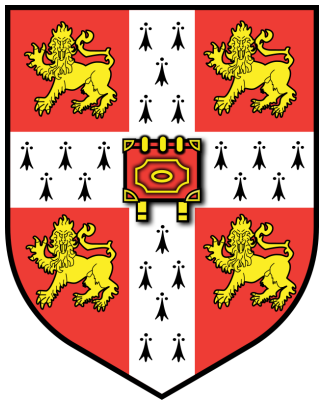
Investigating the spread of ancestry using forward spatial population genetic simulations

A thesis submitted for the
M.Phil. in Computational Biology

Dexter Goodkind



Queens' College
University of Cambridge
August 9, 2022

Word count: 14893

Acknowledgements

I'd like to thank my supervisor Aylwyn Scally for the discussions, suggestions and comprehensive support that he offered me throughout this project.

I'd also like to thank my classmates for their support throughout the year.

Declaration of Authorship

I hereby declare that this dissertation entitled *Investigating the spread of ancestry using forward spatial population genetic simulations* is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of this dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I confirm that I have read and understood the Faculty of Mathematics Guidelines on Plagiarism and the University-wide Statement on Plagiarism.

Abstract

Modelling realistic dynamics of populations is an important aspect of population genetics. Of course, individuals in populations exist on spatial landscapes, and the way that they move through space and interact within it determines much of their dynamics. Forward genetic simulations allow explicit modelling of the complex ways in which populations evolve through time, and the efficiency and accessibility of SLiM, a particular framework for forward genetic simulations, means that biologists are increasingly able to use such realistic simulations as part of their research. The growing popularity of these methods inspired this thesis, in which forward genetic simulations are used to investigate how ancestry spreads through space. Models are developed to explore results from Olalde et al. 2018, which imply that approximately 90% of Britain's gene pool was replaced in the years 2450BC-1000BC. This is done with a view to examine some of the limitations on certain population dynamics that were at play in these years, such as migration rates and the ways that individuals moved through space. I find that the introduction of spatial effects means that ancestry is less able to spread through the population. I also find that to achieve a 90% turnover of Britain's gene pool, it's required that individuals can move a long way through space (on the order of 100km) in their lifetimes in the models. I further find that added dynamics, such as bottlenecks and the addition of a fitness advantage to certain genealogies, may allow ancestry to spread more readily through space.

Contents

Chapter 1

Introduction

In this chapter, some of the ideas and considerations necessary to start exploring spatial population genetics and the spread of ancestry are outlined.

1.1 Why is space important?

Classically, population genetics was done using models of panmixia (in particular, simple Wright-Fisher models), in which all individuals have an equal probability of mating with one another. Such models neglect the possible effects of the geography of the population. For example, where individuals live, how they tend to move across the landscape, and how they choose to interbreed across the landscape.

Spatial effects such as these are particularly relevant in models where migrants are entering a population since presumably they enter at specific regions and are not evenly distributed throughout the population. This is to say that spatial considerations are key in thinking about the spread of ancestry through a country or region.

1.2 What is meant by ancestry?

In trying to model ancestry, it's important to make the distinction between genealogical ancestry and genetic ancestry.

A person's genealogical ancestry is related to their pedigree. It is a function of the 'groups' that their parents, grandparents, etc. belonged to. For example, if an individual had 3 grandparents who belonged to group A and 1 grandparent who belonged to group B, one could say that their genealogical ancestry was three-quarters A and one-quarter B.

While genealogical ancestry can be thought of as a singular value for each individual in this way, genetic ancestry may vary across an individual's genome. During biological reproduction, one parent contributes different parts of their own maternal and paternal genomes. Only half of a parent's genome is passed on to their offspring, and the mechanisms behind which parts are passed on are further complicated by recombination (each chromosome is not just sampled randomly from the parent's maternal and paternal genomes). Even though an individual (diploid) receives 50% of their genome from their mother and 50% from their father, they may not receive 25% from each grandparent, and similarly going back many generations. These effects mean that an individual's genetic ancestors are just a small subset of their genealogical ancestors. In fact, going back in time over generations, the number of genealogical ancestors an individual has increases exponentially (as

2^k) whilst the expected number of genetic ancestors only grows linearly, going back many generations (Coop 2017).

Real-world genetic data is often not informative about either genetic ancestry or genealogical ancestry, so statements often claiming to be about genetic ancestry are truly about genetic similarity. Genetic similarity can only be related to genetic ancestry through assumptions about demography, which are difficult to test (Mathieson and Scally 2020), but Olalde et al. 2018 (see Section 1.3) uses qpAdm which examines patterns of shared genetic drift between populations to test the plausibility of certain admixture models. It then estimates admixture proportions using this information (Harney et al. 2021), and these admixture proportions are used as the effective (genetic) ancestry in this thesis.

The concept of ancestry spreading through space can now be made a little clearer through an example: individuals living in a particular region may have no genetic or genealogical ancestry that originated from a particular group A. If, in the following generations, the new individuals living in that region have genetic or genealogical ancestry originating from group A, ancestry from group A is said to have spread to that area of space.

1.3 Turnover of British ancestry

Here, I consider results from Olalde et al. 2018, which presents an analysis of 400 Neolithic, Copper Age, and Bronze Age Europeans. The paper combines archaeological evidence and aDNA data to draw conclusions regarding populations from previous millennia. A key conclusion of the study, which is the focus of this thesis, is the 90% turnover of the British gene pool within a few hundred years. The results of these types of studies, however, have even wider-ranging implications, both in the fields of archaeology and genomics (Armit and Reich 2021).

Olalde et al. 2018 uses qpAdm (Harney et al. 2021) to investigate the magnitude of population replacement in Britain during these periods, by modelling the genome-wide ancestry of the relevant individuals as a mixture of continental Beaker-associated samples and British Neolithic samples. In this thesis, the continental Beaker-associated ancestry is termed the *migrant* ancestry, whilst the British Neolithic ancestry is termed the *indigenous* ancestry.

Figure 3 in Olalde et al. 2018 shows the key results relevant to this thesis. British genomes dated between 4000BC and 2450BC (Neolithic) have 100% of this indigenous ancestry and no migrant ancestry. Moving through time, one sees that genomes dated between 2450BC and 1500BC (Copper Age and Early Bronze Age) contain some proportion of migrant ancestry – the proportions are extremely variable, however. Then, one sees that genomes dated between 1500BC and 800BC (Middle and Late Bronze Age) show much less variable ancestry proportions (see Section 3.2 for a discussion of some relevant statistics). The paper concludes that by the Middle Bronze Age (approximately 1500BC to 1000BC), there was a minimum of $90 \pm 2\%$ local population turnover, with no significant decrease observed in samples dated in the Late Bronze Age (1000BC to 800BC).

These results are consistent with the relevant migration into Britain starting at a point in time before 2450BC, and continuing into the Copper Age and Early Bronze Age. One can be confident that migration occurred in this period (or soon before it) since there are genomes dated in this period with 100% migrant ancestry. The decrease in variability in the migrant ancestry of genomes is then consistent with the migration rate dropping, perhaps all the way to zero, by the Middle Bronze Age (as a population with variable ancestry proportions among individuals is allowed to evolve with no migration over a longer time, ancestry proportions become more uniform).

Since the purpose of this thesis is to explore *spatial* population genetics, it's worth briefly noting some of the possible confounding spatial effects at play. Figure 1 in Olalde et al. 2018 shows the

original locations of the ancient genomes analysed. One sees that the genomes are certainly not evenly distributed (in a Euclidean sense) across Britain. For example, it's clear that there is a cluster of Beaker-associated genomes taken on the east coast and a larger cluster in the south of England. Therefore, the assumption that the genomes are representative of the whole British population, which is made to simplify the analysis in this thesis, may be rather inaccurate, particularly when these clusters appear at the coasts where migration may have a stronger effect (since migrants who enter Britain at a particular coast are likely to live close to that coast).

1.4 Which considerations are important?

An important parameter in discussing the movement of genetic or genealogical ancestry through space is the mean distance between parents and offspring (Bradburd and Peter L. Ralph 2019). Such a distance may be mediated by the movement that offspring move away from their parents in their lifetime or the distance that an individual is willing to travel to find a suitable mate, for example. It may be possible to measure such a distance for some species, using telemetry for example (Cayuela et al. 2018), but such methods may not be applicable to human populations in which many cultural aspects determine these emergent behaviours.

Here, and throughout this thesis, the word parameter is used in a loose sense to mean an aspect of the population model that one may choose to implement. Such parameters should ideally map to physical and biological reality.

1.5 Aims

In this thesis, I aim to model the forward evolution of populations in continuous space. I aim to use models to investigate how ancestry spreads through space, particularly in the context of results from Olalde et al. 2018 which suggest that roughly 90% of Britain's gene pool was replaced in hundreds of years. This will be done with a view to explore some limitations on certain parameters of models, namely migration rates, which are required to produce realistic behaviours. These inferred limitations may lead to conclusions about the real ways that the British population evolved in the Copper Age and Bronze Age.

Chapter 2

Methods

2.1 Software and hardware

Throughout this thesis, the central framework used for modelling is SLiM 3.0 (Haller and Messer 2019). SLiM is a simulation framework designed to run arbitrarily complex population genetic simulations. Models are written using the Eidos scripting language (Haller 2022), and the fact that SLiM is scriptable is what allows the flexibility to model essentially any relevant evolutionary dynamics.

Shell scripts are also used to run SLiM "recipes" multiple times, perhaps with different parameters or seeds, and to extract the relevant results from these runs.

R scripts and base graphics are used to analyse and visualise data produced by SLiM models.

The simulations throughout this thesis were run on a 2016 MacBook Pro with a 2 GHz Dual-Core Intel Core i5 processor and 8GB of RAM. The discussions of computational cost throughout this thesis are intended to be relative to this level of computing power.

2.2 Forward-in-time simulations

The models in this thesis run forward-in-time, as opposed to the commonly used coalescent framework, which runs backwards-in-time. Although the coalescent framework is often more useful, the behaviours I'm trying to investigate, involving spatial effects, are simply too complicated and emergent to apply coalescent theory.

In this thesis, I don't try to infer parameters or limitations on parameters directly from the relevant data. Instead, I try to simulate possible population dynamics and investigate whether they map well to the observed data from Olalde et al. 2018. Dynamics which map well to the data are more likely to be closer to the real-world population evolution that occurred in 2450BC-1000BC. This is especially true if relevant constraints (such as those discussed in Sections 3.2 and 3.6) are indeed strong.

2.3 Basic modelling principles

Two populations are modelled: one corresponding to the British population and one corresponding to the migrant source population. The dynamics of the source population aren't at all relevant – it only exists so that migrants can enter Britain.

Spatial interactions such as competition for resources and mate choices do not occur between these two populations, just within the British population.

In the simulations, a generation may be thought of as the time taken for offspring to be born, mature and reproduce. Generations are modelled as discrete irreducible units of time throughout this thesis.

2.4 Modelling ancestry

In the simple model of a single population with migration, the genealogical ancestry of indigenous members of the population (corresponding to British Neolithic ancestry in the model of Britain) can be initially defined as 0, and the genealogical ancestry of migrants can be initially defined as 1. Then, as the population evolves, the total genealogical ancestry of a child is just the mean of the genealogical ancestries of the two parents.

An individual's total genealogical ancestry (which I may simply refer to as their ancestry), some real number between 0 and 1, can then be thought of as the fraction of their ancestors that originally belonged to the migrant population.

2.4.1 Genealogical ancestry as a proxy for genetic ancestry

Even though the result of interest from Olalde et al. 2018 quantifies the replacement of genetic ancestry, this is computationally expensive to simulate many times in SLiM for large populations (see Appendix E). For this reason, the choice is taken to simulate the spread and replacement of genealogical ancestry.

The total genealogical ancestry of an individual acts as a proxy for the total genetic ancestry of the individual (where an individual's genetic ancestry is defined as the average origin of a base pair in their genome, with 1 corresponding to full migrant ancestry and 0 corresponding to full indigenous ancestry – see Appendix E for more details). Although the two may be different in practice, the total genetic ancestry can be thought of as being sampled from a probability distribution, with mean equal to the total genealogical ancestry. The expected value of an individual's genetic ancestry is equal to their genealogical ancestry simply because recombination is not biased towards either migrant or indigenous genomic regions.

Modelling genealogical ancestry is less computationally expensive than modelling genetic ancestry because it only requires the calculation and storage of one scalar per individual in the population, as opposed to the tracking of multiple regions on the genome for each individual in the population (see Appendix E).

Theoretical frameworks which more rigorously relate genetic ancestry to genealogy do exist (Buffalo, Mount, and Coop 2016), and future researchers may wish to apply these frameworks in the context of the use of genealogical ancestry as a proxy for genetic ancestry in this thesis.

2.5 Advantages and limitations of approach

Of course, much information is lost in using genealogical ancestry to represent genetic ancestry. Tracking genetic ancestry may be a better approach if a researcher is interested in particular regions of the genome and whether they are passed down through generations.

Despite this, the approach of modelling genealogical ancestry, as a proxy for genetic ancestry, is much more computationally tractable, especially if it's important to run many simulations.

Another disadvantage of the methodology is that many of the model parameters chosen may not map well to physical reality. For example, the mating radius and dispersal used are chosen rather arbitrarily (see Section 3.6).

There are also certain statistical limitations to extracting population genetic information from temporal data (Lynch and Ho 2020). To analyse simulations, I approximate the relevant data from Olalde et al. 2018 as binned temporal data (see Section 3.1 for an explanation of approximations), and this discretisation into two sampling times further limits the inferences one can make.

Simulations are also relatively computationally expensive, which means I'm not able to run many simulations and investigate the average behaviour. The stochasticity in this limited number of simulations may affect the results of this thesis (see Appendix F).

2.6 Reproducibility

To allow the reader to gain some insight into the models and results discussed in this thesis, relevant code is provided on the accompanying Github page. Provided are the SLiM "recipes" used, as well as some example R and Shell code used to generate the data and figures, with the sections that they correspond to in this thesis shown.

I encourage the reader to install SLiM themselves and to explore the models provided, perhaps changing some of the parameters themselves to investigate how this affects the results described throughout this thesis, or to test their own hypotheses. The provided recipes also allow for the reader to clarify any description of parameters that may be unclear to them in the explanations throughout this thesis, as well as allowing them to see how the theoretical models described translate to code.

Chapter 3

Investigating the replacement of Britain's gene pool

In this chapter, which comprises the bulk of this thesis, the mechanisms behind the turnover of British ancestry described in Section 1.3 are explored. I consider the years 2450BC-1000BC, in which the turnover happened and ancestry proportions began to stabilise across the population.

3.1 From abstract model to physical reality

In this section, I begin to consider the relevant physical quantities needed to model the replacement of ancestry on a 2D landscape.

As discussed in Section 1.3, results from Olalde et al. 2018 give two time periods of interest: a period between 2450BC and 1500BC, and a period between 1500BC and 1000BC. In the first period, which I term **P1**, ancestry proportions are variable, and I will assume that migration is happening at a constant rate throughout the whole period. In the second period, which I term **P2**, I'll assume that migration has stopped and ancestry is allowed to spread as the population continues to evolve. For both periods, for the analysis of results, I'll summarise the ancestry values of the population at the end of the relevant period.

To clarify, I discretise the data from Olalde et al. 2018 into two sampling times: the first at the end of **P1** (corresponding to roughly 1500BC, to represent the genomes dated in 2450BC-1500BC); and the second at the end of **P2** (corresponding to roughly 1000BC, to represent the genomes dated in 1500BC-1000BC). Naturally, this is a simplification, since genomes in the original data are dated throughout the periods (with uncertainties between 20 and 100 years), not just at the end. Although assuming that the data can be approximated by this discretisation means that resolution on the dating of genomes is lost, it allows for convenient analysis of the relatively small number of genomes and ancestry proportions in the data as distributions. It then allows for these distributions to be compared to the distributions of ancestry proportions obtained from a single simulation. I'll refer to this binning of the data into two sampling times, at the end of **P1** and the end of **P2**, as a discretisation of the data throughout this thesis.

In a Wright-Fisher model, a generation is the time taken for offspring to be born, mature and replace their parents. A generation can be estimated as 30 years. In a non-Wright-Fisher model (used in Sections 3.7, 3.8 and 3.9), a generation can also be estimated as 30 years, although it does not have the same biological meaning. Individuals can live for longer than one generation in a non-Wright-Fisher model and can reproduce (multiple times) over more than one generation.

This mapping of generations to years means that **P1** lasts 950 years or roughly 32 generations in simulations. It also means that **P2** lasts 500 years or roughly 17 generations in simulations. In the simulations, the beginning of **P1**, corresponding to the year 2450BC, will be indexed as generation 0.

Population size also needs to be considered. Brothwell 1972 used paleodemography to estimate that the population of Bronze Age Britain was 40,000, with a maximum likely variation of estimate of 20,000-100,000. Thus, for this thesis, a population size of roughly 40,000 is assumed. For simplicity, it's also assumed that the population size of Britain stayed roughly the same in the periods of interest (Copper Age and Early-Middle Bronze Age).

3.2 Some criteria for a successful simulation

Throughout this chapter, to probe whether populations are behaving realistically, I rely on many guiding emergent properties of the simulations and test whether these properties satisfy certain criteria. Many of these criteria, such as stable population size and roughly uniform population density, are discussed in later sections (at the beginning of Section 3.6, in particular). Here, a non-exhaustive set of guiding criteria based on Figure 3 and Table S9 in Olalde et al. 2018 is provided, noting again that the data are approximated by discretising into two sampling times, as explained in Section 3.1.

1. Individuals' average migrant ancestry should increase to roughly 90% during **P1**, representing the turnover of British ancestry in the Copper Age and Early Bronze Age. The distribution of ancestry proportions should then tighten during **P2**, representing the tightening of ancestry proportions in the Middle Bronze Age, as outlined in the coming criteria.

2. All 67 relevant samples from the Copper Age and Early Bronze Age in Olalde et al. 2018 have at least 59.7% migrant ancestry. This can be used to say that a good simulation will result in all individuals having at least 59.7% migrant ancestry at the end of **P1**.

3. Of the 67 relevant samples from the Copper and Bronze Age in Olalde et al. 2018, 17 of them have 100% migrant ancestry (where I count individuals calculated to have migrant mixture proportion greater than or equal to 1.0 as having 100% migrant ancestry). This can be used to say that a good simulation will result in a comparable proportion of individuals having 100% migrant ancestry at the end of **P1**.

4. Of the 25 relevant samples from the Middle Bronze Age in Olalde et al. 2018, all of them have at least 78.9% migrant ancestry. This can be used to say that a good simulation will result in all individuals having at least 78.9% migrant ancestry at the end of **P2**.

5. Of the 25 relevant samples from the Middle Bronze Age in Olalde et al. 2018, only 1 of them has 100% migrant ancestry. This can be used to say that a good simulation will result in a comparable proportion of individuals having 100% migrant ancestry at the end of **P2**.

Note that the ancestry proportions mentioned above, calculated using qpAdm in Olalde et al. 2018, are estimates with standard errors on the order of 3-10%. For simplicity, for the analysis in this thesis, they're taken as if they were perfectly accurate.

3.3 An analytic estimate

One can make some further assumptions to try and obtain an analytic estimate for the migration rate required to result in the 90% ancestry turnover. Assuming the migrant genome has the same

fitness as the indigenous genome, but migrants 'replace' non-migrants in some sense, one can imagine a minimum length of incoming migrant genome required for the ancestry turnover. Here, just **P1** (described in Section 3.1), in which migration is happening at a constant rate for 32 generations, is considered.

If it's assumed that all migrants survive and pass on their whole genome once, and this propagation continues throughout all 32 generations (it's assumed that the number of migrant base pairs that enters the population does not increase or decrease, migrant base pairs are just passed down through generations), then $40,000 \cdot 0.9 = 36,000$ whole genomes migrating in are required to achieve the ancestry turnover. This equates to approximately 1125 incoming migrants per generation or 37.5 migrants per year. Although this scale of migration rate seems reasonable, it's likely a lower bound. This is because:

- In reality, incoming migrants won't simply replace individuals with no migrant ancestry. More complex dynamics involving spatial competition and varying survival likelihoods will be at play. Also, it's been assumed that indigenous base pairs are replaced rather than individuals, which breaks down, especially when all individuals have at least some migrant ancestry.
- Migrants entering a small region are forced to compete with each other for resources, so are generally less likely to survive and pass on their ancestry than those far from the region of migration (who have more indigenous ancestry, at least when migration starts).
- In a spatial model, it will be much more difficult for migrant ancestry to spread to regions far from the region of migration since this requires individuals with migrant ancestry to travel a larger distance.

3.4 A minimal non-spatial model

To try to obtain a better intuition for the migration rates necessary to result in the ancestry turnover, I first use a simple neutral Wright-Fisher model (with 2-parent mating) and simulate in SLiM. A population size of 40,000 is used, and the only parameter to investigate is the migration rate per generation m. Full details of the simulations used and the results stated in this section can be found in Appendix B.

In this model, I first look for the migration rate at which the required replacement of ancestry by the end of **P1** occurs. The *critical migration rate* is the term I'll use throughout this thesis to describe the (minimum) migration rate required for the average ancestry to reach 90% by the end of **P1**. **P2** and the stabilisation of ancestry proportions will be considered later in this section.

I find that the critical migration rate in this model is approximately 0.07, which corresponds to a mean of 2800 migrants per generation or 93.3 migrants per year.

Taking this critical migration rate of 0.07 for 32 generations, I continue to evolve the population without migration for 17 further generations, to model **P2** in which ancestry proportions become less variable. I find, however, that in this model, the ancestry proportions become far too uniform by the end of **P2**. For example, in the data from Olalde et al. 2018 one sees a minimum of 78.9% migrant ancestry in the Middle Bronze Age, compared to a minimum of 89.97% migrant ancestry at the end of **P2** in a particular simulation of this model. The random mating in this model (i.e. the lack of spatial differentiation over large distances) causes ancestry to spread too quickly in comparison to the data from Olalde et al. 2018.

One might then hypothesise that this critical migration rate represents a further lower bound for the continuous spatial model, for the following reasons. One might expect migrant ancestry to be able to spread more quickly in a panmictic model than a continuous spatial model where

individuals are likely to mate with those close to them in space since one expects migrants to enter the population in some restricted region. This will mean that individuals being born are less and less likely to inherit this migrant ancestry the further they (and their parents) are from this restricted region. Furthermore, migration happens at the juvenile level in this model, with new migrant children not having to compete for resources with other members of the population, which remains at constant size. By contrast, in the continuous spatial model, new migrants have spatial competition with one another as well as other population members in their limited region of migration, making them less likely to pass on their ancestry over many generations.

3.5 Stepping stone model

As an intermediate step between the non-spatial models described above (Sections 3.3 and 3.4) and the continuous spatial models described below (Sections 3.7, 3.8 and 3.9), I implement a finite stepping stone model, again using SLiM. The stepping stone model works by modelling individuals in a finite number of sub-populations within Britain. Movement between these sub-populations is mediated by "migration rates" (not to be confused with the migration from the migrant population, which still occurs in this model). Within each sub-population there's panmixia, and offspring continue to be generated according to the Wright-Fisher model (as in Section 3.4).

In Section 3.6.1, I'll estimate that the length-to-width ratio of Britain as a rectangle is roughly 4:1, and I choose to model 16 sub-populations within Britain for this reason (16 is divisible by 4). To paint a clearer picture of the regime being modelled, a visualisation of the model during **P1** is shown in Figure 3.1. Here, I consider migrants to be entering Britain on the south coast (only into the two southernmost sub-populations), which I'll continue to do in the coming sections. As expected, one sees in the figure that migrant ancestry spreads more readily in regions closer to the sub-populations in which migrants enter the population.

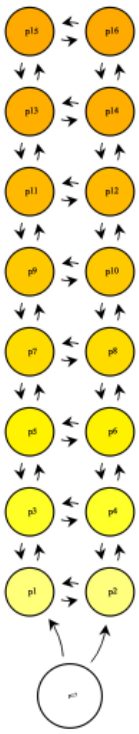


Figure 3.1: Visualisation of the stepping stone model. The very bottom population corresponds to the migrant population (whose dynamics aren't relevant) and the other 16 sub-populations represent regions of Britain. The upper sub-populations correspond to the north of Britain, whilst the lower ones correspond to the south. The sub-populations are coloured according to the average genealogical ancestry of their individuals, with white corresponding to full migrant ancestry and dark red corresponding to full indigenous ancestry.

The purpose of this stepping stone model, having already gained some insights into the possible migrations rates necessary in Section 3.4, is to gain some intuition on the parameter I term *connectedness*. In this model, the connectedness is simply the "migration rate" between adjacent sub-populations within Britain. In a more intuitive and less rigorous sense, it relates to the distance that individuals move through the population in a generation (see the end of this section for a more detailed discussion about how the parameter maps to a continuous spatial model). Clearly, this is a defining consideration in a spatial model of ancestry spread.

I begin to model with the migration rate (into Britain, from the migrant population) estimated in Section 3.5, scaled to 0.56 for both southern sub-populations so that the total mean number of migrants into Britain per generation is still 2800. I run these simulations applying a maximum connectedness (given the migration rate) of 0.22. This maximum connectedness means that for the southernmost British sub-populations, 100% of parents for each generation are selected from outside that sub-population. For all other sub-populations apart from the two northernmost, 66% of parents, on average, are selected from outside that sub-population each generation. I find that even with this maximum connectedness, the average ancestry in Britain at the end of **P1** is approximately three times smaller than that which is required to map to the data in Olalde et al. 2018!

This lack of ability for ancestry to spread even though there are similar numbers of incoming migrants is due to the introduction of spatiality to the model. In the two southernmost sub-populations, for example, although migrants are coming in from the south, there are also individuals

coming in from neighbouring sub-populations to the north, which counter the incoming migrant ancestry. Furthermore, it's now more difficult for ancestry from the south to spread to the north. For example, no migrant ancestry at all can reach the northernmost sub-populations until generation 8 at the earliest (since an eight-by-two grid is used in the model and only adjacent sub-populations are connected).

To explore how the spread of ancestry varies in different spatial regions, and how this is affected by connectedness, the average ancestry of individuals in the south and in the whole population of Britain is plotted in Figure 3.2 for different connectedness values. The figure demonstrates a trade-off when connectedness is increased: ancestry can spread more efficiently through the population, but it's not able to enter the population as readily (since individuals moving into the southernmost sub-populations may force out individuals with more migrant ancestry from previous generations).

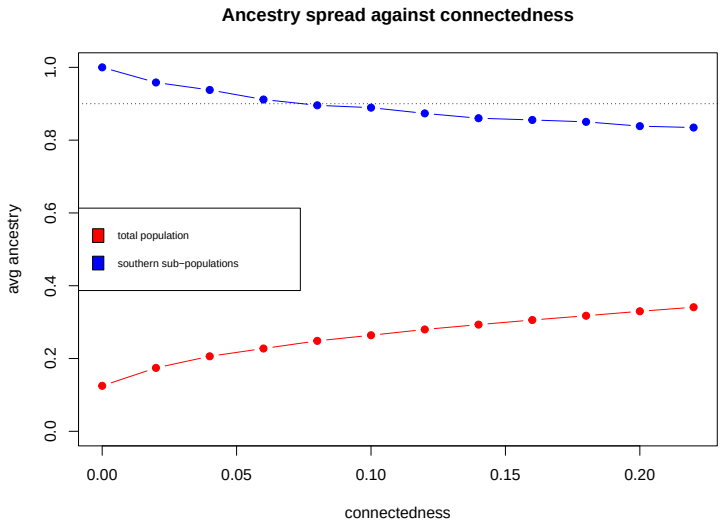


Figure 3.2: Average ancestry at the end of **P1** against connectedness in stepping stone model. The average ancestry of the two southernmost British sub-populations, and the average ancestry of the whole British population, are plotted. A constant migration rate of 0.56 into each of the two southernmost British sub-populations is used as above and connectendness is varied.

To create a stepping stone-like model that would map more closely to the data from Olalde et al. 2018, it's likely that a higher migration rate, as well as a more realistic model of how individuals move throughout space, would be needed. This could be done by connecting more distant sub-populations with smaller "migration rates", for example. More complex dynamics are more explicitly modelled in the continuous spatial models below, in which the connectedness parameter here can map to all of the following:

- The distance that individuals move in their lifetime.
- The distance that offspring move away from their parents between being born and reproducing.
- The mating radius and function of individuals (i.e. the function that specifies the likelihood that an individual will mate with another individual based on the distance between the two).

It may also be necessary for migrant ancestry to move from south to north in some way that's not just specified by random diffusion. This will be further discussed in later sections.

Since there are two columns of sub-populations in this model of Britain, the effective distance between sub-populations is on the order of 0.5 units, or roughly 117km (see Section 3.6.1). Qualitatively, this may give some rough idea of the scales that the three parameters listed above should span in the continuous spatial model.

3.6 Continuous spatial model considerations and parameters

Before beginning to think about the migration rates that allow for the required spread of ancestry through continuous space, it's important to deeply consider the parameters of the model such that realistic behaviour is observed. In this section, I discuss in detail some of the parameter choices taken to model populations in continuous space, aiming to build a model that satisfies the following criteria (in addition to those specified in Section 3.2):

1. Individuals tend to live for a reasonable number of years (somewhere between 0 and 60, say).
2. Individuals produce a reasonable number of offspring each generation (one individual shouldn't be producing 30 offspring in a single generation, for example).
3. Population size should be roughly stable at around 40,000.
4. Migrant ancestry should be able to spread through the population reasonably quickly, to make the required 90% ancestry turnover feasible.
5. The model should be computationally tractable.

The approach taken is to build a model with as few parameters as possible, producing realistic behaviour, so as not to over-complicate the emergent dynamics. Non-Wright-Fisher models, which allow for individuals to (more realistically) live for longer than one generation, will be used. This means that the population size described above will be emergent rather than imposed.

3.6.1 "Rectangular Britain"

Using Google Maps (see Figure 3.3), I estimate that Britain is approximately 898km in height (roughly from Portsmouth to Inshore, Lairg). Dividing the known land area of Britain, $209,331\text{km}^2$, by the height gives an average width of approximately 233km. This means Britain can be roughly modelled as a rectangle with side lengths 1 and $898/233 \approx 3.9$, with one unit in the rectangle corresponding to 233km (see Figure 3.7).

Figure 3.3: Approximate height of Britain using Google Maps.

Approximating the land area as a rectangle in this way allows the boundary effects that naturally arise, such as those described in Section 3.6.5, to be more easily dealt with.

It's worth quickly noting that Britain in the Copper and Bronze Ages didn't include Ireland or Northern Ireland. In this thesis, Britain is just used to mean the geographic region that would today be called Great Britain (England, Scotland and Wales, excluding their component islands).

3.6.2 Carrying capacity and spatial interactions

The carrying capacity of the environment is set to 40,000. This should be thought of as a measure of the environment's ability to support life through resources such as food and water. The carrying capacity is set as constant throughout Britain, and I choose not to explore this parameter further.

The parameter S is set to 0.02 units. S controls the scale of the spatial competition (as explained below) as well as the scale of mating choice (as explained in Section 3.6.4). The reason I choose S to control both the spatial competition and the scale of mating choice is simply to reduce the number of parameters.

S is chosen to be large enough so that all focal individuals in the population (possibly except for a small number of outliers each generation – see Figure C.1) have at least one individual within their mating radius and hence have the possibility of mating, even though this may not lead to the production of any offspring (see Section 3.6.3). Note that each generation, SLiM goes through all individuals in the population synchronously so that they all act as focal individuals for mate choice and spatial competition. Each focal individual chooses at most one suitable mate. They do this by randomly choosing one of their three nearest neighbours (in a Euclidean sense) with equal probability. The neighbours must be within a maximum distance S of the focal individual, and if the focal individual doesn't have any neighbours within this distance they simply do not reproduce in that generation.

I set the spatial competition function, which I term $g(x)$, as a normal distribution (effectively centred at each focal individual). The distribution is scaled to have maximum size 3.7, and standard deviation S, so one can write

$$g(x) = 3.7e^{-\frac{x^2}{2S^2}} \tag{3.1}$$

for a Euclidean distance x between individuals. The distribution is cut off at a distance $S \times 3$ so that no interaction calculations need to be made for individuals further than $S \times 3$ units from each focal individual. This cutoff (along with the smaller cutoff S used for spatial mate choice) makes the model much more computationally tractable since pairwise interaction calculations mean the simulations will run in $O(S^2)$ time and space complexity.

Note that the parameters of the model are set to maintain a population size of roughly 40,000. This balance involves a trade-off between the carrying capacity, the typical strength of the spatial competition (which relates to the maximum size and standard deviation of the normal distribution $g(x)$ and to the denominator of Equation 3.3 described at the end of Section 3.6.6) and the mean number of offspring born in a reproductive event. Whilst the carrying capacity and spatial competition strength are relatively arbitrary parameters, the mean number of children per reproductive event is something that maps more obviously to biological reality (see Section 3.6.3). There are many other possible sets of parameter choices which also maintain a population size of approximately 40,000, and future researchers may wish to investigate whether using different sets influences the dynamics of the model.

3.6.3 Offspring generation

Once a focal individual has chosen a mate (as described in Section 3.6.2), the number of offspring for that reproductive event is sampled from a Poisson distribution with mean 1.

Note that individuals in these simulations are hermaphroditic, meaning that every individual has the opportunity to be chosen as a suitable mate by every other individual.

The fact that (almost) every focal individual has the opportunity to reproduce (see Figure C.1) means that individuals may reproduce more than once in a generation (when they choose a mate and when they are chosen as a mate).

3.6.4 Dispersal and reproductive radius

As mentioned in Section 3.6.2, the reproductive radius is controlled by a parameter S. In discussing the spread of ancestry through space, it's clear that ancestry will be able to spread much more quickly if individuals can mate with others further away from them.

Another related parameter is the function which controls where an individual is born in relation to its focal parent. I choose to sample each of the offspring's coordinates from a normal distribution, with standard deviation 0.8 units, centred at the focal parent. I term this standard deviation the *dispersal*.

The dispersal is a key parameter influencing the spread of ancestry through space since a larger dispersal would allow new offspring with a large amount of migrant ancestry to be placed into areas which currently don't contain a high density of migrant ancestry. This relates to the reproductive radius since this specific scenario, which mediates the spread of ancestry in a significant way, can be equivalently achieved by a focal individual in a region with a low density of migrant ancestry mating with an individual far away (i.e. using a large mating radius) with a large amount of migrant ancestry (presumably in an area with a high density of migrant ancestry). The offspring can then be born near the parent which originally had a small amount of migrant ancestry (due to a low dispersal), thus spreading migrant ancestry. In this scenario, of course, I'm assuming a mating regime different to that described in Section 3.6.2 in which a focal individual doesn't necessarily choose one of its three nearest neighbours to mate with.

For the remainder of this thesis, I choose to run simulations with small S and large dispersal. This is because it's much computationally cheaper to run these models (models have time and space complexity $O(S^2)$, but increasing dispersal does not change time or space complexity) and similar behaviour is obtained.

Since the width of "rectangular Britain" is only 1 unit, the dispersal of 0.8 units means that many new offspring may frequently be placed outside of the British region, unless this is controlled for. As suggested in Haller and Messer 2022, I use reprising boundary conditions to offset this effect. This means that if a new offspring is set to be placed outside of the rectangular region, SLiM simply rejects this location and continues sampling from the normal distribution until a coordinate within the desired region is obtained. Using these boundary conditions makes edge effects less prevalent than if one was using stopping boundary conditions, for example, in which offspring coordinates chosen outside the boundary are moved to their nearest point inside the boundary.

This dispersal of 0.8 units corresponds to approximately 186km, whilst an S of 0.02 corresponds to a distance of approximately 4.7km. Although this dispersal of 186km may seem unrealistically large, in practice it corresponds to movement over a time scale of 30 years. Also, the reality of the 4.7km mating radius is that individuals still tend to mate with individuals much closer than 4.7km away from them since individuals are set to mate with one of their three nearest neighbours (see Section 3.6.2). A more realistic model may use a mating function with a larger mating radius in which individuals don't only mate with their three nearest neighbours, together

with a smaller dispersal. Note that another related parameter is the function controlling the distance that individuals move in their lifetime. For the remainder of this thesis, I choose not to explicitly set individuals to have any movement in their lifetimes, outside the movement in their first generation controlled by the dispersal. This is done to make computation more efficient.

I found in Section 3.5 that connectedness would need to be at least on the order of 0.5 units. Setting dispersal to 0.8 units thus makes sense as a parameter choice.

3.6.5 Further edge effects

One important step in writing a physically realistic continuous spatial model is to consider what happens at the boundaries. One might expect individuals at edges and corners to have artificially high fitness values under the model so far since beyond the edges of the model there are no individuals to compete with, meaning that individuals near the edges are naturally less affected by spatial competition (though this phenomenon is damped by the fact that individuals at these edges don't live for long – see Section 3.6.7). This would cause clusters of higher population density to form at the edges and corners. However, Figure 3.4 shows an almost opposite effect. The figure shows the local population density over a grid, to gain insight into how the population is distributed across the landscape in the simulations. One sees in the figure that towards the south and north of Britain, individuals are more sparse.

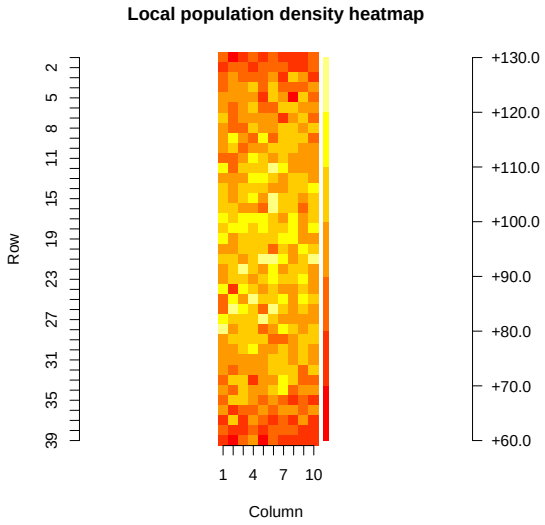


Figure 3.4: Local population density heat map after a 10-generation burn-in period on a grid with squares length 0.1 unit. Dark red corresponds to low population density and yellow corresponds to high population density, as shown in the key to the right of the heat map. The key shows the number of individuals in a particular square unit corresponding to that colour. The rectangular region should be thought of as a representation of Britain, with the bottom corresponding to the south and the top corresponding to the north, as in Figure 3.7. The row should be thought of as a measure of distance from north to south, and the column as a measure of distance from west to east.

One possible explanation for this phenomenon is that the dispersal is large. This can be explained using an example: if a focal individual near the southern border is producing offspring, SLiM samples from the normal distribution until a new point within the boundaries is found (see Section 3.6.4).

Since the standard deviation of the normal distribution is so large, it's more likely that a point further north will be ultimately chosen, since points further south chosen from the distribution are likely to lie outside the southern border and be rejected.

To offset this non-uniform population density, I introduce a slight fitness advantage to individuals in the northern- and southernmost 0.4 units (93.2km) of the rectangle (since these are roughly the regions that are clearly affected in Figure 3.4). I do this by multiplying the fitness scaling by 1.15 if the individuals lie in these northern or southern regions (as in Equation 3.2). This leads to the more uniform population density visualised in Figure 3.5, in which there's no clear clustering towards or away from the edges (compared to Figure 3.4).

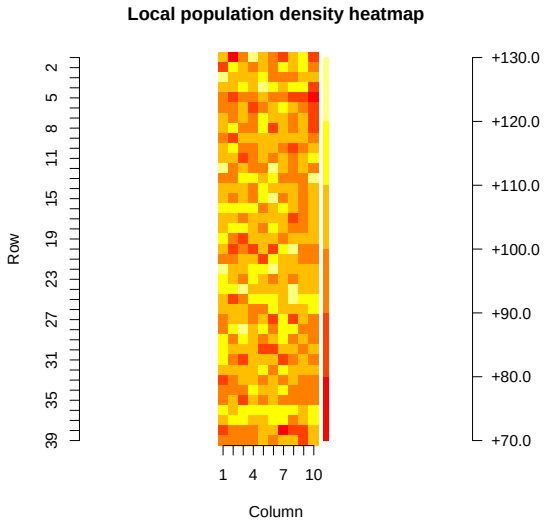


Figure 3.5: Local population density heat map after applying fitness advantage to northern- and southern- most 0.4 units, for comparison with Figure 3.4. Note that the colour scheme in this figure, shown in the key to the right of the heat map, varies slightly from the scheme in Figure 3.4, simply because there are different absolute values for population density in this figure.

Although this scaling is not very systematic and doesn't have an obvious mapping to physical reality (except if resources were more plentiful in the north and south in reality, for example), I choose to use it for the models for its simplicity.

Note that removing or adding this new fitness advantage affects the emergent population size, though the other parameters are set to account for the added fitness advantage.

3.6.6 Full fitness calculation

The fitness calculation determines the probability that an individual will survive to the next generation. For each individual, their fitness is calculated and a random uniform number between 0 and 1 is drawn. If the draw is less than the fitness, the individual survives and gets the opportunity to reproduce in the next generation, otherwise, they die. Note that a fitness greater than or equal to 1 means that the individual is guaranteed to survive that generation. However, an individual can only survive if they're less than 2 generations old (see Section 3.6.7). The fitness calculation should take into account the spatial competition, the expected lifetime and also the fitness advantage discussed in Section 3.6.5. To this end, I use the following fitness calculation, similar to that suggested in

Battey, Peter L Ralph, and Kern 2020:

$$\frac{\eta_i}{1 + n_i/(K(1+L))} \tag{3.2}$$

where η_i is the fitness advantage described in Section 3.6.5 and is given by:

$$\eta_i = \begin{cases} 1.15 & \text{if individual } i \text{ is in upper or lower 0.4 units} \\ 1 & \text{otherwise} \end{cases}$$

In Equation 3.2, n_i is a measure of the spatial competition from all neighbours within a radius $3S$ of individual i (see Section 3.6.2 and Equation 3.3), K is the carrying capacity (see Section 3.6.2) and L is related to the expected lifetime of individuals (see Section 3.6.7), set to 1.8 generations, or 54 years. In the formulation set out in Battey, Peter L Ralph, and Kern 2020, L is the mean lifetime, but other nuances of the model (such as the age cutoff imposed in Section 3.6.7) mean that the resulting mean lifetime of individuals is less than L (see Figure 3.6).

Specifically, n_i is calculated as

$$n_i = \frac{[\sum_j g(d_{ij})] + 1}{2\pi S^2} \tag{3.3}$$

where d_{ij} is the Euclidean distance between individuals i and j, $g(x)$ is the spatial competition function described by Equation 3.1, and the sum is taken over neighbours j within a radius $3S$ of individual i, not including i (Battey, Peter L Ralph, and Kern 2020). Equation 3.3 is related to (but does not map exactly to) the logic laid out in Section 16.10 of Haller and Messer 2022. The added constant 1 in the numerator of Equation 3.3 is somewhat arbitrary (it's between 0 and the maximum size of $g(x)$), and it corresponds to a contribution of the focal individual to its own spatial competition penalty. The division by $2\pi S^2$ is also somewhat arbitrary, though the factor of πS^2 is proportional to the area covered by the spatial competition, and the rest of the parameters of the model are chosen to balance with the pre-factor of 2 to result in the desired (roughly) stable population size of $\sim 40,000$ and other desired behaviours (see Section 3.6.2).

3.6.7 Further dealing with long lifetimes

To further control how long individuals live, I choose to add a cutoff age beyond which individuals cannot live. The reason that this choice is taken is that without this cutoff age (see Figure C.2), one sees that although the average lifetime control of the fitness scaling in Equation 3.2 helps to stop individuals from living beyond approximately 10 generations, there are still some individuals who live beyond 3 generations (corresponding to 90 years), and a small number who live to 8 generations (corresponding to 240 years)!

To control for these unrealistic dynamics, I impose a cutoff of 2 generations (corresponding to 60 years), so that individuals who live to 2 generations are killed off before they have the opportunity to reproduce in the next generation. Although there are other ways to control for these behaviours, such as limiting the fitness of individuals to some number less than one (Battey, Peter L Ralph, and Kern 2020), I choose to work with this cutoff age since one can imagine roughly similar cutoff ages in real human populations. To see how this cutoff affects the ages of individuals, the ages of dying in a simulation are plotted in Figure 3.6 after imposing the cutoff. One sees that the majority of individuals survive for either 1 or 2 generations.

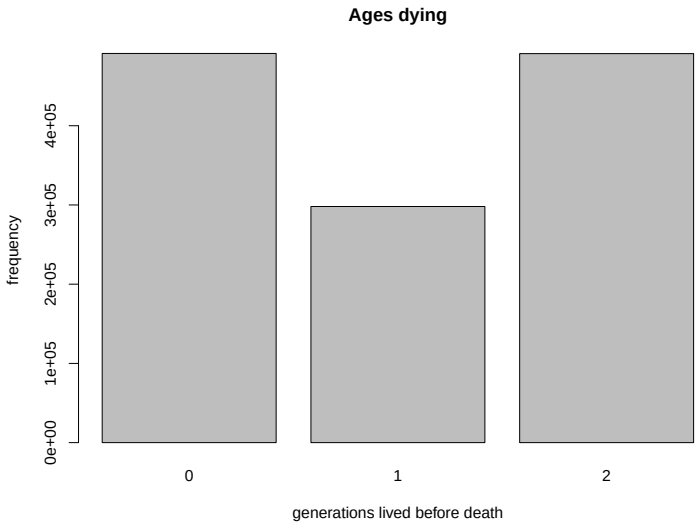


Figure 3.6: Ages that individuals die when a lifetime cutoff is imposed, in a particular simulation run over **P1**, for comparison with Figure C.2. Migration is again set to 1 individual per generation so as not to affect the evolutionary dynamics. The x-axis shows the age of death of individuals (in generations), and the y-axis shows the frequency of occurrences of individuals dying after that many generations.

Interestingly, many individuals in the simulations don't survive for any generations, which effectively means that they don't get the opportunity to reproduce in their lifetimes.

3.6.8 Migration

In the models, I choose to have migration into Britain from the south, which could correspond to migration from Europe. Future researchers may wish to investigate what happens there are multiple or more complex incoming migration regions.

I choose to set a rectangular migration region in the south of Britain, within which migrant individuals spawn, uniformly distributed throughout the region. This is to say that the x and y coordinates of new migrants are chosen according to random uniform distributions within the rectangle. The size of this rectangular migration region is discussed in Section 3.7.

A visualisation of this description is shown in Figure 3.7, to make clear what the migration region is.

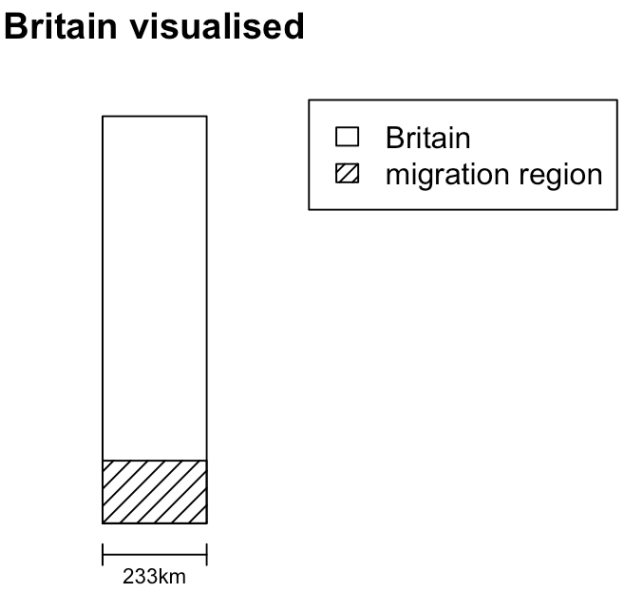


Figure 3.7: Migration region visualised. The rectangular region representing Britain should be thought of as a geographic approximation, with the bottom of the region corresponding to the south of Britain, the top corresponding to the north, the left corresponding to the west and the right corresponding to the east.

3.7 First continuous spatial model

In this section, I investigate the model described by the parameters set out in Section 3.6. In later sections, I'll explicitly add other mechanisms which could help ancestry spread more quickly. I continue to simulate using SLiM.

I start by investigating the relationship between the migration rate and the average ancestry of individuals in the population at the end of **P1**, as in Sections 3.3, 3.4 and 3.5. In this case (as well as in Sections 3.8 and 3.9), in order to get an accurate measure of the ancestry that will remain in the population, I take the measure of this average ancestry 1 generation after the end of **P1**, to give recent migrants living in areas of very high spatial competition the opportunity to die off. Of course, during **P1**, population size increases according to the number of migrants per generation. Since individuals do not live for more than 2 generations, the population size increases by roughly 2 times the number of migrants per generation.

I also investigate the effect that altering the size of the migration region has on the spread of ancestry. As explained in Section 3.6.8, I model using a rectangular region in the south of Britain. The height of this rectangular region (i.e. the distance from the south coast to the northernmost point in the migration region) will clearly affect how quickly ancestry can spread through the population – one expects that a larger height will allow ancestry to spread through the population more quickly for a given migration rate since it will allow more migrant ancestry into regions where it's more difficult for ancestry to spread naturally (further north) upon migration. Furthermore, a larger migration region for a given number of migrants results in less spatial competition among the migrants, meaning that they're more likely to survive and pass on their ancestry. However, increasing the height of the region too much is not in the spirit of the aims of this thesis – if the height of the region were equal to the height of Britain, it wouldn't be as necessary to have a *spatial*

model (this migration scheme also wouldn't be physically realistic).

To investigate the migration rates that are needed to achieve the 90% turnover of British ancestry seen in Olalde et al. 2018, as well as explore how the size of the migration region visualised in Figure 3.7 affects the spread of ancestry, the average ancestry over the whole population after **P1** against the number of migrants per generation for three different heights of migration region is plotted in Figure 3.8. One expects the average ancestry over the population throughout **P2** to remain similar to its value at the end of **P1**, since no additional migrants are added. The three schemes (labelled 0.4, 0.6 and 0.8) correspond to the heights of the migration region (in units, where 1 unit is the width of Britain and 3.9 units is the height of Britain).

One sees in Figure 3.8 that the critical migration rate (required to achieve 90% average ancestry) varies according to the size of the migration region. As expected, larger migration regions result in a smaller critical migration rate. However, the effect of increasing the size of the migration region on the spread of ancestry is smaller than one might expect. One sees that with 9000 migrants per generation, for example, increasing the height of the migration region by 0.2 units (roughly 47km) corresponds to roughly a 0.7% (absolute) increase in the average ancestry after **P1**. Another result is that the rate of increase of average ancestry (with number of migrants) decreases as the number of migrants grows. Again, this is likely because having more migrants in a limited space causes more spatial competition, meaning individuals are less likely to survive and pass on their ancestry. This is to say that the effectiveness of increased migration on the spread of ancestry has its limits due to spatial effects.

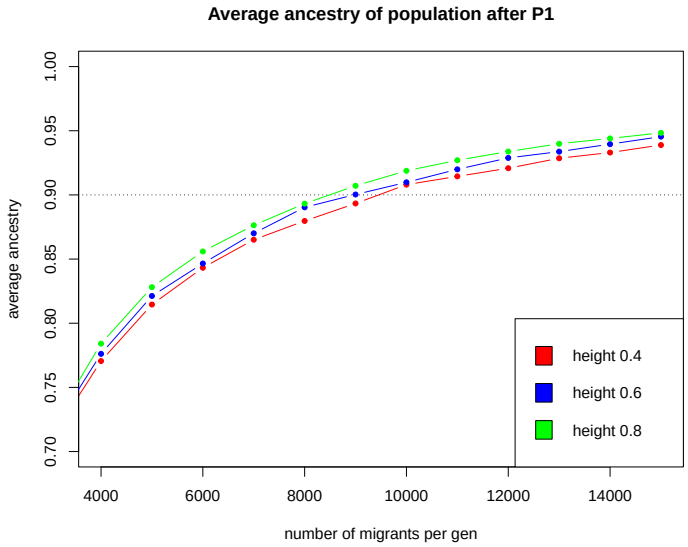


Figure 3.8: Average ancestry of individuals after **P1**. The coloured lines correspond to different heights of migration region. The y-axis gives the average ancestry over the whole population, and the x-axis gives the number of migrants per generation.

For the remainder of this thesis, I assume the largest of the three migration region sizes tested above, 0.8, since it marginally allows for the most modest (i.e. smallest) critical migration rate. Using linear interpolation on the results seen in Figure 3.8, I estimate that this critical migration rate is approximately 8490 migrants per generation or 283 migrants per year.

I now investigate the spread of ancestry in both **P1** and **P2**. To do this, I consider the average ancestry in 6 non-overlapping rectangular bands equal in area, where band 1 is the southernmost, band 2 is immediately north of band 1, and so on. A visualisation of these bands is shown in Figure

3.9. Note that in the model, the migration region covers the whole of band 1 and approximately 23% of band 2. Note also that the bands aren't physically or biologically relevant, they're simply regions labelled for the purpose of investigation.

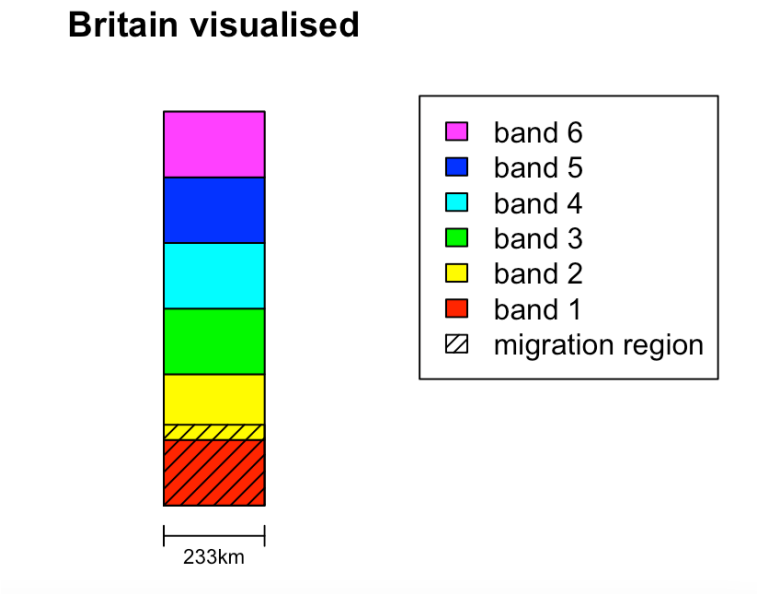


Figure 3.9: Bands visualised, for comparison with Figure 3.7. The 6 bands are non-overlapping and are equal in size.

To examine how well ancestry spreads through different regions in space, the average ancestry of individuals in these 6 bands over **P1** and **P2** is plotted in Figure 3.10.

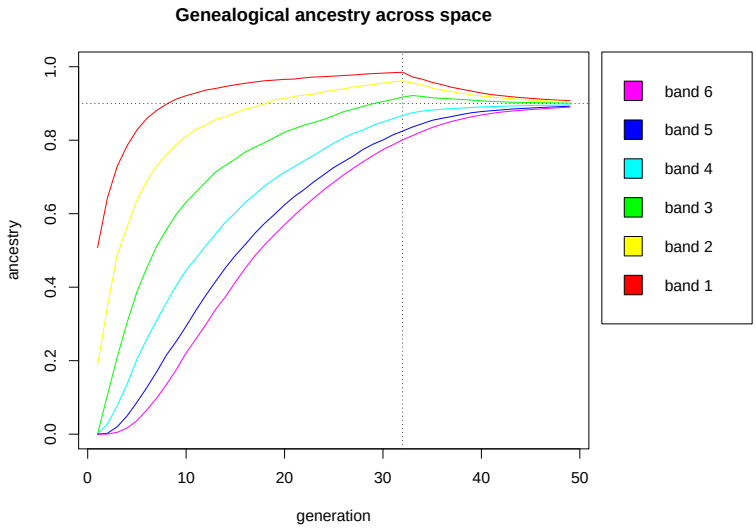


Figure 3.10: Spread of ancestry over **P1** and **P2** in the 6 bands visualised in Figure 3.9. The x-axis corresponds to the time since the beginning of **P1** in generations, and the y-axis corresponds to the average ancestry of individuals in each particular band. A dotted horizontal line showing 90% ancestry and a dotted vertical line showing the end of **P1** and the beginning of **P2** are included.

One sees in Figure 3.10 that the way that ancestry spreads varies across space in this model. One sees that the three northern bands have average ancestry greater than this critical 90% (which is approximately the average ancestry of the whole population) at the end of **P1** and the three southern bands have average ancestry smaller than 90% at the end of **P1**. Interestingly, after **P1** ends, the average ancestry in the three northern bands continues to increase, whilst the average ancestry in the three southern bands begins to decrease, as ancestry diffuses from south to north. As the population continues to evolve, one sees a convergence of the ancestry bands towards the average ancestry of the whole population.

During **P2**, one expects the distribution of ancestry proportions to tighten. To explore how this tightening maps to the tightening observed in Olalde et al. 2018, the distribution of ancestry proportions of the whole population throughout **P2** is plotted as box plots in Figure 3.11. As expected, one sees that the distribution tightens over time. However, further investigation is necessary to explore how this tightening maps to the data in Olalde et al. 2018.

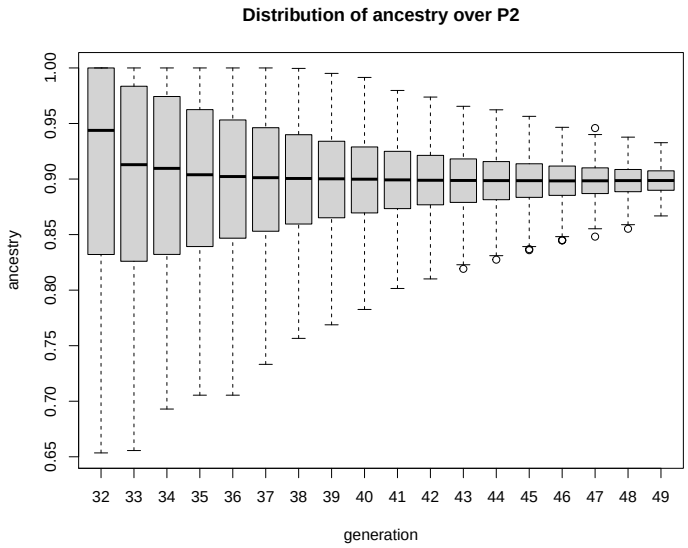


Figure 3.11: Distribution of ancestry proportions over **P2** as box plots. The x-axis shows the generation since the start of **P1**, and the y-axis shows the ancestry proportions of individuals.

Recall from Section 3.2 that the discretisation of the data from Olalde et al. 2018 (explained in Section 3.1) indicates that for a good simulation, at the end of **P1**, ancestry proportions should vary from 59.7% to 100% and at the end of **P2**, ancestry proportions should vary from 78.9% to 100%. To see how well the results of this simulation satisfy these criteria, the distributions of ancestry proportions at the end of **P1** and **P2** are plotted in Figure 3.12 (to compliment Figure 3.11). One sees that at the end of **P1**, many individuals span a wide range of ancestry proportions, and there's a large peak for individuals with 100% migrant ancestry. At the end of **P2**, ancestry proportions have settled into a more compact distribution, with no individuals having below 78.9% migrant ancestry.

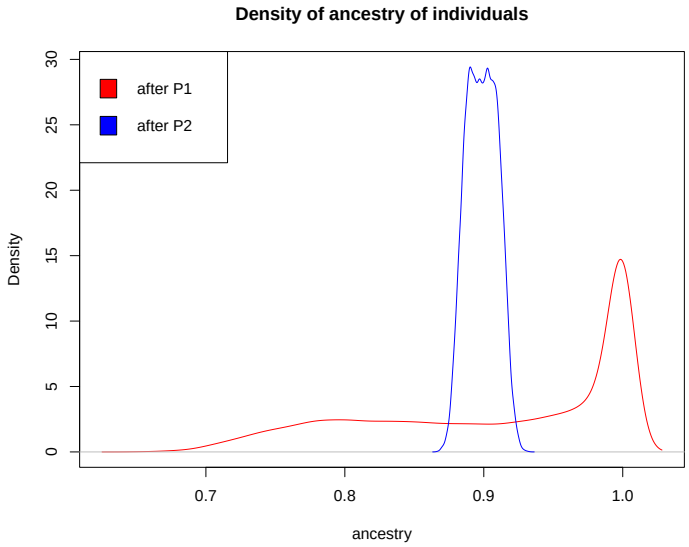


Figure 3.12: Distributions of ancestry values at the end of **P1** and the end of **P2**, as kernel density plots. The x-axis shows the ancestry proportions of individuals and the y-axis estimates the relative frequency with which that ancestry value occurs.

3.7.1 Likelihood of sampling results from this model

Here, I investigate how results from the particular simulation used to create Figures 3.10, 3.11 and 3.12 map to data from Olalde et al. 2018. The analysis relates to the criteria laid out in Section 3.2.

In this particular simulation, 100% of individuals have at least 59.7% migrant ancestry at the end of **P1**. This means that if the true distribution of ancestry values were what one sees in this simulation, it would map to the discretisation of the 67 relevant samples from Olalde et al. 2018 perfectly in this sense. In fact, the minimum migrant ancestry of an individual after **P1** in this particular simulation is 65.3%, which could fall within the standard error of the minimum 59.7% estimated by qpAdm in Olalde et al. 2018.

In this particular simulation, 27.5% of individuals have 100% migrant ancestry after the first period. This means that if the true distribution of ancestry values were what one sees in this simulation, the expected number of genomes with 100% migrant ancestry in a uniform sample of 67 individuals would be 18.4, which is slightly greater than the 17 that seen in the discretisation of the data. The probability that 17 or more of 67 uniformly sampled individuals (from the simulation) have 100% migrant ancestry is approximately 69.3% (See Appendix D for full details of calculation).

One also sees in this particular simulation that 100% of individuals have at least 78.9% migrant ancestry at the end of **P2**. This means that if the true distribution of ancestry values were what one sees in this simulation, it would map perfectly to the discretisation of the 25 relevant samples from Olalde et al. 2018 in this sense. In fact, the minimum migrant ancestry of an individual after **P2** in this particular simulation is 86.7%, which again could fall within the standard error of the minimum 78.9%

Finally, in this particular simulation, no individuals have 100% migrant ancestry at the end of **P2**. This means that if the true distribution of ancestry values were what one sees in this simulation, it would map to the discretisation of the 25 relevant samples from Olalde et al. 2018 perfectly in this sense.

Each of these likelihoods is entirely plausible, and none come close to being significant enough to reject the null hypothesis that the simulation maps to the data. If anything, the ancestry proportions across individuals are too tight in the simulation, compared to the data in Olalde et al. 2018. This may highlight the need to search for a parameter choice in the simulation which leads to a large turnover of average ancestry without too much stabilisation of ancestry proportions after the turnover happens. However, the discretisation into two sampling times is likely at least partially responsible for this apparent over-tightening of ancestry proportions since genomes dated earlier in the original data can likely account for many individuals with smaller migrant ancestry proportions.

In reality, the binary model of fixed migration happening in **P1** and then immediately stopping through **P2** is a simplification. Furthermore, the samples used in Olalde et al. 2018 are not completely independent, and they're certainly not uniformly distributed through Britain, as I've assumed in the analysis. I've already mentioned some of the possible confounding spatial effects in Section 1.3 – one could investigate these spatial effects further by sampling genomes near the coast (in the simulation) and looking into how well these samples map to the original data.

Ideally, one should run multiple simulations and take averages of the statistics calculated above. However, running these simulations is a computationally expensive task, and the smoothness of the distributions (those in Figure 3.12) indicates that stochasticity does not have a significant impact on results (at least in the final generations of the evolution, though it's possible that the probabilistic diffusion of migrants in early generations could have a more significant impact on results – see Appendix F). For these reasons, I choose to take these results without running multiple simulations, and I summarise similar statistics based on singular simulations in Sections 3.8 and 3.9.

3.8 Second continuous spatial model: population bottleneck

In this section and Section 3.9, I move on to talk about some possible mechanisms which would allow for the required spread of ancestry, but with reduced migration rates. I model these mechanisms explicitly.

Such mechanisms might be important if other data, which could be archaeological or aDNA data, for example, were found which suggested that migration rates were smaller than those inferred in Section 3.7.

One may also want to specifically investigate the mechanism of a population bottleneck, as is done in this section, if further genomic evidence were found that pointed to a loss in genetic diversity (similarly to Franks, Pratt, and Tsutsui 2011, for example). Such genomic evidence may point to the scale of the bottleneck event, which may influence some parameter choices of the model described below.

It's impossible to explicitly simulate the whole space of possible mechanisms at play in this section and Section 3.9. For this reason, these sections are based on many additional assumptions, which will be highlighted.

In this section, I consider population bottlenecks and the effects that these could have on the migration rates required to reach the desired ancestry spread. These events correspond to environmental changes, such as earthquakes, disease or famine, which cause a sharp decrease in the population size. Generally, population bottlenecks cause a loss in a population's genetic diversity – this could be useful in allowing ancestry to spread if more indigenous base pairs are lost due to the bottleneck than migrant base pairs, for example. If the population size is reduced at points in space where it's more difficult for ancestry to spread (i.e. towards the north of Britain in the simulations – see Figure 3.10), parents with more migrant ancestry may be able to produce offspring which fill this space readily.

I choose to simulate bottleneck events in the north of Britain, which is where it's most difficult for migrant ancestry to spread. These events are set to happen uniformly in the rectangular area, which I term the *bottleneck region*, 0.9 units (approximately 210km) from the north coast and covering the entire width of the rectangle. A visualisation of this regime is shown in Figure 3.13. I choose to model bottlenecks by scaling the fitness of all individuals in this bottleneck region by a constant I call λ (this scaling occurs after the calculation described by Equation 3.2) – λ can be varied and one can investigate how this affects the spread of ancestry. I assume that the bottleneck event only acts on individuals in a particular generation and that there are no lingering effects (of what caused the bottleneck, e.g. the disease or famine) on the next generation (although it is affected by the smaller population size). The high dispersal parameter means that the bottleneck region re-populates quickly after the bottleneck event (see Figure 3.14).

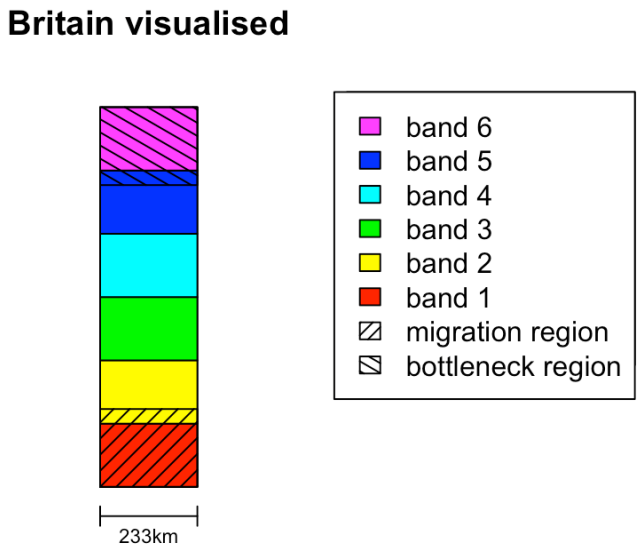


Figure 3.13: Bottleneck region visualised, for comparison with Figures 3.7 and 3.9. I show the 6 bands from Figure 3.9, which I continue to use in this section.

I first investigate the effect of such bottleneck events on population size when migration is set to only 1 migrant per generation, so as to have a negligible effect on the population dynamics. This is done to try and explore how quickly the bottleneck region's population takes to reach roughly its pre-bottleneck size. To this end, the population size in 6 bands (shown in Figure 3.13) over **P1** and **P2** is plotted in Figure 3.14. Note that the bottleneck region covers all of band 6 and 38% of band 5. As expected, one sees that population size is most severely affected by the bottleneck in these two bands. Overall, population size in all the bands climbs back to its pre-bottleneck levels and stabilizes within approximately 10 generations, or 300 years, in this worst-case scenario ($\lambda = 0$).

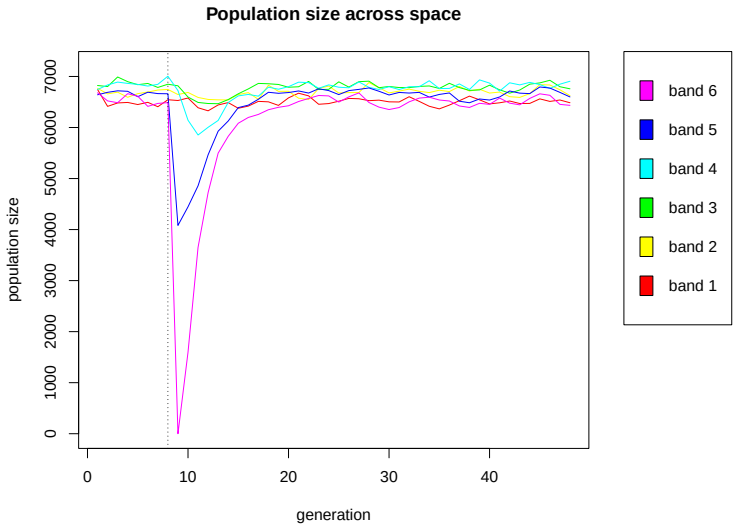


Figure 3.14: Population size over **P1** and **P2** in the same 6 bands as those used to track ancestry in Section 3.7 (as shown in Figure 3.13). In this simulation, λ is set to 0 so that all individuals in the bottleneck region are killed off at generation 9. A dotted vertical line showing the generation at which the bottleneck happens (generation 9) is included.

I continue to simulate bottleneck events that happen at the end of generation 9, or roughly in the year 2180BC. This generation is chosen since it's near the start of migration, before migrant ancestry has had the chance to spread significantly to the north (see Figure 3.10), which should boost the overall spread of migrant ancestry as explained above, yet it's late enough that many migrants are still killed off by the bottleneck event. Furthermore, it's early enough so that population size has the chance to re-stabilise while migration is still happening (see Figure 3.14). This choice of generation, however, is still somewhat arbitrary.

I investigate the dynamics during **P1**, and again look for the critical migration rate, in order to explore how the additional bottleneck affects the spread of ancestry, compared to that seen in Section 3.7. To this end, the average ancestry of the population at the end of **P1** against the number of migrants per generation for different values of λ is plotted in Figure 3.15 (similarly to Figure 3.8).

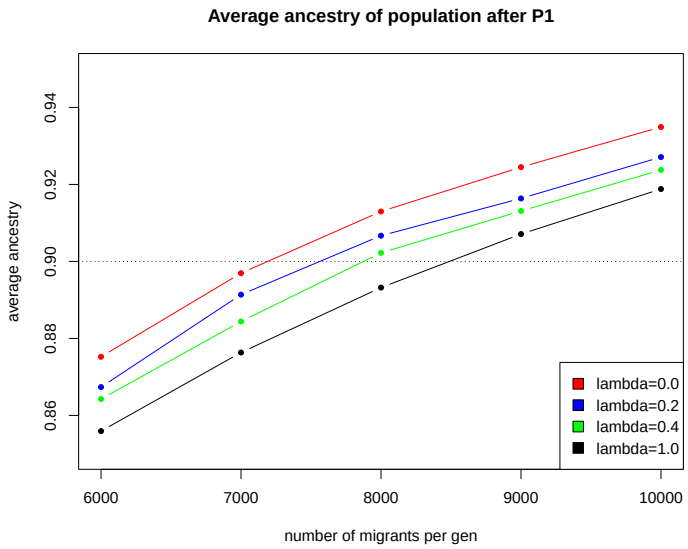


Figure 3.15: Average ancestry of individuals after **P1**. The coloured lines correspond to the different values of λ tested: 0.0, 0.2 and 0.4. The y-axis gives the average ancestry over the whole population, and the x-axis gives the number of migrants per generation. A line with $\lambda = 1$ (i.e. no bottleneck event, as in Figure 3.8) is also included for comparison.

One sees in Figure 3.15 that the critical migration rate decreases as λ increases. This is to be expected since a smaller λ corresponds to a bottleneck event which involves more individuals dying, which means that migrant ancestry can more quickly fill the now empty spaces. It's also clear in the figure that increasing λ linearly does not have a linear effect on the average ancestry after **P1**. For example, one sees in the figure that changing λ from 0.2 to 0 appears to have a greater effect on the spread of ancestry (larger change in average ancestry at the end of **P1** for a given migration rate) than changing it from 0.4 to 0.2.

Clearly, adding bottleneck events decreases the critical migration rate when compared to the results in Section 3.7, where I found a critical number of migrants per generation of approximately 8490. One might also hypothesise that the plateaus seen in Figure 3.8 may begin at higher migration rates when these bottleneck events are added.

For the remainder of this section, I use $\lambda = 0.2$ and the corresponding critical 7560 migrants per generation (estimated by linear interpolation on the results shown in Figure 3.15). This parameter choice is arbitrary but is taken because it gives a population bottleneck which has a significant effect on the spread of ancestry. This migration rate corresponds to approximately 252 migrants per year.

To investigate how well ancestry spreads through regions in space, and how the spread of ancestry is affected by the addition of a bottleneck, the average ancestry over **P1** and **P2** of individuals in the same 6 bands (used in Figure 3.14 and other figures) is plotted in Figure 3.16. Comparing this figure to Figure 3.10, one sees the subtle but visible effect that the bottleneck event has. Namely, one sees a sharp increase in the average ancestry in band 6 after the bottleneck event. This increase is perhaps more significant than it initially seems when one remembers that without the bottleneck event, it's much more difficult for ancestry to spread into northern regions than southern regions.

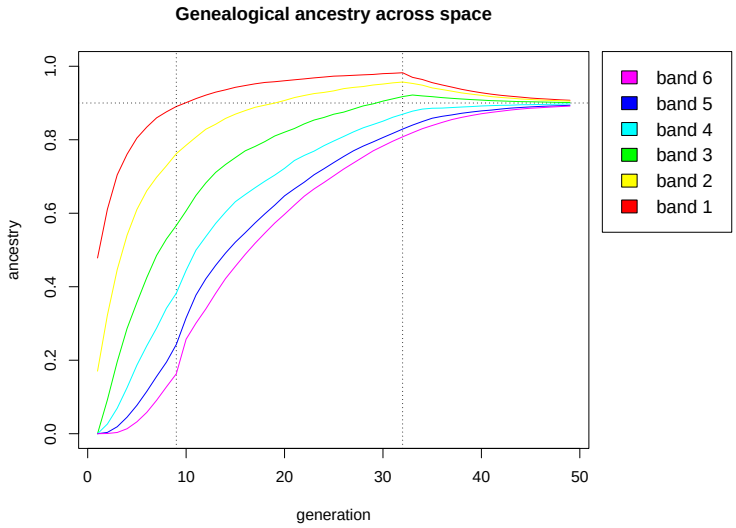


Figure 3.16: Spread of ancestry over **P1** and **P2** in the 6 bands visualised in Figure 3.13. The x-axis corresponds to the time since the beginning of **P1** in generations, and the y-axis corresponds to the average ancestry of individuals in each particular band. A dotted horizontal line showing 90% ancestry, a dotted vertical line showing the generation in which the bottleneck happens, and a dotted vertical line showing the end of **P1** and the beginning of **P2** are included.

To see whether the results of the model map any better to the discretisation of the data from Olalde et al. 2018 with the added bottleneck (compared to Section 3.7), I investigate the tightening of ancestry proportions in this bottleneck model, for one particular simulation (for comparison with results relating to Figure 3.11 – see Figure C.4 for comparable figure, and Figure 3.18 for further comparison). One sees mostly very similar patterns of behaviour, with the small difference that there are more outliers with small migrant ancestry proportions at the end of **P1** and a few generations after that in the bottleneck simulation. Despite this, the distribution of ancestry proportions at the end of **P2** seems to be roughly as tight as the model in Section 3.7. This may suggest that the bottleneck event together with the smaller number of incoming migrants does not significantly affect the longer-term evolution of the population.

Comparing the results of this simulation to the criteria laid out in Section 3.2 and the corresponding statistics from Section 3.7.1, one sees that:

- 100% of individuals have at least 59.7% migrant ancestry at the end of **P1**, as in Section 3.7.1. The minimum ancestry of individuals at the end of **P1** is 67.0%, compared to 65.3% in Section 3.7.1.
- 24.2% of individuals have 100% migrant ancestry at the end of **P1**, compared to 27.5% of individuals in Section 3.7.1.
- 100% of individuals have at least 78.9% migrant ancestry at the end of **P2**, as in Section 3.7.1.
- No individuals have 100% migrant ancestry at the end of **P2**, as in Section 3.7.1.

The fact that the results from this section are so close to those in Section 3.7.1 may further suggest that the additions in this model do not have a significant effect on the long-term evolution of the population.

As a final caveat to this section, note that the results throughout the section may be affected by the artificial scaling described in Section 3.6.5. This artificial scaling has two counteracting effects in the context of the bottleneck model:

- Individuals in the northern 1 x 0.4 unit rectangle are less likely to be killed off by the bottleneck event than they otherwise might be (for $\lambda > 0$). This will slow down the spread of migrant ancestry.
- After the bottleneck event, individuals outside of the bottleneck region are more likely to survive and reproduce in the northern 1 x 0.4 unit rectangle than they otherwise might be. This will speed up the spread of migrant ancestry.

It's unclear which of these effects wins out in general, and it likely depends on the choice of λ. Future researchers may wish to investigate what happens to the results of this section in models where the artificial fitness scaling described in Section 3.6.5 is not imposed.

3.9 Third continuous spatial model: fitness advantage to migrant ancestry

In this section, I consider that migrant genes may have had some fitness advantage, and how implementing this fitness advantage affects the spread of ancestry through space.

To implement this fitness advantage, I take the genealogical ancestry as being directly proportional to some contribution to the fitness of an individual. Explicitly, the scaling seen in Equation 3.2 now becomes

$$\left[\frac{\eta_i}{1 + n_i/(K(1+L))}\right] \cdot [(1-c) + c \cdot a_i] \tag{3.4}$$

where c is the contribution of migrant ancestry to the fitness scaling, some number in $[0, 1)$, and a_i is the genealogical ancestry of individual i, some number in $[0, 1]$. Other methods of scaling are possible, but scaling in this way with a pre-factor determined by the carrying capacity, spatial competition and mean lifetime (as seen in Equation 3.2) ensures that the fitness scaling is kept on the same order as if there weren't any scaling due to ancestry. This scaling also facilitates population growth as ancestry spreads, since as the average migrant ancestry in the population increases, so does the average fitness scaling, meaning that more individuals are likely to survive a given generation. Of course, a higher contribution parameter c will mean this population growth will be quicker and population size will be larger over time.

It's worth noting here that an increase in fitness due to migrant ancestry could map to many population factors. Of course, it could map to certain migrant genes being better suited to the British environment, though it could also map to some of the following cultural factors, for example:

- Migrants may tend to have more children or may continue to have children as they grow older more frequently.
- Migrant communities may be more accustomed to sharing resources such as food (which would map especially nicely to Equation 3.4 in which the harshness of the penalty due to spatial competition, and other factors, is effectively determined by c and a_i).

I now investigate **P1** and once again look for the critical migration rate, in order to explore how the additional fitness advantage to migrant ancestry affects the spread of ancestry, compared to that seen in Sections 3.7 and 3.8. To this end, the average ancestry of the population at the end

of **P1** against the number of migrants per generation for different values of c is plotted in Figure 3.17.

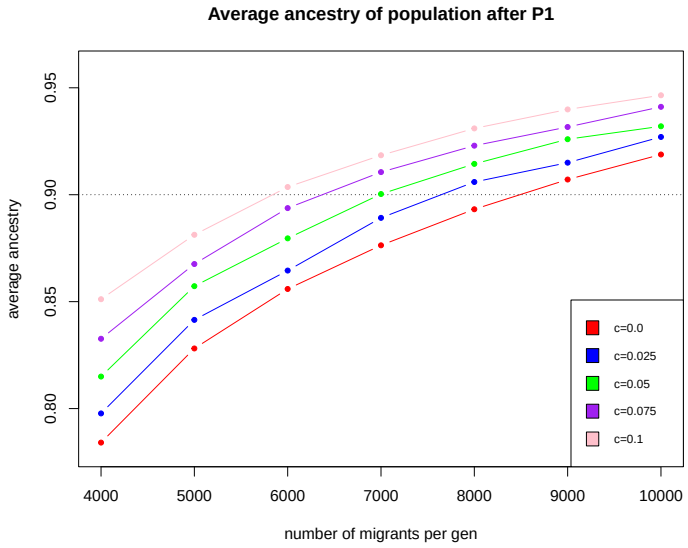


Figure 3.17: Average ancestry of individuals after **P1**. The coloured lines correspond to the different values of c tested: 0.025, 0.05, 0.075 and 0.1. The y-axis gives the average ancestry over the whole population, and the x-axis gives the number of migrants per generation. A line with $c = 0$ (i.e. no fitness advantage to migrant ancestry, as seen in Figure 3.8) is also included for comparison.

One sees in Figure 3.17 that the critical migration rate decreases as c increases. This is to be expected – the more that migrant ancestry positively contributes to the fitness of individuals, the quicker migrant ancestry will be able to spread. One can conclude from these results that including even a small fitness advantage (corresponding to a small value of c) to migrant genealogy results in a much faster spread of ancestry through space.

For the remainder of this section, I use $c = 0.1$ and the corresponding critical 5840 migrants per generation (estimated by linear interpolation on the results in Figure 3.17). This choice of c is somewhat arbitrary but is taken because it gives a contribution to the fitness scaling that has a significant effect on the spread of ancestry. This migration rate corresponds to approximately 195 migrants per year.

The spread of ancestry in the same 6 bands (seen in Sections 3.7 and 3.8) has no obvious differences compared to the spread seen in Figure 3.10 (see Figure C.3 for a comparable figure). This suggests that the spread of ancestry through space may not be significantly affected by introducing this fitness advantage to migrant ancestry while decreasing the migration rate appropriately.

To explore whether the results of the model map any better to the discretisation of the data from Olalde et al. 2018 with the added fitness advantage to migrant ancestry (compared to Section 3.7), I investigate the tightening of ancestry proportions during **P2** in this particular regime (for comparison with results relating to Figure 3.11 – see Figure C.5 for comparable figure, and Figure 3.18 for further comparison). One sees very few distinguishable differences between the results of the two models, which further suggests that this additional fitness advantage to migrant ancestry, together with the smaller number of incoming migrants, may not significantly affect the longer-term evolution of the population. One difference to be aware of, however, is the slightly smaller resultant population size from this fitness advantage regime. This is particularly the case in **P2** where incoming migrants do not contribute to the population size (this phenomenon is discussed

in more detail at the end of this section). This results in the population being approximately 1000 individuals smaller, on average, during **P2** than the population in Section 3.7.

In fact, comparing the results of this simulation to the criteria laid out in Section 3.2 and the corresponding statistics from Section 3.7.1, one sees that:

- 100% of individuals have at least 59.7% migrant ancestry at the end of **P1**, as in Section 3.7.1. The minimum ancestry of individuals at the end of **P1** is 68.6%, compared to 65.3% in Section 3.7.1.
- 18.7% of individuals have 100% migrant ancestry at the end of **P1**, compared to 27.5% of individuals in Section 3.7.1.
- 100% of individuals have at least 78.9% ancestry at the end of **P2**, as in Section 3.7.1.
- No individuals have 100% migrant ancestry at the end of **P2**, as in Section 3.7.1.

The slight differences between the results in the two models here may suggest that the added fitness advantage of migrant ancestry allows migrant genealogy to spread more quickly through the population in earlier generations.

To qualify the discussion of this final spatial model, remember that adjusting the fitness scaling as described in Equation 3.4 affects the population size, as briefly mentioned above. Of course, when there's little migrant ancestry in the population the scaling means that the average fitness of individuals would be less than if Equation 3.2 were applied. This means that the population size could only reach the levels seen in Section 3.7 if all individuals in the population had full migrant ancestry. However, during migration (which begins at the start of the simulation), new migrants counter this decreased population size simply by entering Britain. As ancestry spreads and the average ancestry of individuals gets closer to 1, the fitness scaling in Equation 3.4 approaches what it would be under Equation 3.2, which means that the effect of reducing population size is dulled. It would be interesting to explore how different values of c and different migration rates affect the population size and whether this subsequently affects the spread of ancestry, but I choose not to do so in this thesis for the sake of time.

3.10 A brief comparison of results

In this section, I make some comparisons of results from the sections of this chapter.

To explore how using different models affects the critical migration rate, the number of migrants per year estimated throughout the sections in this chapter is compared in Table 3.1. The table acts as a comparison of models for **P1**. One sees in the table that the introduction of spatial effects results in increased critical migration rates.

Model	Estimated migration rate (migrants per year)
analytic	38
non-spatial WF	93
stepping stone*	> 93
continuous spatial	283
continuous spatial with bottleneck	252
continuous spatial with fitness advantage	195

*I don't estimate a migration rate in Section 3.5, but for the maximum connectedness given this migration rate, it's impossible to obtain the desired ancestry spread. This means that 93 migrants per year can be viewed as some form of lower bound for migration rate in this type of stepping stone model.

Table 3.1: Resulting migration rates in different models. Results are taken from Sections 3.3, 3.4, 3.5, 3.7, 3.8 and 3.9.

To elucidate the extent to which ancestry spreads through continuous space when either a bottleneck event or a fitness advantage to migrant ancestry is added to the model, the ancestry proportions after **P2** obtained from simulations in Sections 3.7, 3.8 and 3.9 are overlaid in Figure 3.18. The figure acts as a comparison of continuous spatial models for **P2**.

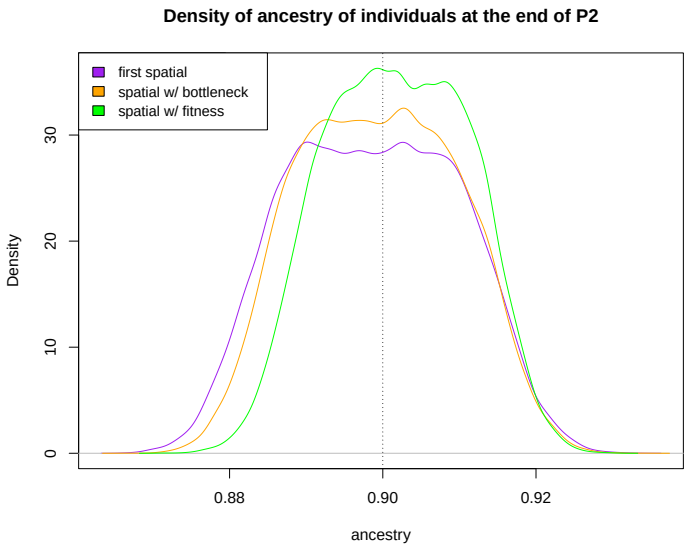


Figure 3.18: Comparison of ancestry distributions at the end of **P2** in models from Sections 3.7, 3.8 and 3.9. For each model, I take a random sample of 39,000 individuals so that the relative densities are comparable. The x-axis shows the ancestry proportions of individuals and the y-axis estimates the relative frequency with which that ancestry value occurs.

One sees in Figure 3.18 that the ancestry proportions corresponding to the model in Section 3.7 follow a more similar distribution to those corresponding to the model in Section 3.8 than those corresponding to the model in Section 3.9.

One also sees in Figure 3.18 that the ancestry proportions corresponding to the model in Section 3.9 are distributed more tightly than those corresponding to the model in Section 3.8, which are distributed more tightly than those corresponding to the model in Section 3.7. This further suggests that the added fitness advantage to migrant ancestry, along with the decreased migration rate, allows

migrant ancestry to spread more quickly through the population (with a similar but milder effect observed with the addition of the bottleneck along with the decreased migration rate). In the case of this specific parameter choice, the increased rate of spread in fact makes the distribution of ancestry proportions even tighter than is necessary to map to the discretisation of the data from Olalde et al. 2018, noting that the distribution of ancestry proportions was, if anything, already too tight in Section 3.7 (though this apparent over-tightening may again be a product of the discretisation – see Section 3.7.1).

Chapter 4

Discussion

4.1 Conclusions

In this thesis, I used forward simulations to investigate the turnover of British ancestry in the Copper Age and Early-Middle Bronze Age. I found that the introduction of spatial effects, in particular continuous spatial effects, resulted in a larger critical migration rate.

I used a stepping stone model, as an intermediate between non-spatial models and continuous spatial models, to find that the distance between parents and offspring over a generation should be on the order of 117km. I found that the spread of ancestry seen in Olalde et al. 2018 required a similarly large distance between parents and offspring (in the absence of a large mating radius).

Modelling, I found that a non-spatial simulation resulted in a critical migration rate over twice as large as the critical migration rate in the analytic estimate. I also found that introducing continuous spatial effects resulted in a critical migration rate over three times as large as in the non-spatial simulation.

I also investigated two mechanisms by which the critical migration rate could be reduced in the simulations: bottlenecks and fitness advantages to migrant ancestry.

I found that bottleneck events which caused the population size to decrease by approximately 7300 individuals ($\lambda = 0.2$) decreased the critical migration rate by approximately 31 migrants per year or roughly 11% (i.e. roughly 30,000 fewer incoming migrants required throughout **P1**).

I identified that if the amount of migrant ancestry that an individual had contributed 10% to their overall fitness (with the other 90% being determined by spatial competition, expected lifetime and carrying capacity), the critical migration rate decreased by approximately 88 migrants per year or roughly 31%. This is to say that even a modest contribution of migrant ancestry to total fitness resulted in a significantly faster spread of ancestry.

These numbers loosely give an idea of the relative scale of effects of the two additional mechanisms of bottlenecks and fitness advantages.

I found that adding a fitness advantage to migrant ancestry (together with decreased migration rate) or adding a bottleneck event (together with decreased migration rate) resulted in a tighter spread of ancestry at the end of the simulation.

In designing simulations, it became apparent that modelling a population with realistic dynamics, such as realistic lifetimes and relatively stable total population size, is quite a strong constraint in itself, even before introducing migration.

4.2 Scope of these results

Although I hope that the models and methodology set out in this thesis will be useful to those wishing to explore similar problems in the future, it's important to realise that the scope of the results in this thesis is still profoundly limited. The results found are only true under the very limited space of possible parameters that I chose to explore, and results are also built on many simplifications on the data, such as the formulation in terms of **P1** and **P2** and the discretisation into two sampling times, which may not map well to physical or biological reality. Furthermore, the assumptions made may make it unfair to compare the results of different models, such as those in Sections 3.8 and 3.9, completely like-for-like (when comparing critical migration rates, for example).

Despite this, the results still give some intuition into what is necessary to build a realistic model and into some of the scales that mechanisms, such as bottlenecks, have on the resulting evolution of the population. If one finds that the parameters required for the 90% replacement of Britain's gene pool, such as the large dispersal, are unrealistic, they may wish to run simulations with more modest parameters, perhaps with added mechanisms which allow ancestry to spread more quickly (such as bottlenecks and fitness advantages to migrant ancestry).

This thesis should be viewed as a pilot study – I've set out some models and considerations that are needed to understand how forward spatial simulations can be used to explore the mechanisms behind the turnover of British ancestry from 2450BC-1000BC and subsequently investigate limitations on parameters, but clearly more data are needed, and further research needs to be undertaken, in order to draw conclusions that may be more informative.

For example, one would need a better understanding of which parameters of the population simulation are reasonable, such as the mating radius and the dispersal. This understanding could come perhaps from archaeological research or aDNA studies.

This thesis demonstrates the advantages and limitations of using forward simulations to understand population genetic data. In a wider sense, it also shows the usefulness of simulations in science, especially in an age where high-performance computing is becoming more and more accessible. It would be impossible to run many real-world experiments, on the scale of population evolution, for example, to infer results. Therefore, for a system where many of the features are known in detail (e.g. if one could be confident in their choice of dispersal, mating radius and other model parameters), simulation makes sense as a method of investigation.

Finally, note that since I've focused on genealogical ancestry throughout this thesis (rather than genetic ancestry), I haven't taken full advantage of SLiM, which is designed to efficiently simulate genomes explicitly. I lay out a framework for using genetic ancestry in the context of this analysis, and hence taking better advantage of SLiM, in Appendix E.

4.3 Further research

4.3.1 Extensions to the methods used

Throughout this thesis, I've given some suggestions as to directions for further research. Here, I highlight a subset of further ideas more explicitly.

It would certainly be ideal to run spatial simulations with larger mating radii, in which individuals can mate with others at a larger distance, than the simulations above. This would allow for a smaller, perhaps more realistic, dispersal. Such simulations are not difficult to program but would require more computing power than is currently available to me (see Section 3.6.4).

In Section 3.8, I chose that the bottleneck event would occur in generation 9. One could explore the effect on the spread of ancestry if this choice of generation is varied, if there are multiple

bottleneck events, or if fitness effects last for multiple generations, for example.

One could consider the fact that the carrying capacity of the environment would not be constant across the whole of Britain in a more realistic simulation. For example, it's likely that Scotland was much more sparsely populated than England, and this could be modelled by using a smaller carrying capacity in the north of Britain than in the south. One could then investigate how this affects the spread of ancestry, and in particular how ancestry spreads across the boundaries separating regions with different carrying capacities (perhaps England and Scotland).

One could think about uncertainties in migration rates and the range of migration rates that produce a particular range of outcomes (such as average ancestry falling within certain bounds).

One could also add certain features to the simulations which map to physical and biological reality. For example, the introduction of sexes so that males can only mate with females, the migration of children with their parents, or a changing migration rate over time. Exploring the addition of certain societal phenomena such as cities may also be interesting.

These suggestions for further research are certainly not exhaustive, and the reader is encouraged to explore any ideas they might have by adding to the frameworks set out and the code provided.

4.3.2 Requiring new methods

Future researchers may wish to explore similar questions using the *slendr* R package (Petr et al. 2022 – currently a preprint). *slendr* uses SLiM as a back-end and allows for simulation across complex geospatial landscapes, which would allow one to use a more realistic map of Britain on which to run simulations. The package also allows for convenient analysis of tree sequences, which may be useful in the exploration of how genetic ancestry spreads across space (see Appendix E).

An obvious flaw of this thesis, touched on in Section 3.6.2, is that only a specific parameter set is simulated, and I tended to investigate what happened when individual parameters were varied. It's possible, however, that multiple sets of parameters, changed together, could produce similar results. One could perhaps use machine learning methods (e.g. deep learning), or other methods, to tune parameters *simultaneously*, with the goal of simulations satisfying criteria such as stable population size and a minimum average ancestry over the population.

Finally, future researchers may wish to investigate similar questions, letting go of the discretisation into two sampling times, and sampling genomes throughout the evolution of the population, perhaps over multiple simulations. One could also let go of the assumption that migration started in 2450BC, and try to estimate the year in which migration started by mapping results of sampling throughout the evolution to the genomes in Olalde et al. 2018 with their corresponding dates.

Bibliography

Brothwell, Don (1972). "Palaeodemography and earlier British populations". In: *World Archaeology* 4.1. PMID: 16468215, pp. 75–87. DOI: 10.1080/00438243.1972.9979521. eprint: https://doi.org/10.1080/00438243.1972.9979521. URL: https://doi.org/10.1080/00438243.1972.9979521.

Franks, Steven J., Paul D. Pratt, and Neil D. Tsutsui (2011). "The genetic consequences of a demographic bottleneck in an introduced biological control insect". In: *Conservation Genetics* 12.1, pp. 201–211. DOI: 10.1007/s10592-010-0133-5. URL: https://doi.org/10.1007/s10592-010-0133-5.

Buffalo, Vince, Stephen M Mount, and Graham Coop (Sept. 2016). "A Genealogical Look at Shared Ancestry on the X Chromosome". In: *Genetics* 204.1, pp. 57–75. ISSN: 1943-2631. DOI: 10.1534/genetics.116.190041. eprint: https://academic.oup.com/genetics/article-pdf/204/1/57/42221787/genetics0057.pdf. URL: https://doi.org/10.1534/genetics.116.190041.

Coop, Graham (Dec. 2017). *Where did your genetic ancestors come from?* URL: https://gcbias.org/2017/12/19/1628/.

Cayuela, Hugo et al. (2018). "Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review". In: *Molecular Ecology* 27.20, pp. 3976–4010. DOI: https://doi.org/10.1111/mec.14848. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14848. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14848.

Olalde, Iñigo et al. (2018). "The Beaker phenomenon and the genomic transformation of northwest Europe". In: *Nature* 555.7695, pp. 190–196. DOI: 10.1038/nature25738. URL: https://doi.org/10.1038/nature25738.

Bradburd, Gideon S. and Peter L. Ralph (2019). "Spatial Population Genetics: It's About Time". In: *Annual Review of Ecology, Evolution, and Systematics* 50.1, pp. 427–449. DOI: 10.1146/annurev-ecolsys-110316-022659. eprint: https://doi.org/10.1146/annurev-ecolsys-110316-022659. URL: https://doi.org/10.1146/annurev-ecolsys-110316-022659.

Haller, Benjamin and Philipp Messer (Mar. 2019). "SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model". In: *Molecular biology and evolution* 36, pp. 632–637. DOI: 10.1093/molbev/msy228.

Battey, C J, Peter L Ralph, and Andrew D Kern (May 2020). "Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data." eng. In: *Genetics* 215.1, pp. 193–214. ISSN: 1943-2631 (Electronic); 0016-6731 (Print); 0016-6731 (Linking). DOI: 10.1534/genetics.120.303143.

Lynch, Michael and Wei-Chin Ho (Mar. 2020). "The Limits to Estimating Population-Genetic Parameters with Temporal Data". In: *Genome Biology and Evolution* 12.4, pp. 443–455. ISSN: 1759-6653. DOI: 10.1093/gbe/evaa056. eprint: https://academic.oup.com/gbe/article-pdf/12/4/443/33161388/evaa056.pdf. URL: https://doi.org/10.1093/gbe/evaa056.

Mathieson, Iain and Aylwyn Scally (Mar. 2020). "What is ancestry?" In: *PLOS Genetics* 16.3, pp. 1–6. DOI: 10.1371/journal.pgen.1008624. URL: https://doi.org/10.1371/journal.pgen.1008624.

Armit, Ian and David Reich (2021). "The return of the Beaker folk? Rethinking migration and population change in British prehistory". In: *Antiquity* 95.384, pp. 1464–1477. DOI: 10.15184/aqy.2021.129.

Harney, Éadaoin et al. (Apr. 2021). "Assessing the performance of qpAdm: a statistical tool for studying population admixture." eng. In: *Genetics* 217.4. ISSN: 1943-2631 (Electronic); 0016-6731 (Print); 0016-6731 (Linking). DOI: 10.1093/genetics/iyaa045.

Haller, Benjamin (Feb. 2022). *Eidos: A Simple Scripting Language*. Eidos version 2.7.1. Dept. of Computational Biology, Cornell University. Ithaca, NY 14853.

Haller, Benjamin and Philipp Messer (Feb. 2022). *SLiM: An Evolutionary Simulation Framework*. SLiM version 3.7.1. Dept. of Computational Biology, Cornell University. Ithaca, NY 14853.

Petr, Martin et al. (2022). "slendr: a framework for spatio-temporal population genomic simulations on geographic landscapes". In: *bioRxiv*. DOI: 10.1101/2022.03.20.485041. eprint: https://www.biorxiv.org/content/early/2022/03/21/2022.03.20.485041.full.pdf. URL: https://www.biorxiv.org/content/early/2022/03/21/2022.03.20.485041.

Appendices

A Code

The relevant SLiM recipes, with some example R and Shell scripts, used to generate the results in this thesis are given on this thesis' accompanying Github page (https://github.com/dg622cam/thesis_code).

B Further details for the minimal non-spatial model seen in Section 3.4

In this model, migration occurs at the juvenile stage. This means that parents are chosen from each migration source (there's only one migration source in this model, but multiple migration sources are used in the stepping stone model in Section 3.5) with a probability equal to the corresponding migration rate, and chosen from the original population with probability

$$1 - \sum (\text{incoming migration rates})$$

i.e. $1 - m$ in the minimal non-spatial model. This is to say that migration is stochastic in this model and the stepping stone model (Haller and Messer 2022), in that the number of migrants per generation (or individuals moving between sub-populations in the stepping stone model) is effectively chosen from a binomial distribution. This notion of migration is similar to that used in Section 3.3, in that much of the gene pool from the previous generation is forced out of the population by migrants (although this now happens randomly and is no longer biased towards forcing out indigenous base pairs). This assumption is relaxed, however, in the continuous spatial models. One could alternatively choose to use a non-spatial non-Wright-Fisher model, in which population size is emergent rather than enforced, to prevent this effect of much of the gene pool being forced out.

As explained in Section 3.1, I assume a constant migration rate during **P1** and I assume it lasts for 32 generations, and look for the critical migration rate.

To explore the effects of increased migration rate, I look for the time taken for a simulation to reach 90% average ancestry for different migration rates, shown in Figure B.1. As expected, one sees that as m increases, the time taken for the simulation to meet the stopping criterion (the average ancestry over the whole population reaching 90%) decreases. The figure appears to show that the critical migration rate is approximately 0.07. This corresponds to a mean of 2800 migrants per generation with a standard deviation of $\sqrt{40000 \cdot 0.07 \cdot (1 - 0.07)} \approx 51.0$ migrants per generation, or a mean of approximately 93.3 migrants per year and a standard deviation of approximately 1.7 migrants per year (calculations based on the fact that the stochastic migration regime explained above corresponds to a binomial distribution).

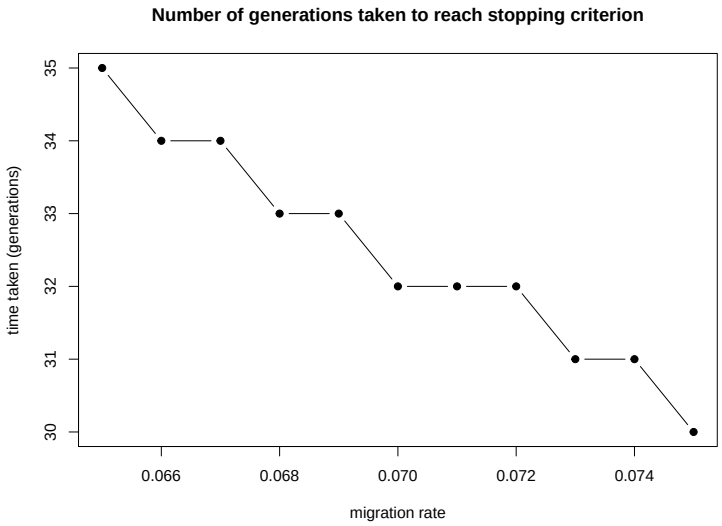


Figure B.1: Number of generations taken for the average ancestry over the whole population to reach the threshold in the non-spatial Wright-Fisher model. The x-axis shows the different values of migration rate m (in a small window around the critical migration rate), and the y-axis shows the corresponding number of generations taken for the average ancestry over the whole population to reach 90%.

With a view to investigate whether the ancestry proportions in the model map to those seen in the discretisation of the data from Olalde et al. 2018, the distribution of ancestry proportions in the simulation at the end of generations 32 and 49 (i.e. at the end of **P1** and **P2**) is plotted in Figure B.2. During **P2**, after migration has stopped, one expects ancestry proportions to become more uniform over the population, and this is indeed the behaviour one observes in the figure. However, the ancestry proportions in the data from Olalde et al. 2018 are much less uniform in the Middle Bronze Age than at the end of **P2** in this figure. In fact, the proportions appear to have a wider spread in the original data (one sees a minimum of 59.7% migrant ancestry in the Copper Age and Early Bronze Age – see Section 3.2) when compared to the proportions at the end of **P1** in this simulation, which is to say that **P2** is somewhat obsolete in this model if the goal is solely to map the simulation to the data. This behaviour is likely a result of the panmixia in the model, which causes ancestry to spread more readily. In a more complex spatial model, it's likely that **P2** will still be necessary to obtain results similar to those seen in the discretisation of the data from Olalde et al. 2018.

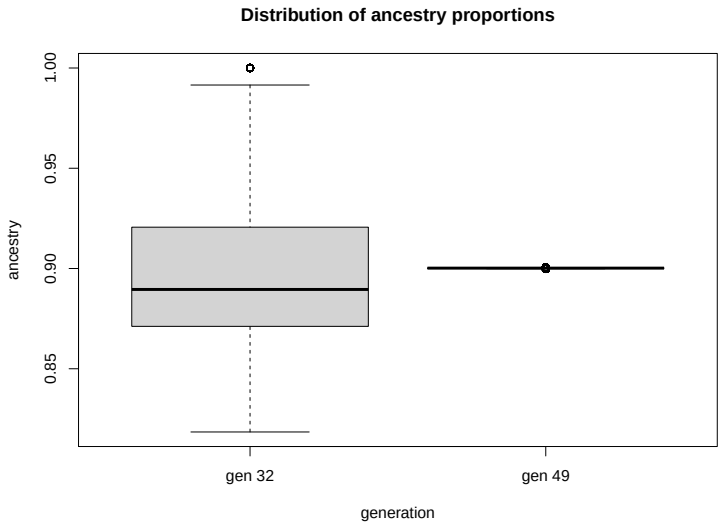


Figure B.2: Box plots showing the distributions of ancestry proportions of individuals in non-spatial Wright-Fisher model at the end of **P1** (gen 32) and the end of **P2** (gen 49). The ancestry proportion is shown on the y-axis.

C Additional figures

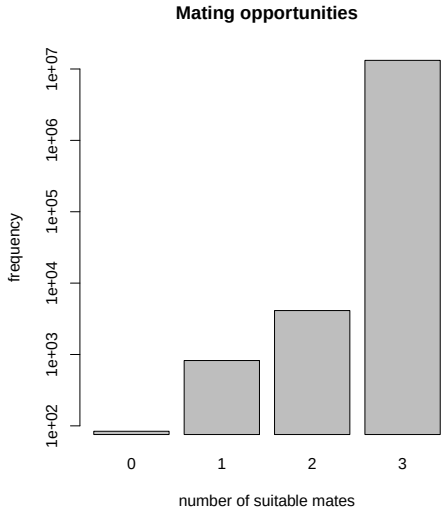


Figure C.1: Mating opportunities in spatial simulations, with results over 10 simulations with different random seeds summed, only simulating **P1**. SLiM goes through each focal individual in each generation and counts the number of suitable mates (number of individuals within a distance S, capped at 3 individuals) that the focal individual has (shown on the x-axis). The y-axis shows the number of occurrences of individuals having this number of suitable mates and is scaled logarithmically. One sees that the choice of S taken ($S = 0.02$) is sufficiently large so that almost all focal individuals have the opportunity to mate in each generation. Note that this also implies that 3 or more individuals strongly affect how each focal individual suffers from spatial competition in the vast majority of cases.

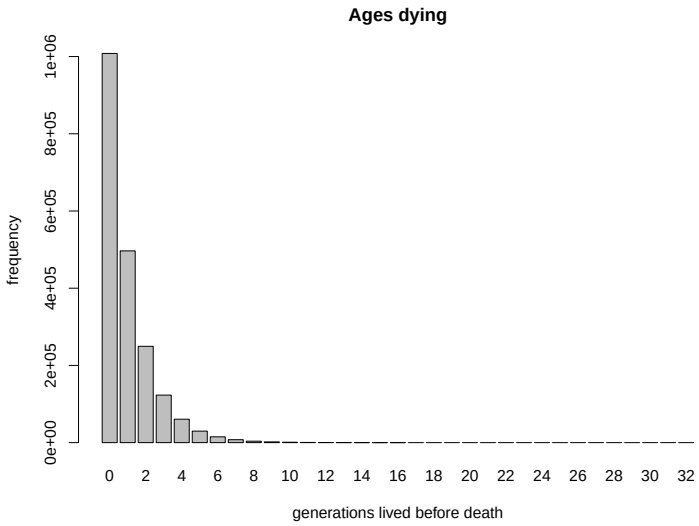


Figure C.2: Ages that individuals die if age cutoff is not imposed, in a particular simulation run over **P1**, for comparison with Figure 3.6. Migration is set to 1 individual per generation, so as not to affect the evolutionary dynamics. The x-axis shows the age of death of individuals (in generations), and the y-axis shows the frequency of occurrences of individuals dying after that many generations. One sees that many individuals live beyond realistic ages.

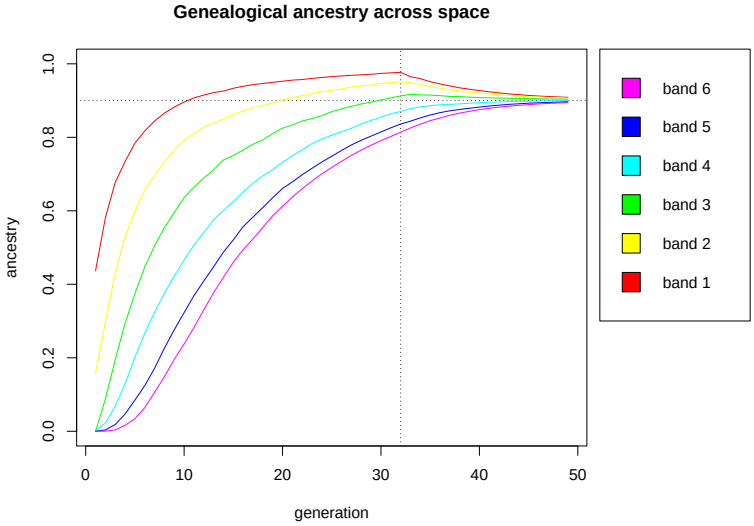


Figure C.3: Spread of ancestry over **P1** and **P2** in the 6 bands used in Sections 3.7 and 3.8, for one particular simulation of the model with an added fitness advantage to migrant ancestry, seen in Section 3.9. c is set to 0.1 and the migration rate is set to the critical migration rate of 5840 migrants per generation seen in Section 3.9. The x-axis corresponds to the time since the beginning of **P1** in generations, and the y-axis corresponds to the average ancestry of individuals in each particular band. A dotted horizontal line showing 90% ancestry, and a dotted vertical line showing the end of **P1** and beginning of **P2** are included. For comparison with Figures 3.10 and 3.16.

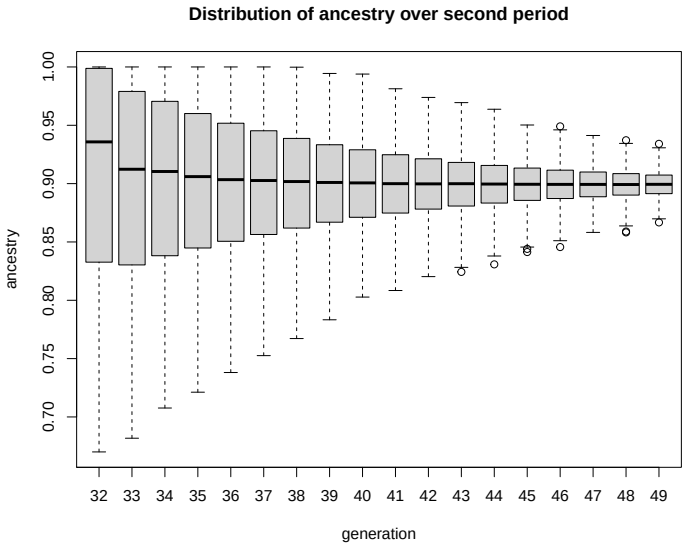


Figure C.4: Distribution of ancestry proportions over **P2** as box plots in a particular simulation of the bottleneck model, seen in Section 3.8. λ is set to 0.2 and the migration rate is set to the critical migration rate of 7560 migrants per generation seen in Section 3.8, The x-axis shows the generation since the start of **P1**, and the y-axis shows the ancestry proportions of individuals. For comparison with Figures C.5 and 3.11.

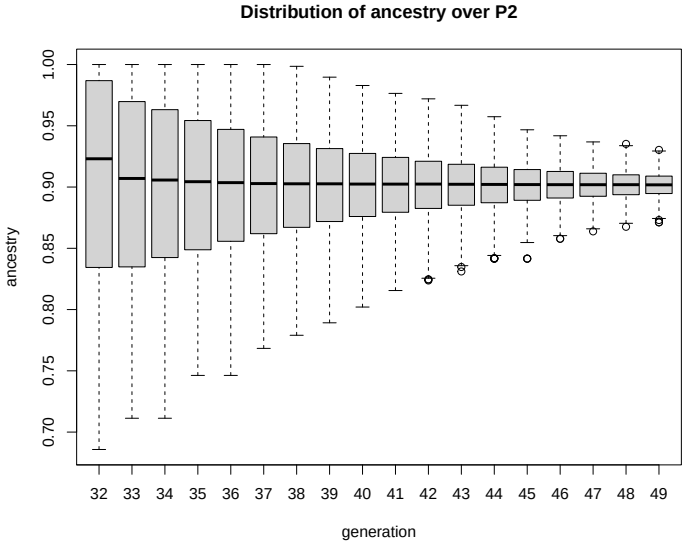


Figure C.5: Distribution of ancestry proportions over **P2** as box plots in a particular simulation of the model with an added fitness advantage to migrant ancestry, seen in Section 3.9. c is set to 0.1 and the migration rate is set to the critical migration rate of 5840 migrants per generation seen in Section 3.9. The x-axis shows the generation since the start of **P1**, and the y-axis shows the ancestry proportions of individuals. For comparison with Figures C.4 and 3.11.

D Details of the calculation in Section 3.7.1

To calculate the probability of sampling at least 17 out of 67 genomes with 100% migrant ancestry, given that 27.5% of genomes have 100% migrant ancestry, I make the following simplifying assumptions:

- Samples happen independently of one another and uniformly across the population.
- The number of individuals in the population ($\sim 40,000$) is large enough compared to the number of samples (67) that the probability of sampling a genome with 100% migrant ancestry does not significantly change as one continues to take samples.

With these relatively reasonable assumptions, one can say that the number of samples with 100% migrant ancestry can be described by a binomial distribution, with the number of trials $n = 67$ and the probability of success $p = 0.275$.

Then, one can use the known properties of the binomial distribution to proceed with the calculation. Namely, that the probability of a certain outcome (x sampled genomes having 100% migrant ancestry) is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

which leaves that the desired probability that at least 17 genomes have 100% ancestry is given by

$$\begin{aligned} P(X \geq 17) &= \sum_{k=17}^{n} P(X = k) \\ &= \sum_{k=17}^{n} \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

E Spread of *genetic* ancestry through space

Here, I discuss the modelling of genetic ancestry, as opposed to genealogical ancestry which has been modelled throughout this thesis.

E.1 Measuring genetic ancestry

To model the spread of genetic ancestry, one needs to track the ancestry of points along the genome. Here, I track only one chromosome (which should be thought of as an autosome) in diploid individuals, but it's easy to conceptualise how the analysis could be extended to more chromosomes (just by adding more independent base pairs explicitly). In the model of Britain with a dummy migrant population, one can initially define the genetic ancestry of each base pair of each individual in Britain at the start of **P1** as 0, and the genetic ancestry of each base pair of each new migrant as 1. As the population evolves, a child's genetic ancestry depends on whether the first base pair of each of the child's genomes was inherited from the corresponding parent's maternal or paternal genome, and where recombination events happened (which are tracked as explained in Appendix Section E.2). Throughout the simulation, each base pair thus has a corresponding genetic ancestry value of either 1 or 0.

An individual's total genetic ancestry can then be thought of as the average genetic ancestry of a base pair in their two genomes.

E.2 Tracking recombination events

To efficiently keep track of the genetic ancestry of individuals over many generations, rather than tracking the origin of every base pair, I choose to simply track the recombination events that occur and use these to infer the origin of the haplotypes, which is to say that my code keeps track of base pairs which are inherited together.

The method developed supports a simple "crossover breakpoints" model, which is SLiM's default way of modelling recombination. The only parameter for this model of recombination is the recombination rate per base pair per generation (which can vary along the chromosome but is kept constant here). In the model, one of the two parental genomes is chosen as a copy strand and the other as the opposite strand, and crossover breakpoints are drawn according to the recombination rate(s). At each breakpoint, the gametic genome switches from its current strand, taking base pairs from the corresponding parental genome until it reaches the next breakpoint (Haller and Messer 2022).

The method tracks genetic ancestry by storing vectors of ancestry values and vectors of breakpoints for each genome for each individual. The value of a haplotype can be either 0, corresponding to a chromosomal region derived from the indigenous population, or 1, corresponding to a chromosomal region derived from the migrant population. I use marker mutations at the first base pair of the chromosome to track which parental genomes SLiM chooses as copy strands.

The code required for this implementation is much more complex than the code used to track genealogical ancestry in this thesis and requires many more lines than the models seen. Code is available on the Github page.

This method is much more space and time efficient than the method described in Section 14.7 of Haller and Messer 2022, and allows for whole-chromosome-scale simulations (at least for simulations with a smaller number of individuals), compared to ~ 1000 bp simulations for the method in Haller and Messer 2022.

In the models, I use the standard recombination rate of 10^{-8} crossover events per base pair per generation (Haller and Messer 2022). I also use a chromosome length of 10^8 base pairs. To make simulations more efficient, one has the option to scale down the chromosome length whilst scaling up the recombination rate by the same factor (resolution may become a relevant consideration if this factor is set to a very large number, and depending on what one is trying to measure). In the example in Appendix Section E.3, this factor is set to 10^5.

One could extend this framework by modelling gene conversion, or exploring more realistic recombination maps, for example.

E.3 Example: spread of a deleterious mutation

This example model demonstrates the types of questions that one can think about answering using this framework.

I continue to model the spread of ancestry through Britain, and use the critical migration rate of 8490 migrants per generation, as in Section 3.7. I give all migrant individuals an allele with selection coefficient -0.8 (such a strong selection coefficient is needed to observe visible effects since so many migrant individuals are being forced into the population), placed at the final base pair of both their chromosomes (the right-most base pair, in the context of Figure E.1). I allow the population to evolve and measure the average ancestry along the chromosome across bands, using the same 6 bands as in Sections 3.7, 3.8 and 3.9. This is done to explore whether there are any noticeable differences at different regions in space in the way the genetic ancestry propagates, and the results are shown in Figure E.1.

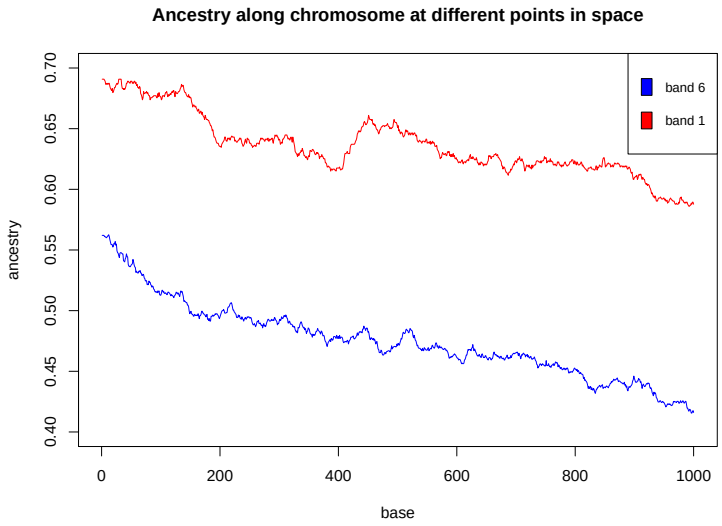


Figure E.1: Genetic ancestry along chromosome at the end of **P1**, for one particular simulation. To take these measurements, the code first identifies the individuals in the appropriate bands. For each of the base pairs of each of the genomes of each of these individuals, the code finds whether that base pair originated from the migrant or indigenous population. A value of 1 corresponds to the migrant population and a value of 0 corresponds to the indigenous population. The value on the y-axis of the graph corresponds to the average origin population of that base pair for individuals in that band. The x-axis corresponds to the position of the base pair on the chromosome, where the length of the chromosome is scaled down, as explained in Appendix Section E.2.

One sees in Figure E.1 that it's more difficult for migrant ancestry in base pairs towards the right to spread further north (into band 6). This is almost certainly due to the deleterious allele that migrants have in both their chromosomes at their right-most base pair. Recombination allows more migrant ancestry at base pairs towards the left to spread without this deleterious mutation.

F Differences due to stochasticity

Here, I briefly investigate the extent to which stochasticity affects the results of this thesis. To do this, in order to limit computational expense, I just investigate the first spatial model, seen in Section 3.7.

I look at two statistics which are relevant and likely to be relatively sensitive to stochasticity:

1. The average ancestry in band 6 at the end of **P1** (this relates to Figures 3.10, 3.16 and C.3).
2. The variance in ancestry proportions over the whole population at the end of **P2** (this relates to Figure 3.18).

For the first statistic, I use the same 6 bands as used in Sections 3.7, 3.8 and 3.9. I choose to focus on band 6 because it's likely that the northernmost areas of space will be more sensitive to stochasticity. Areas in space further south are less likely to be affected because migrants are placed in the population in the south in the models. However, it's possible the proportions in the north will be affected by individuals with larger ancestry proportions who spawn towards the north at early generations in the simulations, as touched on in Section 3.7.1.

For the second statistic, I look at the variance in ancestry proportions over the whole population at the end of **P2**, since it's been clear throughout this thesis that this is an important statistic in determining how well the results of the simulation map to the discretisation of the data from Olalde et al. 2018.

I choose to set the migration rate to the critical migration rate of 8490 migrants per generation, as in Section 3.7. I investigate how the average ancestry in band 6 (using the same 6 bands as in Sections 3.7, 3.8 and 3.9) at the end of **P1** varies over trials in Figure F.1, to see whether the choice of random seed affects the results of the simulations. One sees a small amount of variation in this statistic over trials. In fact, the largest and smallest values for the statistic obtained over the trials have a relative difference of approximately 0.9%. This small variation suggests that stochasticity over different trials does not have a significant effect on the spread of ancestry into band 6.

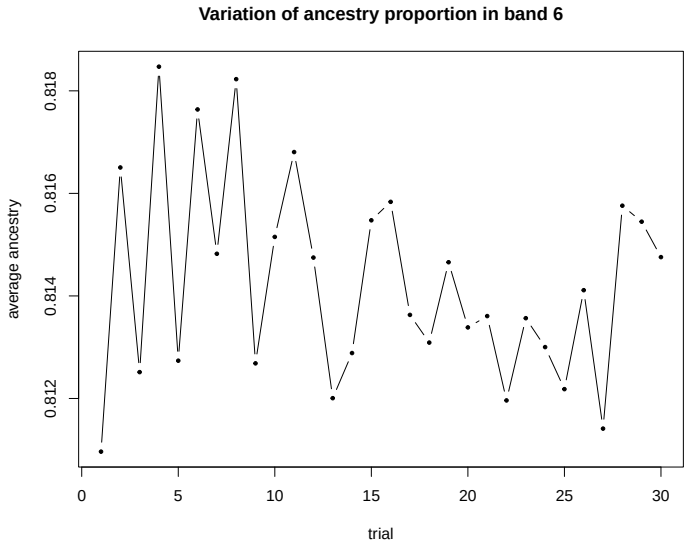


Figure F.1: Average ancestry in band 6 at the end of **P1**, over 30 trials run with different random seeds. The y-axis shows the average ancestry of all individuals in band 6.

I investigate how the variance in ancestry proportions over the whole population at the end of **P2** changes over trials in Figure F.2, to see whether the choice of random seed affects the results of the simulations. Again, one sees a small amount of variation in this statistic over trials, but a significant amount of relative variation. In fact, the largest and smallest values for the statistic obtained over the trials have a relative difference of approximately 17%. This suggests that stochasticity over different trials does have some effect on the variance in ancestry proportions at the end of **P2**. However, the fact that these variances remain very small indicates that they likely won't exceed the variance of proportions observed in the discretisation of the data from Olalde et al. 2018 over trials.

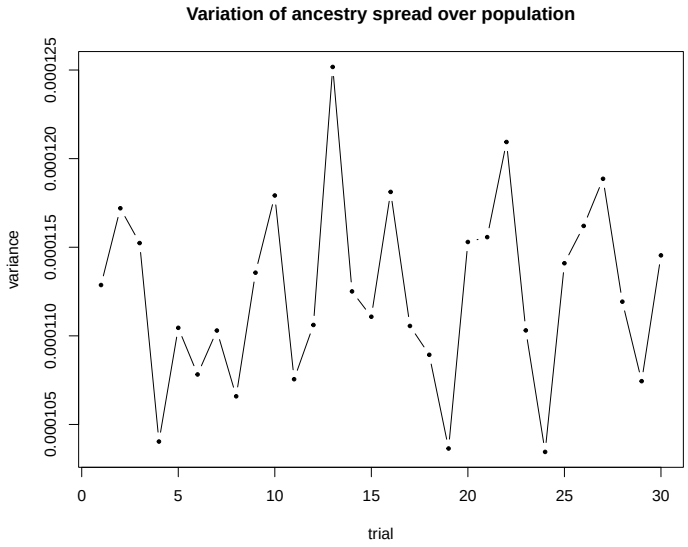


Figure F.2: Variance in ancestry proportions over whole population at the end of **P2**, over 30 trials run with different random seeds. The y-axis shows the variance of ancestry proportions over the whole population.