G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320, 2002.

P. Endicott and S.Y.W. Ho. A Bayesian evaluation of human mitochondrial substitution rates. *The American Journal of Human Genetics*, 82(4):895–902, 2008.

L. Excoffier, I. Dupanloup, E. Huerta-Sánchez, V.C. Sousa, and M. Foll. Robust demographic inference from genomic and SNP data. *PLOS Genetics*, 9:e1003905, 2013.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: models, methods and applications*. Springer, New York, 2013.

J. Felsenstein. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, 35:1229–1242, 1981.

J. Felsenstein. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetics Research*, 59: 139–147, 1992.

J. Felsenstein. *Inferring phylogenies*. Sinauer, Sunderland, Massachusetts, 2004.

J.N. Fenner. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 128:415–423, 2005.

S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

A. Gelman, D. B. Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

R.E. Giles, H. Blanc, H.M. Cann, and D.C. Wallace. Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 77:6715–6719, 1980.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

R.C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B*, 344:403–410, 1994.

R.N. Gutenkunst, R.D. Hernandez, S.H. Williamson, and C.D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics*, 5:e1000695, 2009.

T. Hall. Bioedit: an important software for molecular biology. *GERF Bulletin of Biosciences*, 2:60–61, 2011.

D. L. Hartl and A. G. Clark. *Principles of population genetics*. Sinauer, Sunderland Massachusetts, 1997.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

M. Hasegawa, H. Kishino, and T. Yano. Man's place in hominoidea as inferred from molecular clocks of DNA. *Journal of Molecular Evolution*, 26:132–147, 1987.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55:757–779, 1993.

A. Helgason, A.W. Einarsson, V.B. Guðmundsdóttir, Á. Sigurðsson, E.D. Gunnarsdóttir, A. Jagadeesan, S.S. Ebenesersdóttir, A. Kong, and K. Stefánsson. The Y-chromosome point mutation rate in humans. *Nature Genetics*, 47(5):453, 2015.

R. Heller, L. Chikhi, and H. R. Siegismund. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PlOS ONE*, 8(5):e62992, 2013.

L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer, New York, 2012.

M. Ingman, H. Kaessmann, S. Pääbo, and U. Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408:708–713, 2000.

ISGG International Society of Genetic Genealogy. Generation length, 2015. URL `https://isogg.org`. (visited on 03/04/2017).

M. Jobling, E. Hollox, M. Hurles, T. Kivisild, and C. Tyler-Smith. *Human Evolutionary Genetics, 2nd edn.* Garland, New York, 2014.

T. H. Jukes and C. R. Cantor. *Evolution of protein molecules. In H.N. Munro (ed.)*, volume Mammalian Protein Metabolism. Academic Press, New York, 1969.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

M. Kimura and J. F. Crow. The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738, 1964.

M. Kimura and G.H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576, 1964.

J. F. C. Kingman. On the genealogy of large populations. *Journal of applied probability*, 19: 27–43, 1982.

T. Kivisild, M. Metspalu, H.-J. Bandelt, M. Richards, and R. Villems. The world mtDNA phylogeny. In *Human mitochondrial DNA and the evolution of Homo sapiens*, pages 149–179. Springer, 2006.

R.G. Klein. The archeology of modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews*, 1:5–14, 1992.

P. Kokoszka and M. Reimherr. *Introduction to Functional Data Analysis*. Chapman and Hall, London, 2017.

Q-P. Kong, Y-G. Yao, C. Sun, H. Bandelt, C-L. Zhu, and Y-P. Zhang. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *American Journal of Human Genetics*, 73:671–676, 2003.

M.K. Kuhner, J. Yamato, and J. Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, 149:429–434, 1998.

R.C. Lewontin and J.L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, 54:595–609, 1966.

H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475:493–496, 2011.

D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227 (4693):1435–1441, 1985.

J.S. Lopes and M.A. Beaumont. ABC: a useful Bayesian tool for the analysis of population data. *Infection, Genetics and Evolution*, 10:825–832, 2010.

N. Maca-Meyer, A.M. González, J.M. Larruga, C. Flores, and V.M. Cabrera. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics*, 2:13, 2001.

V. Macaulay, C. Hill, A. Achilli, C. Rengo, D. Clarke, W. Meehan, J. Blackburn, O. Semino, R. Scozzari, F. Cruciani, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308:1034–1036, 2005.

D.R. Maddison, D.L. Swofford, and W.P. Maddison. NEXUS: an extensible file format for systematic information. *Systematic biology*, 46(4):590–621, 1997.

A.M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74:560–564, 1977.

I. Mayrose, N. Friedman, and T. Pupko. A gamma mixture model better accounts for among site rate heterogeneity. *Bioinformatics*, 21:151–158, 2005.

P. Mellars. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proceedings of the National Academy of Sciences*, 103:9381–9386, 2006.

D. Mishmar, E. Ruiz-Pesini, P. Golik, V. Macaulay, A.G. Clark, S. Hosseini, M. Brandon, K. Easley, E. Chen, M.D. Brown, et al. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences*, 100:171–176, 2003.

P.A.P. Moran. *The statistical processes of evolutionary theory*. Clarendon Press, Oxford., 1962.

NIH National Human Genome Research Institute. The human genome project, 2019. URL `https://www.genome.gov/human-genome-project`. (visited on 04/10/2019).

S. Nee, E. C. Holmes, A. Rambaut, and P. H. Harvey. Inferring population history from molecular phylogenies. *Philosophical Transactions of the Royal Society of London. Series B*, 349: 25–31, 1995.

H.C. Neu. The crisis in antibiotic resistance. *Science*, 257:1064–1073, 1992.

R. Nielsen and J. Wakeley. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158:885–896, 2001.

R. Opgen-Rhein, L. Fahrmeir, and K. Strimmer. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*, 5:6, 2005.

J.A. Palacios and V.N. Minin. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics*, 69:8–18, 2013.

E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

A. Polanski, M. Kimmel, and R. Chakraborty. Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. *Proceedings of the National Academy of Sciences*, 95:5456–5461, 1998.

O. G. Pybus, A. Rambaut, and P. H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155:1429–1437, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

A. Rambaut, A.J. Drummond, D. Xie, G. Baele, and M.A. Suchard. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67:901–904, 2018.

J. Ramsay, G. Hooker, and S. Graves. *Functional Data Analysis with R and MATLAB*. Springer, New York, 2009.

J.O. Ramsay. Functional data analysis. *Encyclopedia of Statistical Sciences, Wiley online*, 2004.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

J.O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2018. URL `https://CRAN.R-project.org/package=fda`. (R package version 2.4.8).

J.A. Reuter, D.V. Spacek, and M.P. Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58:586–597, 2015.

J.A. Rice. *Mathematical statistics and data analysis (3rd ed.)*. Brooks/Cole, CA, 2006.

T. Rito, M.B. Richards, V. Fernandes, F. Alshamali, V. Cerny, L. Pereira, and P. Soares. The first modern human dispersals across Africa. *PLOS ONE*, 8:e80031, 2013.

A.R. Rogers and H. Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, 9:552–569, 1992.

F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74:5463–5467, 1977.

M. Slatkin and B. Rannala. Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics*, 60:447–458, 1997.

L.M. Smith, J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B.H. Kent, and L.E. Hood. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321:674–679, 1986.

G.W. Snedecor and W.G. Cochran. *Statistical methods*. Iowa State University Press, IA, 1980.

P. Soares, F. Alshamali, J.B. Pereira, V. Fernandes, N.M. Silva, C. Afonso, M.D. Costa, E. Musilová, V. Macaulay, M.B. Richards, et al. The expansion of mtDNA haplogroup L3 within and out of Africa. *Molecular Biology and Evolution*, 29:915–927, 2011.

C.C. Spencer, J.E. Neigel, and P.L. Leberg. Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. *Molecular Ecology*, 9:1517–1528, 2000.

K. Strimmer and O. G. Pybus. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18:2298–2305, 2001.

M.A. Suchard, P. Lemey, G. Baele, D.L. Ayres, A.J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4: vey016, 2018.

F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105: 437–460, 1983.

K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10: 512–526, 1993.

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Mathematical Questions in Biology*, pages 57–86, 1986.

S. Vijayraghavan, S. G. Kozmin, P. K. Strope, D. A. Skelly, Z. Lin, J. Kennell, P. M. Magwene, F. S. Dietrich, and J. H. McCusker. Mitochondrial genome variation affects multiple respiration and non-respiration phenotypes in saccharomyces cerevisiae. *Genetics*, 211, 2018.

L. Waits, P. Taberlet, J.E. Swenson, F. Sandegren, and R. Franzén. Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (Ursus arctos). *Molecular Ecology*, 9:421–431, 2000.

J. Wakeley. *Coalescent theory: an introduction*. Roberts & Company, Colorado, 2009.

B.L. Waltoft and A. Hobolth. Non-parametric estimation of population size changes from the site frequency spectrum. *Statistical Applications in Genetics and Molecular Biology*, 17, 2018.

J.D. Watson. The human genome project: past, present, and future. *Science*, 248(4951):44–49, 1990.

G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.

I.J. Wilson and D.J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150(1): 499–510, 1998.

C. Wiuf. Recombination in human mitochondrial DNA? *Genetics*, 159:749–756, 2001.

M. Woodward. *Epidemiology: study design and data analysis*. Chapman and Hall, London, 2013.

S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

H. Xu and Y.X. Fu. Estimating effective population size or mutation rate with microsatellites. *Genetics*, 166:555–563, 2004.

Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

Z. Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39: 105–111, 1994.

Z. Yang. Statistical properties of a DNA sample under the finite-sites model. *Genetics*, 144: 1941–1950, 1996.

Z. Yang. *Computational Molecular Evolution*. Oxford University Press, Oxford, 2006.

J.-T. Zhang. *Analysis of variance for functional data*. Chapman and Hall, London, 2013.

A. Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39:315–329, 1994.

E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H.J Vogel (eds). Evolving Genes and Proteins. pages 97–166. Academic Press, New York, 1965.