universität wien

MASTERARBEIT / MASTER THESIS

Titel der Masterarbeit / Title of the Master Thesis

„Coalescent Effective Population Size in Structured Populations: Between the Stepping Stone and the Island Model“

verfasst von / submitted by

Pia Gober, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Jitka Polechová, Privatdoz. PhD

Abstract (english)

For the german version, see Appendix A.
The effective population size N_e is often regarded as the size of an idealized population, which undergoes the same effects of drift or inbreeding as the focal one. In the field of coalescent theory, it is equivalent to the expected coalescence time of a random lineage pair in this population. This thesis focuses on the coalescent effective population size in the continuum between the one-dimensional, circular stepping stone and the island model. While the total population size is kept constant, the continuum is explored concerning the percentage of local/global migration for different migration rates and numbers of demes in the population. The formula for the effective population size of a more general migration model with changed migration rates is used to describe the behaviour of the effective size in the continuum. To represent the expected coalescence time of the stepping stone and island model, mathematical formulas as well as the population genetics simulator *msprime* are used. When it is appropriate to use the island model and when one should rather use the stepping stone model depends on the percentage of local migration to the neighboring demes. The results within this thesis suggest that this percentage of local migration, from which on the use of the stepping stone model can be justified, decreases for increasing migration rates. The interval, in which the island model is a better approximation, is generally much larger than the one, in which the circular stepping stone model should be used. Apart from that, the expected coalescence time seems to increase approximately linearly for increasing deme numbers for both the stepping stone and the island model.

Acknowledgements

I would like to thank the VSC-Team of the Technical University of Vienna for providing me the access to the VSC-4 cluster. In this regard, I especially want to thank my supervisor Jitka Polechová and Ben Wölfl of the University of Vienna for helping me with the technical implementation to run the Python code in the cluster. Without the VSC-4, it would not have been possible for me to get the plots with $d = 20$ that are presented below. Apart from this I want to thank Jitka Polechová for providing me with a variety of works related to the topic of my thesis as well as inputs on the structure of it. Last but not least, I also want to thank my family and friends, who supported me in different ways through the process of writing my master thesis.

Contents

Notations vii

1. Introduction **1**

I. Mathematical basics and derivation of the effective population size **3**

2. Models **7**
2.1. The (neutral) Wright-Fisher model 7
2.2. The Island model 8
2.3. The circular Stepping Stone model 8
2.4. The "General Migration model" 9

3. Coalescent Theory **11**
3.1. The structured coalescent 12
3.2. Coalescent effective population size 13
3.2.1. What does an effective population size say about the population? . 13
3.2.2. Derivation of a coalescent effective population size for spatially structured populations 14
3.2.3. Specific examples for the structured coalescent 16

II. Results: Between Stepping Stone and Island model **19**

4. Between Stepping Stone and Island model **21**
4.1. Weak migration 22
4.2. Arbitrary migration 28
4.3. Equal deviation to the GMM for IM and STST 32

5. Discussion **35**

III. Appendix **41**

A. Abstract (german) **43**

B. The coalescent tree for a Wright-Fisher population **45**

C. The Python code **49**
C.1. The VSC-4 cluster 50

D. Additional plots **51**
D.1. Weak migration 51
D.2. Arbitrary migration 51

Notations

Notations	Meaning
m_{ij}	probability that gene in i migrated from deme j
N	the deme size
N_T	$= N \cdot d$ the total population size
d	the number of demes
T_{ij}	the coalescence time for two lineages in deme i and deme j resp.
T_w	the "within-deme" coalescence time for two lineages, which are in the same deme
T_b	the "between-deme" coalescence time for two lineages, which are in different demes
a	the rate of migration going to the local (neighboring) demes
b	=1-a, i.e. the rate of migration going to the global (non-neighboring) demes
$\tilde{a}$	percentage of local migr., for which STST and IM deviate equally from the GMM

1. Introduction

The effective population size was first introduced by Wright (1931) and is a widely used tool in population genetics. Wang (2005) wrote a review about how the effective population size can be estimated from genetic data. There are various types of effective population size, each of which depends on a certain parameter like for example the variance in reproductive success, the level of inbreeding or the geographic structure in a population. In coalescent theory the effective population size can be calculated with the expected coalescence time of a pair of lineages in the genealogy of a considered population.

In unstructured populations the coalescent effective population size depends, amongst others, on the number of breeding individuals N_T, the ploidy of the population, the variance in reproductive success and the considered locus. In structured populations, other parameters that influence the coalescent effective population size N_e are the number of demes d and their size N (as they determine the population size) as well as the migration rate and pattern. Two fundamental models for a structured population are the island model, first stated by Wright (1943), and the stepping stone model, first stated by Kimura (1953). The difference between these models is that in the stepping stone model only neighboring demes exchange migrants, whereas in the island model this happens between each pair of demes. This will be explained in more detail in the section 2. The stepping stone and the island model both have restrictive assumptions concerning their general migration pattern. They are useful for analytical predictions, however, the biological reality is likely to lie in between them. In this thesis, I therefore assume that there was a certain fraction of migrants going to the neighboring (local) demes and the rest of the migrants would be going to the other $d - 3$ (global) demes. The continuum between the two models concerning the coalescent effective population size can be studied.

In the literature, studies like (Notohara 1990 and Cherry and Wakeley 2003) focus on exploring either of them separately in more detail. To my knowledge, there are no works, which study the continuum between the stepping stone and the island model concerning the coalescent effective population size in the setting of this thesis.

In this thesis, a more general migration model which can be found in (Nagylaki 1998, page 1600) will be adapted to investigate the continuum between the two models. The main focus will be on two different deme numbers, i.e. 8 and 20 demes. In the first sections, the used models and the concept of the effective population size will be introduced. It will be discussed how this concept connects to the expected coalescence time in the setting of a structured population and why it leads directly to the answer of the questions about the continuum between the island and the stepping stone model.

Part I.

Mathematical basics and derivation of the effective population size

2. Models

2.1. The (neutral) Wright-Fisher model

The most basic model to consider, when one is considering populations and how they change in time, is the Wright-Fisher model. It has the following underlying assumptions:

1. Isolation. The population is closed, i.e. isolated from the outside, meaning that there is no migration from elsewhere.

2. The generations are assumed to be discrete, so there are no overlapping generations. One generation is counted each breeding time.

3. No mutations, no selection and no migration. There is no mutation and no selection, all genes have the same fitness. Theoretically, there is assumed to be an infinitely large gene-pool, from which the genes for the descendent are picked.

4. Panmixia. The population is assumed to be panmictic, meaning that there is no subdivision and and each parental individual has the same probability of contributing a gamete to an individual that will breed in the next generation. The individuals are hermaphroditic, meaning that they can equally likely reproduce by selfing or by mating with another individual.

5. The population size is fixed, stays the same over time. It is often assumed to have $2N_T$ diploid individuals.

Summed up this means that the Wright-Fisher model is equal to considering selectively neutral, autosomal variants in a diploid, randomly mating hermaphrodite population of constant size, in which all genes have the same chance of contributing to the next generation. New individuals are formed each generation by random sampling (with replacement) of gametes produced by the parents. In a restricted setting like this, the only force influencing the gene frequencies is genetic drift (B. Charlesworth and D. Charlesworth 2010, page 199). The change is just by chance an therefore the gene frequencies will never be at equilibrium and one gene will finally take over and erase all the others. The population will become more homozygous. That is why it is a good starting point to define the effective population size, a measure of genetic drift, which will be defined later on in this thesis. Populations that follow the restrictions of the Wright-Fisher Model are said to be idealized.

2.2. The Island model

The island model, first mentioned by Wright (1931), is the most fundamental model of population structure. The only deviation from the Wright-Fisher model is that the population is subdivided into so-called islands or classes, which leads to violation of the panmixia - assumption. There is no internal structure assumed within the islands, so each of the subpopulations are assumed to be reproducing according to the Wright-Fisher population. Reasons for this type of population structure could for example be a large geographical distance between the subpopulations.

We assume to have d demes each of which contains $2N_i = 2N \ \forall i = 1, \ldots, d$ genes. A proportion m of these genes (in each deme) migrates to the other demes, such that each of the $d-1$ other demes receive a proportion of $m/(d-1)$ immigrants from this deme. Migration therefore is conservative, as each deme disperses m individuals and receives $m/(d-1) \cdot (d-1)$ individuals from the $d-1$ other demes in each generation.
Each island/subpopulation is assumed to be equidistant from all the other ones so there is total symmetry in the geography as well as in the migration matrix.

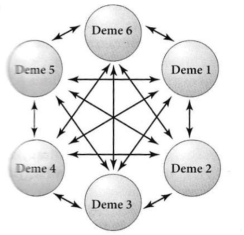

Figure 2.1.: The island model (after B. Charlesworth and D. Charlesworth 2010, page 317).

2.3. The circular Stepping Stone model

The other model under consideration is a version of the stepping stone model, which was first mentioned by Kimura (1953). Assume that the d subpopulations are arranged in a regular order in space and that the migration rate m_{ij} between each deme pair (i, j) depends on the distance between the considered demes. The simplest example is that the demes are arranged in a circular order in space and that each deme has two direct neighbors, see plot (2.2). Migration is assumed to happen to the two neighboring demes with rate $m/2$ in each direction resulting in a fraction of $m/2$ immigrants in each deme from its neighboring demes. All the demes have the same properties: N individuals and panmixia within the deme and they have the same number of neighbors. Within each deme the population fulfills the properties of the Wright-Fisher model. Therefore, migration is assumed to be conservative, like in the island model, leading to a symmetric migration matrix as well.

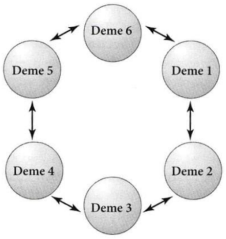

Figure 2.2.: The circular stepping stone model (after B. Charlesworth and D. Charlesworth 2010, page 317).

2.4. The "General Migration model"

In order to connect these two models, the general migration model was used, which was developed by Nagylaki (1998, page 1600). Like the name already induces, there is one deviation from the Wright-Fisher model, as there is no panmixia. The simplest form is based on a very similar setting to the island model or stepping stone model but with a more general migration rate. The deme sizes and properties are equal for each deme. Within each deme the population is ideal, so it has Wright-Fisher properties. Each generation a fraction of m_{ij} immigrants in deme i came from deme j, so that there is a total of $m = \sum_{j=1;i\neq j}^{d} m_{ij}$ immigrants in deme i. The migration matrix is given as follows:

$$M = \begin{bmatrix} m_{11} & \dots & m_{1d} \\ \vdots & \ddots & \vdots \\ m_{d1} & \dots & m_{dd} \end{bmatrix}.$$

3. Coalescent Theory

Before one could understand the coalescent effective population size N_e, it should be clear what coalescent theory is and what the term *coalescence* means.

The first person that described the standard coalescent is Kingman, which is why it is also often called the Kingman's coalescent, see (Kingman 1982, Dutheil 2020 page 5). Usually in population genetics, one looks at the current gene frequencies in the population and wants to investigate how these frequencies will change in the next time-step (forward in time). In coalescent theory, instead of asking what the gene frequencies will look like in the next generation, one rather poses the question of what happened in the history of this sample in order to create the current gene frequencies. The intention is to trace the ancestry of the genes along lineages back in time in order to find out more about the present and the future. It is a very powerful tool to investigate the efficiency of reproduction, the rate of genetic drift. It is also used in conservation genetics to detect hints for loss of heterozygosity early and prevent a species from extinction.

If we look at a part of the DNA for which the recombination rate is low, called haplotype, we can also form a gene tree of the considered population (Felsenstein 2005, page 464). Throughout this thesis, it is assumed that the population size is constant in time.

Generally, there is not a single *true coalescent-tree* for a set of individuals. The genealogy is different for each loci. It is not the goal to get the *true coalescent-tree*, but to see which of the predicted models fits best for which sample. If the sample and the haplotype is fixed, there is further always some level of uncertainty about which model is the *correct* one in that sense. Therefore, one has to look at the likelihood of the data under the possible coalescence histories. This approach is also used in the Python Code for Part II of this thesis.

If we look at a certain gene-tree and fix the population size N_T over time, it might happen that one ancestor has more than one offspring and another one has no offspring. When then picking a pair of lineages, which derived from this common ancestor, what happened when looking at the gene-tree backwards in time is that these two lineages merge at the time of this most recent common ancestor. This is called a coalescent event.

More about the standard coalescent can be found in Appendix B.

3.1. The structured coalescent

This chapter is based on (Notohara 1990 page 61f, Herbots 1994 page 16, Nordborg 2019 page 13f).

Consider a diploid population of $2N_T$ individuals, but from now on the individuals are not equivalent anymore, they are subdivided into classes. The classes are assumed to be discrete, so that each individual belongs to exactly one fixed class.

Remark 3.1.0.1. *In coalescence theory, it is common to write $2N_T$ instead of N_T, when talking about a diploid population. This way it is easier to switch to the haploid case.*

It is appropriate to use the structured coalescent, when each subpopulation is very large and there is minimal flow m_{ij} between the subpopulations. Then there will be a chance for a lineage to coalesce with one of the lineages of the same class, but not with a lineage from a different class. As we will see, structure will often increase the mean and variance of the time until the MRCA is reached.

There are certain conditions under which the considered process converges to a structured coalescent process. Let N_i denote the deme size. It is assumed that

1. the number of classes d
2. the relative subpopulation sizes $c_i = \frac{N_i}{N_T}$, $i = 1, \ldots, d$
3. and the number of backward migration events $M_{ij} = 2N_T \cdot m_{ij}$, $i \neq j$

remain constant as N_T approaches infinity. The second assumption assures that the probability of coalescence in the previous generation is $O(1/N_T)$ and the third assumption assures that the migration probabilities fulfill the same. Then when time is measured in units of N_T generations, the process converges to the so-called structured coalescent process, because in the limit $N_T \to \infty$ the event of coalescence of more than two lineages or of migration of more than one lineage or similar events can be neglected. In the structured coalescent, each pair of lineages in subpopulation i has a probability $1/c_i$ of coalescing in the previous generation and migrates independently to another subpopulation j at rate $M_{ij}/2$.

Remark 3.1.0.2 (The strong migration limit)**.** *The violation of the third assumption above would lead to the scenario of the strong migration limit. This means that the migration events happen on a much faster timescale. One can say that the timescale between coalescence and migration is separated. This leads to a panmixia-like behaviour.*

There are many possible ways to define classes and population structure such as sex-classes, age-classes and geographical-classes. In this thesis, the focus will be on the last one.

3.2. Coalescent effective population size

3.2.1. What does an effective population size say about the population?

The effective population size, first introduced by Wright (1931), is a tool in conservation genetics to quantify how a certain population is influenced by drift or inbreeding. To get an intuition of this measure, one should bear in mind that populations of different species, which all have the same size N_T, will not necessarily have the same rate of genetic drift. N_e is often regarded as the size of an idealized population, which undergoes the same effects of drift or inbreeding as the regarded one, whose census size is N_T. That is why it is seen as a measure of genetic drift. In an idealized population there is nothing influencing the change in gene frequencies but drift.

One should note that there is no such thing as a *general* effective population size, as it strongly depends on which property is considered. The formula for the coalescent effective population size can be derived respective to several properties or forces, which are influencing the considered population, such as the sex ratio, the change in population size over time and the offspring variance. Examples on how to do this can be found in (Nordborg and Krone 2002, Harmon and Braude 2010, and Nordborg 2019).

Example 3.2.1.1 (Offspring variance)**.** *Any variance in reproductive success which is larger than expectations (one child per parent) reduces N_e.*
The following explanation of this can be found in (Nordborg and Krone 2002, page 4):
In the Wright-Fisher model, the number of offspring contributed by each individual is 1*. Therefore, there will be N_T individuals in the offspring generation. This means that the expected coalescence time per gene pair will be N_T generations, as the probability of coalescence is $\frac{1}{N_T}$. Therefore, it is a good approach to set the timescale to N_T.*
Let now the variance be σ^2. In this case one would also need to adjust the expected coalescence time for a pair of lineages. When the variance increases, then there have to be more individuals that do not contribute offspring to the next generation (forward in time). If N_T stays the same for all generations, this automatically means that the resulting population comes from fewer parents. Backwards in time, this then means that there will be a greater number of coalescence events as there are more siblings in the considered offspring generation. The expected mean coalescence time decreases. Therefore, the expected mean coalescence time for this would be N_T/σ^2.

From the fact that offspring variance influences N_e negatively, it can be deduced that this is also the case for strong directional selection (Hare et al. 2011, page 440).

The big advantage of the effective population size is that by replacing the census size N_T by the effective population size N_e of a structured population, many of the formulas that were derived for the standard coalescent can be recovered. With this new timescale, under certain conditions, the structured coalescent process converges to the standard coalescent process, as described in section 3.1). The coalescent effective population size can then be defined by noting that the coalescent timescale in a diploid Wright-Fisher

population is $2 \cdot N_T$. Therefore, when the underlying model has the scaling N_c, one gets that $N_c/2 = N_e$ (Nordborg and Krone 2002, page 31). This is why it is so important to know the formulas for the expected time of coalescence for the standard Wright-Fisher model.

A disadvantage of this concept is that it only approximates the rate of genetic drift. As mentioned earlier in this section, it depends on a large variety of properties of the considered population like the type(s) of structure, the level of inbreeding in the population or the number of males and females (B. Charlesworth and D. Charlesworth 2010, page 216).

3.2.2. Derivation of a coalescent effective population size for spatially structured populations

The whole chapter is based on (Nagylaki 1998, page 1600).

In the following we will derive the formula for the coalescence time of two lineages in the *general migration model* from section 2.4, from which the expected mean coalescence time of two lineages in the island model can be deduced.

Let the migration matrix M be ergodic (i.e. irreducible and aperiodic) and denote T_{ij} as the number of discrete generations to coalescence for a gene in deme i with a gene in deme j. Let us look at the expected value of this and say w.l.o.g. that the two genes will coalesce in deme k. Then the following scenarios could happen in the next time-step:

1. that one of the two lineages migrates to a different deme first, say deme l, in which case they can not have coalesced. The coalescence time then is $(1 + E(T_{kl}))$.

2. that the two lineages end up in the same deme and coalesce. The coalescence time then is 1.

3. that the two lineages end up in the same deme and do no coalesce. The coalescence time then is $(1 + E(T_{kl}))$.

Summed up this leads to

$$E(T_{ij}) = \underbrace{\sum_{k,l:k\neq l} m_{ik}m_{jl} \cdot (1}_{=\sum_{k,l} m_{ik}m_{jl} - \sum_k m_{ik}m_{jk}} + E(T_{kl})) + \sum_k m_{ik}m_{jk} \cdot \underbrace{(\frac{1}{2N_k} + (1 - \frac{1}{2N_k}) \cdot E(T_{kk}))}_{=1+E(T_{kk})\cdot(1-\frac{1}{2N_k})} =$$

$$= 1 - \sum_k m_{ik}m_{jk} + \sum_{k,l:k\neq l} m_{ik}m_{jl} \cdot E(T_{kl}) + \sum_k m_{ik}m_{jk} + \sum_k m_{ik}m_{jk} \cdot E(T_{kk}) - \sum_k m_{ik}m_{jk} \cdot \frac{E(T_{kk})}{2N_k} =$$

$$= 1 + \sum_{k,l} m_{ik} m_{jl} \cdot E(T_{kl}) - \sum_{k} m_{ik} m_{jk} \cdot \frac{E(T_{kk})}{2N_k} \tag{3.1}$$

The second equality holds due to $\sum_j m_{ij} = 1$.

By now defining the global and local mean of the coalescence times $\bar{T}$ and $\bar{T^0}$ by

$$\bar{T} = \sum_{i,j} \nu_i \nu_j T_{ij} \tag{3.2}$$

$$\bar{T^0} = \beta \cdot \sum_{i} \frac{\nu_i^2}{c_i} T_{ii},$$

In the above formula we have that

1. $\beta = (\sum_i \nu_i^2 / c_i)^{-1}$ fulfills that $\beta N_T = N_e$, the migration effective population size by Nagylaki (1980)
2. ν_i is the stationary distribution of the coalescent process, which is a Markov Chain. It is defined by $0 < \nu_i < 1$, $\sum_i \nu_i = 1$, $\nu^T M = \nu^T$, where M again is the migration/transition matrix of the underlying model
3. T is the transposition operator for matrices.

We then get that the expected value of $\bar{T}$ is

$$E(\bar{T}) = 1 + \sum_{i,j} \nu_i \nu_j \sum_{k,l} m_{ik} m_{il} E(T_{kl}) - \sum_{i,j} \nu_i \nu_j m_{ik} m_{jk} \cdot \frac{E(T_{kk})}{2N_k} =$$

$$= 1 + E(\bar{T}) + \frac{E(\bar{T^0})}{2N_T \beta}$$

By the definition of β and by rearranging the equation, the above becomes

$$E(\bar{T^0}) = 2N_e.$$

3.2.3. Specific examples for the structured coalescent

Island model

The island model, which was introduced in chapter 2, is an appropriate model to use when one wants to investigate how migration/mutation and drift work together. Migrations and mutations make the subpopulations more different and drift makes the subpopulations more similar (as drift works towards one random allele in the respective subpopulation).

The following derivation of the expected mean coalescence time can be found in (Slatkin 1991 page 169f and B. Charlesworth and D. Charlesworth 2010 page 317f). Consider again equation (3.1). Recall that in the island model, all demes are equal and all deme pairs are equal. Therefore, the expected mean coalescence time for the island model for small m (the weak migration limit) can be derived by using the facts that

1. $T_{ii} = T_w$ and $T_{ij} = T_b$ for $i \neq j$
2. $m_{ij} = m/(d-1)$ for $i \neq j$

The formula for the general migration model 3.1 can then be simplified for T_w and T_b respectively to

$$1 \approx (2m + \frac{1}{2N_e})T_w - 2mT_b$$

and

$$1 \approx 2mT_b - \frac{2m}{(d-1)}T_w - \frac{2m(d-2)}{d-1}T_b$$

Solving this one gets the following expressions for T_w and T_b:

$$T_w \approx 2dN_e = 2dN$$

$$T_b \approx T_w + \frac{d-1}{2m}$$

Remark 3.2.3.1. *Note that when d is large, the within- and between-deme coalescence time are very similar.*

We get the mean coalescence time by noting that the probability of drawing two genes from the same subpopulation is $1/d$, as we just need the second gene to pick the same subpopulation as the first. The probability of drawing them from different demes is the converse probability $1 - 1/d = (d-1)/d$. In case they come from different demes, they would have to pick the same deme for coalescence, and they do so with probability $1/d$.

$$E(T_c) = \frac{E(T_w)}{d} + \frac{E(T_b)(d-1)}{d}.$$

Remark 3.2.3.2. *The above aligns with equation* (3.2)*. Note that the migration matrix* M *for the island model is symmetric and doubly-stochastic (which means that* $\sum_i m_{ij} = \sum_j m_{ij} = 1$*). Due to the definition of* ν*, one can see that in this case* $\nu_i = 1/d$*.*

Definition 3.2.3.1.

One then gets the following expression for the expected mean coalescence time for small m

$$E(T) \approx 2dN(1 + \frac{(d-1)^2}{4Nmd^2}) \tag{3.3}$$

By Tajima (1983, page 442), it is known that $E(T_c) = 4N_e(1-1/n)$ holds in randomly mating populations. In this case, $n = 2$ as we sample two lineages from the population. So we have the following formula for the effective population size:

$$N_e \approx dN(1 + \frac{(d-1)^2}{4Nmd^2})$$

As B. Charlesworth and D. Charlesworth (2010, page 318) noted, equation (3.3) aligns with the fact from Slatkin (1991) that the average time until two lineages, which are in different demes, are in the same deme is $\frac{(d-1)^2}{dm} \approx \frac{d}{2m}$.

Remark 3.2.3.3. *The above formula holds for a sample size of two genes. But Takahata (1991) even derived this formula for a sample size of n:*

$$E(T) = 4dN(1 + \frac{(d-1)^2}{4Nmd^2})(1 - \frac{1}{n})$$

The circular Stepping Stone model

In the circular stepping stone model the demes are assumed to built a circle such that the system is invariant and each deme has two neighboring demes with which they are exchanging migrants.

The following derivation is based on (Slatkin 1991, page 170). When looking at two lineages, which are located in different demes, one can also think of them as being i demes apart. The average time for such a pair of lineages to end up in the same deme is $(d-i)\cdot i/2m$, see (Feller 1957). As in the circular stepping stone model there is no differentiation between the demes, one can say that from that time on, it takes $T_{ii} = T_w$ time-steps until they coalesce. Therefore, the average coalescence time of the two genes that are i demes apart from each other is

$$E(T(i)) = T_w + \frac{(d-i)\cdot i}{2m} \tag{3.4}$$

where $T_w = 2dN$.

Averaging over i, the formula for the mean expected coalescence time can be derived by $E(T) = \sum_{i=0}^{d-1} E(T_c(i))/d$. The effective population size can be obtained with the formula $E(T_c) = 2N_e$ (Tajima 1983, page 442) as before.

Part II.

Results: Between Stepping Stone and Island model

4. Between Stepping Stone and Island model

In this part of the thesis, the continuum between the stepping stone and the island model will be studied. The research questions are the following:

What if there was a fraction of **a** of the migration happening on a local basis between the demes and a fraction of **b** of the migration happening on a global basis? Which model would be more appropriate to use for which values of **a** and **b**?

In order to answer this questions, the formulas (3.3), (3.1) and the - according to (4.1) - averaged results of the formulas (3.4) will be considered for the case of weak migration. Furthermore, the population genetics simulator *msprime* was used to explore the continuum for arbitrary migration. More about this simulator can be found in (Kelleher and Lohse 2020) as well as in Appendix C. The code that was used within this thesis is also available on Github, via the Link: *https://github.com/krissiliv/Coalescent-Theory.*

To my knowledge, this has not been done before in this setting. If I am mistaken, I appreciate any hint, which can be sent to my E-mail address *a01503122@unet.univie.ac.at.*

Alcala et al. (2019) studied the set of all possible migration motifs, amongst others, the stepping stone and the island model. They looked at effective population size in different migration motifs and assumed $d \leq 4$. The main focus was on 4 demes and how the effective population size changes with different migration motifs, such as the island model and the stepping stone model. Their results suggest that even for larger deme numbers the within-subpopulation diversity generally increases with the number of connections of an average subpopulation.

4.1. Weak migration

Consider again formula (3.1) from (Nagylaki 1998, page 1600):

$$E(T_{ij}) = 1 + \sum_{k,l=1}^{d} m_{ik}m_{jl} \cdot E(T_{kl}) - \sum_{k} m_{ik}m_{jk} \cdot \frac{E(T_{kk})}{2N_k}$$

The migration probabilities m_{ij}, for the island model as well as for the stepping stone model, fulfill the following:

1. $m_{ii} = (1 - m)\ \forall i = 1, \ldots, d$ for both models
2. $m_{i,i\mp1} = m_{i\mp1,i} = \frac{m}{f}\ \forall i = 1, \ldots, d$, where $f = 2$ in the stepping stone and $f = d-1$ in the island model
3. and else $m_{ij} = 0$ for the stepping stone and $m_{ij} = \frac{m}{d-1}$ for the island model.

As a next step assume that a fraction of **a** of the migration is local, i.e. to (the) two direct neighboring demes and a fraction of **b=1-a** of the migration is global, i.e. to the $d - 1 - 2$ other demes. The migration rates in this case are the following:

1. $m_{ii} = (1 - m)\ \forall i = 1, \ldots, d$ stays the same
2. $m_{i,i\mp1} = m_{i\mp1,i} = a \cdot \frac{1}{2} \cdot m\ \forall i = 1, \ldots, d$ is the local migration rate
3. and else $m_{ij} = b \cdot \frac{1}{d-3} \cdot m$ is the global migration rate

Remark 4.1.0.1. *Note that in the case of the island model the neighboring demes can just be any two demes, as there is nothing inducing distance.*

In order to get the global mean coalescence time, formula (3.2) from chapter 3.2.2 can be used:

$$\bar{T} = \sum_{i,j} \nu_i \nu_j T_{ij}$$

From remark 3.2.3.2 it is known that for migration conservative models, which is the case for island and stepping stone model, $\nu_i = \frac{1}{d}$. Using this the above formula reduces to

$$\bar{T} = \sum_{i,j=1}^{d} \frac{1}{d^2} T_{ij}. \tag{4.1}$$

Remark 4.1.0.2. *From now on the island model will be called IM, the general migration model GMM and the stepping stone model will referenced as STST.*

The expected coalescence times were derived with Python using the libraries sympy and numpy. The goal of this section is to explore the continuum between the IM and the STST by using the following formulas.

In the following plots

1. for the IM the plot shows the result of the formula (3.3).
2. for the GMM the plot shows the result of the formula (3.1) with the adapted migration rates.
3. for the STST the plot shows the - according to (4.1) - averaged results of the formulas (3.4).

The deme size was chosen to be $N = 1000/d$ (which was also assumed throughout this thesis).

First, look at the expected mean coalescence time dependent on the migration rate m and assume $d = 8$. The results for $a = 1$ and $a = 2/(d-1)$ look as follows.

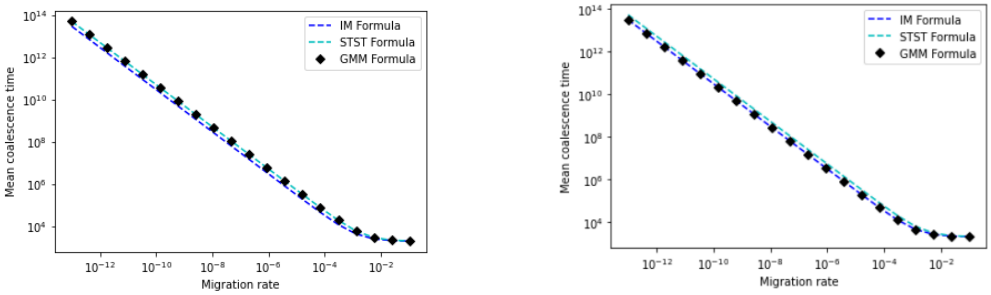

Figure 4.1.: For this plot, the parameter $d = 8$ was chosen and the migration rate ranges from $m = 10^{-14}$ to $m = 1$. It shows the results of the formulas for the expected mean coalescence times of the IM in dark blue, the STST in light blue and the GMM with $a = 1$ (left) and $a = 2/(d-1)$ (right) in black, all dependent on the migration rate m. With a small deme number like in this case, the expected mean coalescence times for the IM and STST are similar and therefore the light blue and the dark blue line are close together.

On the left side of plot (4.1) 100 % migration to the neighbors (i.e. two arbitrary demes in the case of the IM) and 0 % migration to the global demes is assumed. Therefore, the black line, the GMM, represents the light blue line for the STST. On the other hand, $a = 2/(d-1)$ leads to IM-like results, represented in the dark blue line, see right side of plot 4.1. This is because in the case of the IM, migration goes equally to two demes as well as to the (d-1) other demes (excluded the considered deme).

In the following plots, the results will be shown dependent on the parameter a. First, m was chosen to be $m = 10^{-14}$. This m is the smallest possible m such that sympy does not lead to errors. When choosing a small deme number, say $d = 5$, the results are given by plot 4.2. The other parameters were chosen like before.

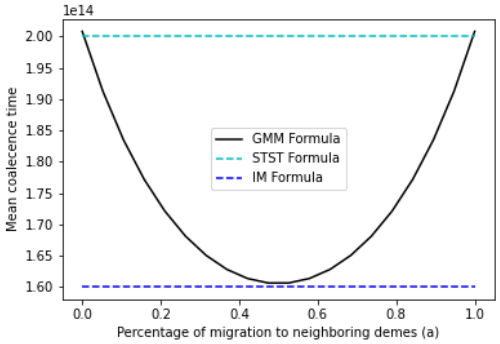

Figure 4.2.: For this plot, the deme number $d = 5$ and the minimal migration rate $m = 10^{-14}$ were chosen in order to explore the behaviour from a mathematical point of view. As the previous plot, it shows the results of the formulas for the expected mean coalescence times of the IM, the STST in light blue and the GMM in black, but in this case dependent on the percentage of local migration a.

To get IM results, $a = 2/(d-1) = 0.5$ is needed and for STST results $a = 1$ is needed, as before. For the case where there is no migration to the neighboring demes ($a = 0$), the GMM gives the same results like the STST. This indicates that the two neighboring demes were just exchanged by the two demes that normally would not receive any migrants in the STST. This can happen because in the STST with $d = 5$ all the demes have equal properties as for this case the number of neighboring demes is the same as the number of global demes. For a larger number of demes, say $d = 8$, the plot is given by 4.3.

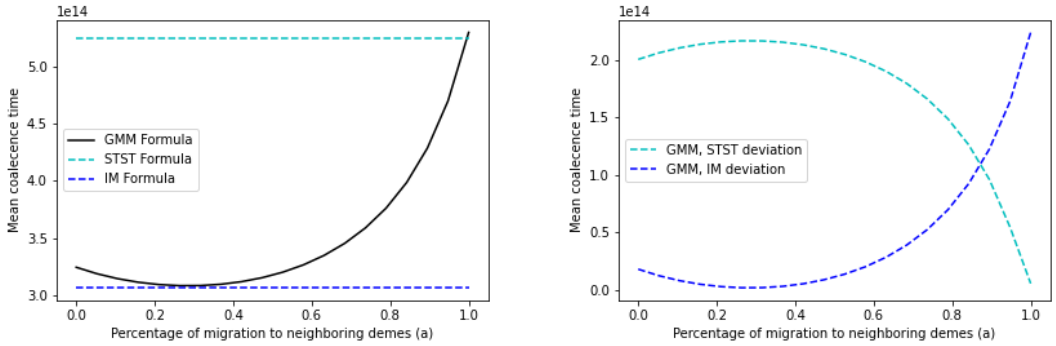

Figure 4.3.: Here, $d = 8$ and $m = 10^{-14}$ were chosen. The plot on the left side again shows the results of the formulas for the expected mean coalescence times of the IM, the STST and the GMM dependent on the percentage of local migration a. The plot on the right side shows the respective deviation of the line for the IM or the line of the STST to the line for the GMM.

To get IM results, $a = 2/(d-1) = 0.\dot{2}$ is needed. The highest deviation from the STST to the GMM is for $a = 2/(d-1) = 0.\dot{2}$, where the local migration goes to two demes and the global migration goes to the other 5 demes. This is exactly the case of the IM. If a is smaller than that, the expected mean coalescence time is larger than for the IM. This is because in that case, the local demes are more excluded from the migration than the demes that are further away. Therefore, when choosing a pair of lineages, where one of them is in a neighboring deme to the other, the expected coalescence time of that pair is larger than for a pair, where both lineages are in the same deme. This increases the average of all coalescence times, the expected mean coalescence time, and therefore gives a larger number than for the IM.

If $a > 2/(d-1) = 0.\dot{2}$ increases, the deviation to the IM gets larger and the deviation to the STST gets smaller. So the more important the two (neighboring) demes get, the more appropriate it becomes to use the STST. When looking at the right plot, the deviation dependent on a, one can note that for $a \approx 0.875$ the deviation is equal. With this choice of a, about 85 % of the migration goes to the neighboring demes and 15 % to the rest of the demes. This means that $2 \cdot N_T \cdot 0.875 \cdot 1/2 \cdot 10^{-14}$ immigrants go to each of the neighboring demes. The robustness of this result is studied for different migration rates and deme sizes in section 4.3.

If m is larger, say $m = 10^{-4}$ the picture looks pretty similar, except for the fact that the expected mean coalescence time is much smaller. This is due to the increased migration rate. Generally, it can be noted that the expected coalescence time depends stronger on the between-deme coalescence time than on the within-deme coalescence time, see D.1 in the Appendix D.1.

In the next step, the VSC-4 was used, which is a cluster provided by the Technical University of Vienna. More about this can be found in Appendix C.1. The following part will be focused on 20 demes in order to make the Python code more feasible. It should show how the behaviour and precision of the IM and STST change, when the deme size is very large. Plot (4.1) looks as follows for $d = 20$ (plot 4.4).

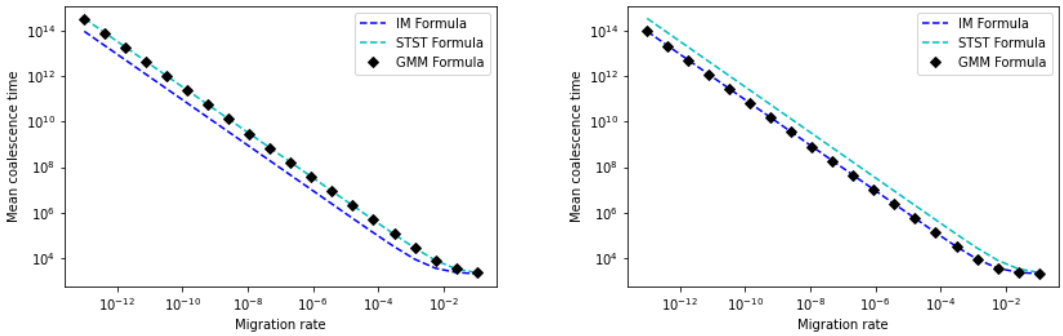

Figure 4.4.: In this plot the deme number is assumed to be $d = 20$ and the migration rate again ranges from $m = 10^{-14}$ to $m = 1$. It shows the results of the formulas for the expected mean coalescence times of the IM, the STST and the GMM with $a = 1$ (left) and $a = 2/(d-1)$ (right) dependent on m.

Compared to plot (4.1) it can be noted that the line for the IM and the STST drifted apart. This is due to the fact that with a larger number of demes, the circle of demes in the circular STST gets bigger, so that there is a larger average distance between the demes. When choosing a large distance i between the two demes where the lineages are drawn from, their expected coalescence time is large compared to the expected coalescence time for two lineages in neighboring demes. Therefore, for more demes the expected coalescence time for the IM increases much slower than for the circular STST, see plot 4.8.

Plot (4.3) but with 20 demes looks as follows:

Remark 4.1.0.3. *Note the change of the scaling of the y-axis in the following plots.*

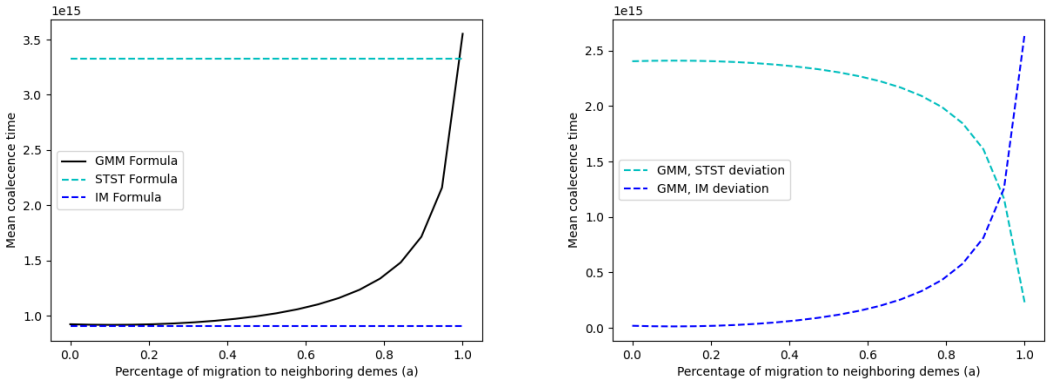

Figure 4.5.: The right hand plot shows the formula results for the IM, STST and GMM (left) and the respective deviation of the line for the IM and the line of the STST to the GMM. The deme number $d = 20$ with $m = 10^{-14}$ was chosen. As can be seen, the black line does not overlap with the light blue line for $a = 1$. This error does not appear in any other parameter setting that was explored (see for example 4.8), which is why the reason for this seems to be an approximation error in the code.

Here as well you can see on the right plot that the lines for the deviation to the GMM drifted apart. For the STST, there is now a much higher deviation to the GMM for small a. For the IM, the opposite is the case. Their deviation to the GMM for this larger deme number is equal at about 0.9, see section 4.3. For larger m, the plot looks similar.

4.2. Arbitrary migration

In the last subsection, weak migration was assumed and therefore the formula for IM as well as for STST and GMM were used. The goal of this section is to study the continuum between the expected coalescence time of the IM and STST with the help of the population genetics simulator *msprime*, which is further described in (Kelleher and Lohse 2020). More about this can be found in Appendix C. To compare the results of the Python code to the results of the IM and STST formulas, they were as well included in the plots.

Remark 4.2.0.1. *It is worth noting that the Python code computes the actual values by averaging over all the possible genealogies and lineage pairs, whereas the formulas compute the expected values. More about this can be found in Appendix C.*

Consider again formula (3.1) from (Nagylaki 1998, page 1600):

$$E(T_{ij}) = 1 + \sum_{k,l=1}^{d} m_{ik}m_{jl} \cdot E(T_{kl}) - \sum_{k} m_{ik}m_{jk} \cdot \frac{E(T_{kk})}{2N_k}$$

with the adapted migration rates from section 4.1.

The deme size was again chosen to be $N = 1000/d$.

The plot for the expected mean coalescence time dependent on the parameter m with the parameter $d = 8$ looks very similar to the plot of the formulas, see Appendix D.2, plot D.2.

Now fix $m = 10^{-4}$ and look at the expected mean coalescence time dependent on a in plot 4.6.

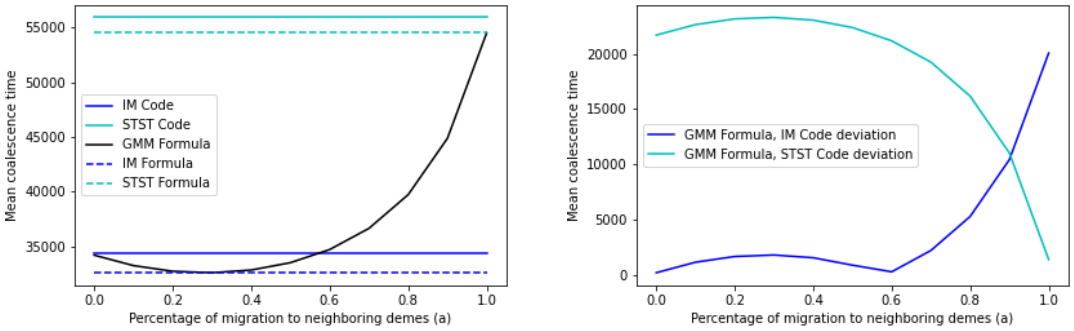

Figure 4.6.: For the left plot the results for the formulas (3.3), (3.1) with the adapted migration rates and the - according to (4.1) - averaged results of the formulas (3.4) were evaluated, as in the previous section. Apart from that the results from *msprime* for the IM and the STST were added to the plot. The right hand side represents the deviation of the IM and STST lines resulting from the *msprime* simulator to the GMM. Here the parameters $d = 8$ and $m = 10^{-4}$ were chosen.

In the left plot, the dashed lines, as before, represent the formulas for the IM and the STST. The GMM overlaps them for $a = 2/(d-1)$ and $a = 1$ respectively, as for example in plot 4.3. The difference is that we have a higher migration rate in plot 4.6.

There seems to be a small deviation to the actual values (from the Python-simulation). Calculating the absolute value of the deviation one gets the plot on the right-hand-side. The difference between the expected mean coalescence time of the IM code and STST code and the one predicted from the respective formulas are small. It seems like the green line is shifted downwards meaning that the actual expected mean coalescence time is greater than the formulas suggest. For smaller m the behaviour stays the same. On the right plot, one can note that again for a between 0.85 and 0.9 the deviation to the continuum, i.e. the GMM-formula with the adapted migration rates, is equal for both models.

If there are more migrants per generation, e.g. if $m = 0.99$, one gets plots 4.7.

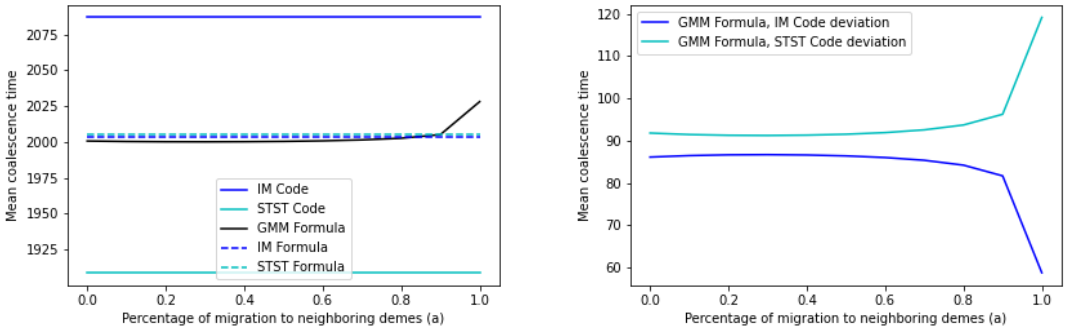

Figure 4.7.: In the previous plot, but with a higher migration rate, $m = 0.99$, the expected mean coalescence times according to the formulas lie between the values, which are obtained by the *msprime* simulator. Because of the large migration rate, the line for the GMM appears flatter than before. It only starts increasing for some $a > 0.8$, but the exact point is not clear due to the chosen step size.

Compared to plot 4.6 the lines for the GMM, STST and IM formulas get closer together and end up between the lines for STST and IM that the code gives. The black curve for the GMM flattens as m increases, because if there is a very high rate of migration it is decreasingly important whether or not migration is more on a local or global basis. It can also be noted, especially on the right plot(s), that there is less deviation between the GMM and the two lines for the IM code and the STST code. This is due to the more similar results for the IM and STST.

Another difference to plot 4.6, where $m = 10^{-4}$, is that the expected mean coalescence time for each of the models is now much closer to the expected within deme coalescence time $E(T_w)$. When there are lots of migrants it is much more likely that two random lineages from different demes end up in the same deme after the next time-step. Last but not least something unusual can also be noted in plot 4.7: The line for the expected mean coalescence time according to the IM code is above the line for the STST. In Appendix D.2, there is also the plot D.4 for $m = 1$, where the two lines switched back to their usual positions.

For consistency reasons, the assumption on the deme number $d = 20$ in section 4.1 will be continued. However, it is worth noting that it would also be possible to compile them for larger deme numbers in VSC-4, but this would increase the run-time of the code.

Plot (4.6) and plot (4.7), which represent the (expected) mean coalescence time of the models dependent on the parameter a, look as follows for d=20 (4.8):

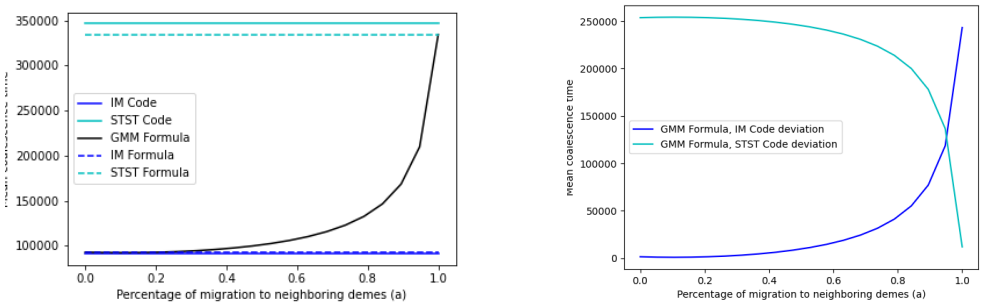

Figure 4.8.: As in the previous plots, the results of the formulas and the *msprime* simulator respectively for the expected mean coalescence time for each of the three models were evaluated. The deme size is chosen to be larger, i.e. $d = 20$ and the migration rate is $m = 10^{-4}$. Comparing the right-hand side of this plot to the previous plots one can note that the two lines now intersect for a larger percentage of local migration $\tilde{a}$.

The respective plot to $d = 20$ and $m = 0.99$ looks analogously to the above.

The percentage of deviation between the GMM and the IM and STST respectively seems to be less. It already can be noted from the above pictures that there seems to be a quasi linear change of the expected mean coalescence time, when the deme numbers change. In order to get a clearer picture of the changes due to the parameter d, the next plot 4.9 shows the expected mean coalescence time dependent on the deme number d with parameter $m = 10^{-4}$.

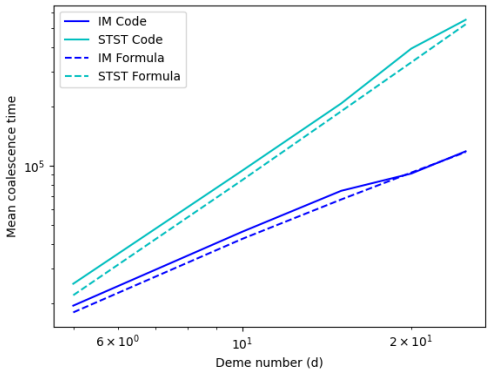

Figure 4.9.: On this plot one can see the expected mean coalescence time of the IM code (*msprime*) and the IM formula results as well as the same for the STST. The chosen parameters are $d = 5, 10, 15, 20, 25$ and $m = 10^{-4}$. The limitations in compiling the plots concerning the deme size can be found in Appendix C.1.

This plot suggests that the expected mean coalescence time increases linearly for in-

creasing deme sizes. Furthermore, the increase for the STST is much steeper than for the IM, which indicated once more that the between-deme coalescence time of the STST is much more sensitive to the deme number than the one of the IM.

4.3. Equal deviation to the GMM for IM and STST

In the previous section, especially the plots 4.3, 4.5, 4.6 and 4.8 suggested, that from a certain percentage $\tilde{a}$ of local migration on, it is more appropriate to use the STST than the IM. In order to study the robustness of these results, the following shows how much the two lines for the deviation from the GMM to IM and STST respectively differ from each other when a changes.

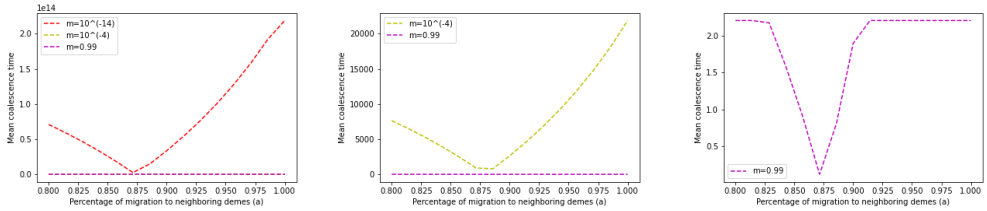

Figure 4.10.: Here the parameter $d = 8$ was chosen and the formula results were taken into account. In the right hand sides of plots 4.3, 4.5, 4.6 and 4.8 one can see the deviation of the IM to the GMM in dark blue and of the STST to the GMM in light blue. This plot shows the difference between the light blue and the dark blue line in these plots. Therefore, the minimum represents the intersection point between the dark blue and the light blue line, which is the point from which on the use of the STST can be justified. Up to that point, it is more reasonable to use the IM, as we saw in section 4.1 an 4.2.

As it can be seen, for $d = 8$ for the expected coalescence time calculated with the formulas, the point from which on it is better to use the STST model is robust concerning the migration rate. When looking at the results for the *msprime* calculation of the expected mean coalescence time, plot 4.10 looks as follows:

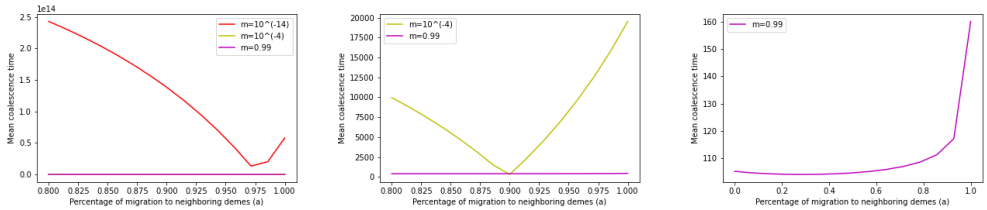

Figure 4.11.: Here the parameter $d = 8$ was chosen and the *msprime* results were taken into account. The minimum in these plots again represents the point from which on the use of the STST can be justified. Up to that point, it is more reasonable to use the IM.

Note the different sections on the a-axis. Even though the parameters are chosen equally as in plot 4.10, the percentage $\tilde{a}$ decreases for increasing migration rates. This makes sense, because when there are generally more migrants per generation, then there is a higher chance of coalescence in the STST. As we noticed above, the expected mean coalescence time of the IM is more robust to migration. Therefore, the deviation for the light blue line for the IM and the dark blue line for the STST in the above deviation plots 4.3 and 4.6 get closer together. When this happens, the intersection point $\tilde{a}$ decreases.

Comparing the line for $m = 0.99$ in plot 4.10 and plot 4.11 one can see an odd change of the shape of the line. The shape is in accordance with plot 4.7 even though the values are not the precise ones. This is due to the stochasticity of the results from the simulator *msprime*.

When the deme number is larger, say $d = 20$, the plot for the formulas looks as follows:

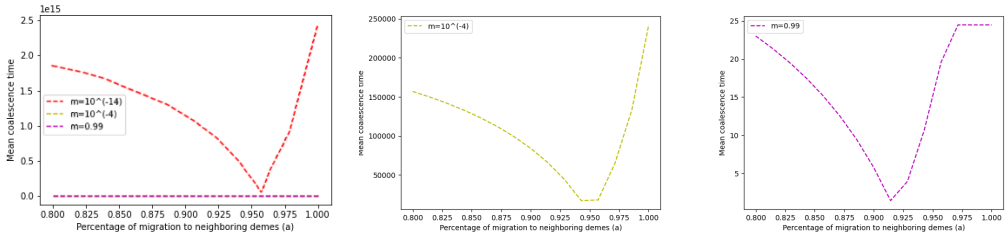

Figure 4.12.: For these plots the parameter $d = 20$ was chosen and the formula results were taken into account. Like the above plots, this one indicates that the point $\tilde{a}$ decreases as m increases.

When it comes to the deme number d, it can be noted that the point increases for increasing deme numbers. The reason is analogous to the behaviour due to migration: the IM is more robust to change of not only migration rates but as well deme sizes as the STST. The plots again suggest that the point $\tilde{a}$ decreases for increasing deme numbers. The following shows plot 4.11 but for $d = 20$.

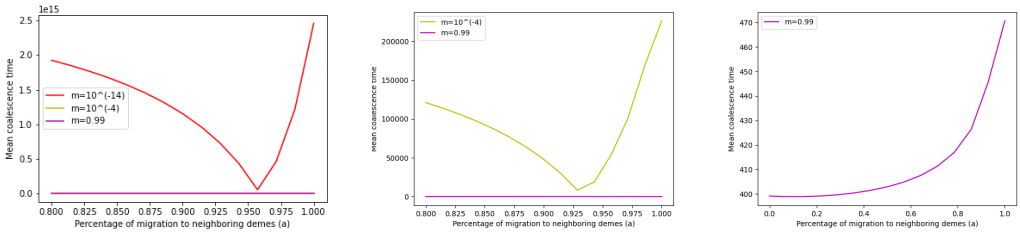

Figure 4.13.: For these plots the parameter $d = 20$ was chosen and the *msprime* results were taken into account. The right plot is the analogous to the right plot in 4.11 and has equal shape. The type of change of the point, from which on it is reasonable to use the STST, aligns with the observations from the above plots.

Note again that the percentage of local migration that is needed to justify the use of the STST decreases with an increasing number of migrants per generation. But this time, when comparing the 4.11 and 4.13 for $m = 10^{-14}$ respectively, the case where $d = 20$ shows a higher value of $\tilde{a}$. The shape of the line for $m = 0.99$ resulting from the simulations with *msprime* is very similar to the one in plot 4.11 and has an analogous explanation.

5. Discussion

The important thing about the effective population size is that it is useful to assess the influence of genetic drift on the genetic variability of the considered population. Unlike it sometimes appears in the literature, there is no such thing as *the* effective population size. In fact, this quantity can be defined in various ways depending on the proprieties that one wishes to investigate. There is for example the variance effective population size, the inbreeding effective population size or the eigenvalue effective population size, whose definitions can, for example, be found in (Harmon and Braude 2010, page 126f).

While the variance effective population size relates to the change of allele frequencies and the eigenvalue effective population size relates to the loss of unique alleles, the inbreeding effective size relates to the degree of pedigree inbreeding. In particular, the inbreeding effective population size can be calculated via the harmonic mean of the population sizes over time and focuses on the amount of pedigree inbreeding, which occurs when two gene copies in the offspring generation are identical by descent. It is therefore closely related to the coalescent effective population size.

In coalescence theory the effective population size is determined by the expected mean coalescence time of the respective population. In the case of structured populations, it is strongly influenced by the expected between-deme coalescence time rather than the within-deme coalescence time. Therefore, changes in the migration pattern influence the effective population size. In the island model migration is rather globally focused, which means that it goes to each of the demes at equal rate and there are no demes, which are excluded from this migration pattern. In the one-dimensional, circular stepping stone model it is locally focused, i.e. migrants only go to the neighboring demes. This indicates that for small deme numbers their coalescent effective population size is very similar and diverges for large deme numbers.

As it is known from (Wright 1943, Laporte and B. Charlesworth 2002, page 501), the effective population size increases, when there is population structure. In particular, this holds for stable population structure as in the models subject to this thesis. Unstable population structure can for example be associated with a bottleneck event, where the size of the entire population is substantially decreased, e.g. by a catastrophic event. This reduces the effective size by reducing genetic variability. Furthermore, the estimate of the effective population size used depends on the way of sampling - both across a genome, and from the species' range. Notably, when the island model approximation is not an accurate description of the connectivity within the population in question, sampling from just part of the range will have a large effect on the obtained estimate of the effective

population size.

The coalescent effective population size, which is determined by the expected mean coalescence time, is in general greater in the stepping stone model than in the island model with the same parameters. The numerical analysis demonstrates that the island model provides a good approximation unless most of the migration is local, whereas even a small amount of global migration can lead to a large deviation from the stepping-stone model approximation.

Furthermore, the results suggest that the difference between the coalescent effective population size of the stepping stone model and the one of the island model increases with the number of demes. This is most likely due to the fact that in the stepping stone model, the expected mean coalescence time of lineages, which are many demes apart, increases the average expected coalescence time. In the island model, all the demes are directly connected, which is why the increase is not that large. Therefore, in the continuum between these two models, the more the migration pattern resembles the stepping stone model, the more the effective population size increases. The point from which on the stepping stone model is a better approximation than the island model varies for different migration rates. The above results suggest that it decreases if m increases, which seems to be a good approximation for the parameters $N = 1000/d$, $d = 8, 20$ and $m = 10^{-14}, 10^{-4}, 0.99$.

This thesis focuses on a single neutral haplotype and thus does not consider the effects of selection and recombination. Furthermore, it is assumed throughout that the population size is fixed. For simplicity, I focused one the continuum between one-dimensional habitats with stepping stone properties and the island model. When comparing the island model to the two-dimensional stepping stone model, the difference would be less than in the above setting. This is because in two dimensions, the demes are much more connected than in one dimension, making the model more similar to the island model.

The review of Nordborg (2019) emphasizes that a very wide range of biological phenomena can be treated as a simple linear change in the time scale of the coalescent. Examples for this are overlapping generations, separate sexes and mating systems. B. Charlesworth, D. Charlesworth, and Barton (2003) explain that this is due to the fact that processes, such as movements of alleles between sexes or age classes, happen very fast compared to coalescence of alleles within demes, and the migration of alleles between demes.

There are different types of selection, each of which influence genetic diversity differently. Most often, directional selection is considered: it can be positive or negative. When considering two alleles, positive directional selection works towards one beneficial mutant. This is why it is often related to selective sweeps, where favoured alleles then become fixed in the population. This is similar to a bottleneck event, which reduces the effective

population size. Negative directional selection favours the non-beneficial allele and balancing selection benefits heterozygote advantage, as it works towards a balance between these two alleles (Wakeley 2009, page 170). This is why balancing selection increases the effective population size. Selection also affects nearby (neutral or nearly neutral) linked loci, which is called background selection. This type of selection can reduce genetic diversity, which especially happens in inbreeding populations (B. Charlesworth, Morgan, et al. 1993). As directional selection works towards one allele, it as well decreases genetic diversity.

In the island model, the effective selection coefficient (which is just the re-scaled selection coefficient for the equivalent panmictic population) is smaller than the actual selection coefficient. This explains why the product $N_e s_e$ is not altered by subdivision, despite the fact that subdivision increases N_e (Cherry and Wakeley (2003)). Background selection in the island model increases the population differentiation due to genetic structure by decreasing the within-deme diversity. Balancing selection increases the expected coalescence time compared to the expected within deme coalescence time, so that when these two act together, the expected within-deme coalescence time is reduced, while the between-deme coalescence time is unaffected (B. Charlesworth, Nordborg, et al. 1997, B. Charlesworth, D. Charlesworth, and Barton 2003 page 113). This leads to a reduction of the effective population size.

Bibliography

Alcala, Nicolas et al. (2019). "Coalescent theory of migration network motifs". In: *Molecular biology and evolution* 36.10, pp. 2358–2374.

Charlesworth, Brian and Deborah Charlesworth (2010). *Elements of evolutionary genetics*. Roberts and Company.

Charlesworth, Brian, Deborah Charlesworth, and Nicholas H Barton (2003). "The effects of genetic and geographic structure on neutral variation". In: *Annual Review of Ecology, Evolution, and Systematics* 34.1, pp. 99–125.

Charlesworth, Brian, MT Morgan, and Deborah Charlesworth (1993). "The effect of deleterious mutations on neutral molecular variation." In: *Genetics* 134.4, pp. 1289–1303.

Charlesworth, Brian, Magnus Nordborg, and Deborah Charlesworth (1997). "The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations". In: *Genetics Research* 70.2, pp. 155–174.

Cherry, Joshua L and John Wakeley (2003). "A diffusion approximation for selection and drift in a subdivided population". In: *Genetics* 163.1, pp. 421–428.

Dutheil, Julien Y (2020). *Statistical Population Genomics*. Springer Nature.

Feller, W (1957). *An Introduction to Probability Theory and its Applications, vol. 1, second edn.(1957); vol. 2, first edn.(1966)*.

Felsenstein, Joseph (2005). "Theoretical evolutionary genetics joseph felsenstein". In: *University of Washington, Seattle*.

Hare, Matthew P et al. (2011). "Understanding and estimating effective population size for practical application in marine species management". In: *Conservation Biology* 25.3, pp. 438–449.

Harmon, Luke J and Stanton Braude (2010). "Conservation of small populations: effective population sizes, inbreeding, and the 50/500 rule". In: *An introduction to methods and models in ecology, evolution, and conservation biology*, pp. 125–138.

Herbots, Hilde Maria Jozefa Dominiek (1994). "Stochastic models in population genetics: genealogy and genetic differentiation in structured populations". PhD thesis. Queen Mary University of London.

Kelleher, Jerome and Konrad Lohse (2020). "Coalescent simulation with msprime". In: *Statistical Population Genomics*. Humana, New York, NY, pp. 191–230.

Kimura, Motoo (1953). "'Stepping stone'model of population". In: *Annual Report of the National Institute of Genetics Japan* 3, pp. 62–63.

Kingman, John Frank Charles (1982). "The coalescent". In: *Stochastic processes and their applications* 13.3, pp. 235–248.

Laporte, Valérie and Brian Charlesworth (2002). "Effective population size and population subdivision in demographically structured populations". In: *Genetics* 162.1, pp. 501–519.

Nagylaki, Thomas (1980). "The strong-migration limit in geographically structured populations". In: *Journal of mathematical biology* 9.2, pp. 101–114.

– (1998). "The expected number of heterozygous sites in a subdivided population". In: *Genetics* 149.3, pp. 1599–1604.

Nordborg, Magnus (2019). "Coalescent theory". In: *Handbook of Statistical Genomics: Two Volume Set*, pp. 145–30.

Nordborg, Magnus and Stephen M Krone (2002). "Separation of time scales and convergence to the coalescent in structured populations". In: *Modern developments in theoretical population genetics: The legacy of gustave malécot* 194, pp. 2–40.

Notohara, M (1990). "The coalescent and the genealogical process in geographically structured population". In: *Journal of mathematical biology* 29.1, pp. 59–75.

Slatkin, Montgomery (1991). "Inbreeding coefficients and coalescence times". In: *Genetics Research* 58.2, pp. 167–175.

Tajima, Fumio (1983). "Evolutionary relationship of DNA sequences in finite populations". In: *Genetics* 105.2, pp. 437–460.

Takahata, Naoyuki (1991). "Genealogy of neutral genes and spreading of selected mutations in a geographically structured population." In: *Genetics* 129.2, pp. 585–595.

Wakeley, John (2009). *Coalescent theory: an introduction*. 575: 519.2 WAK. Roberts and Company.

Wang, Jinliang (2005). "Estimation of effective population sizes from data on genetic markers". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459, pp. 1395–1409.

Wright, Sewall (1931). "Evolution in Mendelian populations". In: *Genetics* 16.2, pp. 97–159.

– (1943). "Isolation by distance". In: *Genetics* 28.2, p. 114.

Part III.

Appendix

A. Abstract (german)

Die effektive Populationsgröße N_e wird oft als die Größe einer idealisierten Population angesehen, die den gleichen Auswirkungen von Drift oder Inzucht ausgesetzt ist, wie die betrachtete Population. Auf dem Gebiet der Koaleszenztheorie entspricht sie der erwarteten Koaleszenzzeit eines zufälligen Linienpaares in dieser Population. Diese Masterarbeit konzentriert sich auf die effektive Populationsgröße der Koaleszent Theorie im Kontinuum zwischen dem eindimensionalen, kreisförmigen Stepping Stone und dem Insel Modell. Bei konstanter Zensusgröße, wird das Kontinuum bezüglich dem Prozentsatz an lokaler und globaler Migration erforscht. Bezug genommen wird dabei auf verschiedene Migrationsraten sowie auf verschiedene Deme Anzahlen. Die Formel für die effektive Populationsgröße für ein generelleres Migrationsmodell wird, mit abgeänderten Migrationsraten, verwendet, um das Verhalten der effektiven Populationsgröße im Kontinuum zu beschreiben. Um die erwartete Koaleszenzzeit des Stepping Stone und Insel Modells zu repräsentieren, wurden mathematische Formeln sowie der populationsgenetische Simulator *msprime* verwendet. Wann es passend ist das Insel Modell zu verwenden und wann es mehr Sinn macht, das Stepping Stone Modell zu verwenden, hängt von dem Prozentsatz an lokaler Migration ab. Die Resultate in dieser Masterarbeit lassen annehmen, dass dieser Prozentsatz an lokaler Migration, von welchem an der Gebrauch des Stepping Stone Modells gerechtfertigt ist, sinkt, wenn man die Migrationsrate vergrößert. Das Intervall, in welchem das Insel Modell eine bessere Annäherung darstellt, ist im Allgemeinen viel größer als jenes, in welchem das Stepping Stone Modell verwendet werden sollte. Außerdem scheint die erwartete Koaleszenzzeit bei größerwerdender Deme Anzahl in etwa linear zu wachsen, sowohl für das Stepping Stone als auch für das Insel Modell.

B. The coalescent tree for a Wright-Fisher population

The Wright-Fisher model has two very important properties:

1. As populations fulfilling the Wright-Fisher model are idealized, the equations for the expected coalescence time can be easily derived for this case.

2. It is useful to understand the basics of coalescence theory, which this thesis is based on.

The first person that described the standard coalescent is Kingman, which is why it is also often called the Kingman's coalescent, see (Kingman 1982), (Dutheil 2020, page 5).

The following is based on (Nordborg and Krone 2002, page 2ff).

Let us now consider $2N_T$ individuals from which two distinct lineages are sampled at random. Denote by T_{ji} the transition probability from the state, where we have j genes to the state where we have $i < j$ genes. We can calculate the probability of coalescence for these two lineages after t generations. Note that the probability of coalescence in the previous generation is $1/(2N_T)$, as there are $2N_T$ parents that could be chosen for the second gene but just one of them is also the parent of the first gene.

$$P(T_{21} > t) = (1 - \frac{1}{2N_T})^t$$

If $2N_T$ gets larger and we re-scale time to the measure $t = 2N_T\tau$, time can be treated as continuous and the geometric distribution can be exchanged by the appropriate exponential distribution.

$$P(T_{21} > 2N_T\tau) = P(\frac{T_{21}}{2N_T} > \tau) = (1 - \frac{1}{2N_T})^{\lfloor 2N_T\tau \rfloor} \xrightarrow[N_T \to \infty]{} e^{-\tau} = P(\tilde{T_{21}} > 1).$$

It follows that

$$\frac{T_{21}}{2N_T} \to^d \tilde{T_{21}}$$

After letting $N_T \to \infty$ the mean $E(\tilde{T}_{21})$ becomes 1, so coalescence for a the last pair of genes takes on average one timestep = $2N_T$ generations. The variance of the coalescence time is $(2N_T)^2$ generations.

Analogously the probability of one coalescence event in the next generation $P(T_{n(n-1)})$, if there are n lineages sampled from the $2N_T$ lineages, can be calculated. There are n over 2 possibilities to pick a pair of lineages from a population of n lineages. Each of these pairs has probability $1/(2N_T)$ of picking the same parent, because the second lineage just has to pick the same parent like the first.

$$P(T_{n(n-1)}) = \binom{n}{2} \cdot \frac{1}{2N_T} + O(N_T^{-2}) \approx \frac{n(n-1)}{2 \cdot 2N_T} = \frac{n(n-1)}{4N_T}.$$

When there are n lineages present at the time, the mean of the time until the next coalescence event (of any pair of lineages) is $E(T_{n(n-1)} = 4N_T/(n(n-1))$.

The above also shows that initially, when the number of lineages n is still large, coalescence events occur on a much faster rate than later, when n is small. When $n = 2$, we can calculate that the mean time of coalescence of this last pair of lineages is $4N_T/(2 \cdot 1) = 2N_T$.

The probability that more than one pair of lineages coalesces is negligible and if the number of lineages changes, it will only decrease by 1 almost surely (i.e. the probability of this event goes to 1 as $N_T \to \infty$).

Therefore,

$$P(\frac{T_{n(n-1)}}{2N_T} > \tau) = (1 - \frac{n(n-1)}{2N_T}\frac{1}{2})^{\lfloor 2N_T\tau \rfloor} \to_{N_T \to \infty} e^{-\frac{n(n-1)}{2}\tau} = P(\tilde{T}_{n(n-1)} > 1).$$

It follows that

$$\frac{T_{n(n-1)}}{2N_T} \to^d \tilde{T}_{n(n-1)}$$

After letting $N_T \to \infty$ the mean of this exponentially distributed random variable $E(\tilde{T}_{n(n-1)})$ becomes $\frac{2}{n(n-1)}$ and the variance $V(\tilde{T}_{n(n-1)}) = (\frac{2}{n(n-1)})^2$.

The two below results can be found in (Tajima 1983, page 442). Consider the coalescent timescale and let T_{n1} be the time to the MRCA, when there are n lineages initially present. Summing up over all these expected times, we can calculate formally

$$E(\tilde{T}_{n1}) = \sum_{k=2}^{n} E(\tilde{T}_{k(k-1)}) = \frac{2}{n(n-1)} + \frac{2}{(n-1)(n-2)} + \cdots + \frac{2}{2} = \cdots = 2(1 - \frac{1}{n})$$

This corresponds to

$$E(T_{n1}) = 4N_T(1 - \frac{1}{n}) \tag{B.1}$$

generations. By the same argument we get for the variance of the length of the gene tree, that

$$V(\tilde{T}_{n(n-1)}) = \sum_{k=2}^{n} V(\tilde{T}_{k(k-1)}) = \sum_{k=2}^{n} \frac{(2)^2}{n(n-1)^2}$$

Which, in generation time, is

$$V(T_{n(n-1)}) = (4N_T)^2 \sum_{k=2}^{n} \frac{(2)^2}{n(n-1)^2}$$

C. The Python code

The population genetics simulator *msprime* with Python was used in order to compare the results from the code with the results from the GMM formula (3.1), to see how accurate the formula is. A close description of *msprime* can be found in (Kelleher and Lohse 2020), which was also the starting point for the Python code used for the plots.

The code computes the results for the within-deme and the between-deme coalescence time respectively and then averages over the results from each replicate. A random genealogy is taken each replicate with the use of a random seed. It finally computes the mean of all these values.

In order to make a feasible program, it is appropriate to first consider the influence of the parameters onto the model. The plot C.1 represents the variance, the code-results and the formula-results for the expected mean coalescence time of the IM, when 10^4 replicates are taken for the code. The number of replicates is the number of times that the code creates a random genealogy, which will then be taken into account for the average coalescence time. The population is assumed to have $N_T = 1000$ individuals, just like in the above results. The line IM-formula represents the results from 3.3. The variance is scaled-down by a factor of 10^{10}.

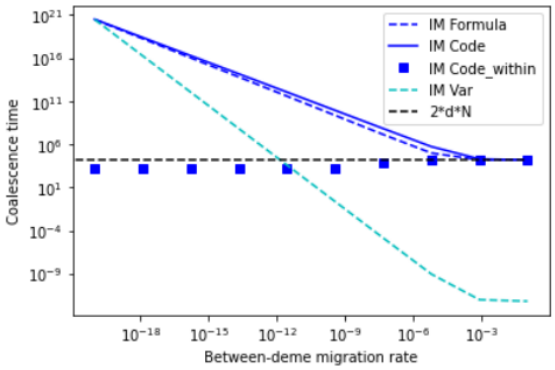

Figure C.1.: Python results for the IM with d=8.

One can see that the variance decreases as m goes to 1, meaning that the number of replicates can be decreased for increasing m. Also the within-deme coalescence time T_w converges to the case, where there is just one large Wright-Fisher population with expected coalescence time $2dN$ (which is the case for $m = 1$). It is close to that value already for very small m. This aligns with plot D.1: The between-deme coalescence time

T_b influences the expected coalescence time much more.

The same picture for the STST is represented by plot C.2.

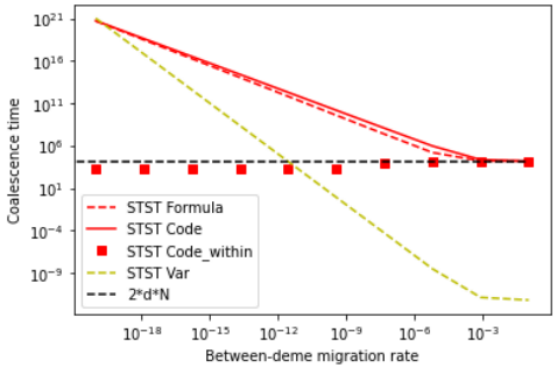

Figure C.2.: Python results for the STST with d=8.

The two above pictures persist for increasing deme numbers d. However, for a large deme number, running the code with enough replicates according to above becomes infeasible for a normal computer. This is why for $d = 20$ the VSC-4 was used, which is described in Appendix C.1.

C.1. The VSC-4 cluster

The VSC-4 is a cluster provided by the Technical University of Vienna. More about it can be found in *https://www.tuwien.at/tu-wien/aktuelles/news/news/der-vsc-4-oesterreichs-neuer-supercomputer/*.

It was used to compute the plots for the parameter $d = 20$.

For the settings that were used for this thesis, there are too many recursion equations of formula 3.1 from section 3.2.2 to solve in the case of $d = 50$. Therefore, it gave the following error-message C.3:

```
    for monom, coeff in f.items():
RecursionError: maximum recursion depth exceeded while calling a Python object
Job ended at
Sat Mar 19 23:54:27 CET 2022
```

Figure C.3.: Error-message of VSC-4 when compiling the Code with d=50.

To my knowledge this issue can be solved and the plots in the above sections could also be compiled for a larger number of demes than $d = 20$. For this thesis, however, it was sufficient to examine the change of the continuum between the STST and IM due to deme size.

D. Additional plots

D.1. Weak migration

In the following plot you can see the line for the expected within-deme coalescence time and expected between-deme coalescence time of the IM (D.1).

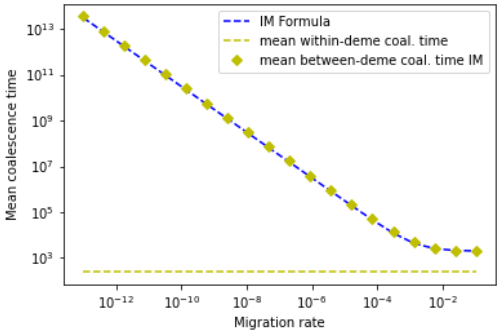

Figure D.1.: IM formula results for $d = 8$.

The expected within-deme coalescence time is very small compared to the expected between-deme coalescence time. On the plot it can be seen that the between-deme coalescence time influences the mean expected coalescence time more than the within-deme coalescence time.

D.2. Arbitrary migration

The expected mean coalescence time dependent on the parameter m with the parameter $d = 8$ can be seen in plot D.2.

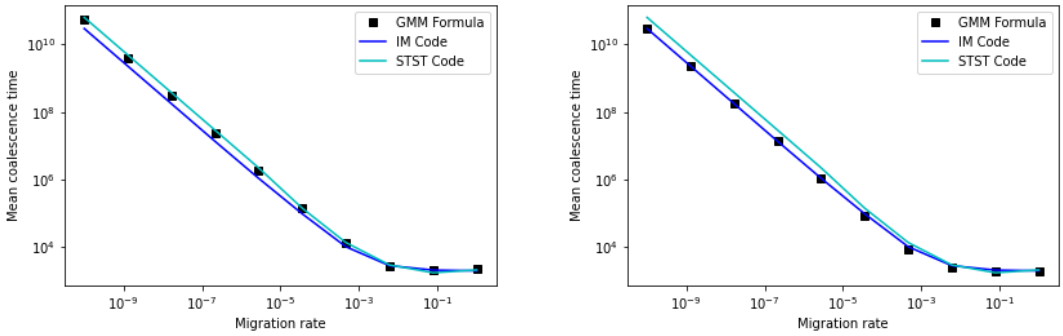

Figure D.2.: Python results for d=8 with $a = 1$ (left) and $a = 2/(d-1)$ (right).

Just like when using the formulas for STST and IM, the GMM formula gives the results for the IM code, when $a = 2/(d-1)$ and for the STST code, when $a = 1$.

Now, plot (D.2) is re-compiled with the parameter $d = 20$ resulting to plot (D.3).

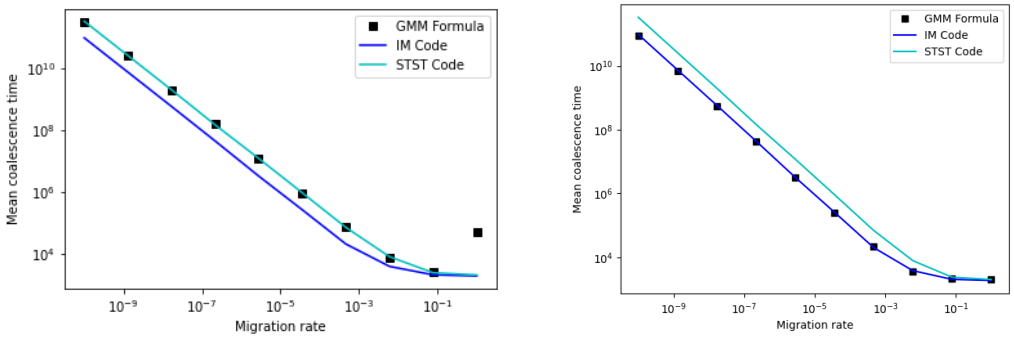

Figure D.3.: Python results for d=20 and $a = 1$ (left) and $a = 2/(d-1)$ (right).

Comparing plot D.2 and D.3 the following can be observed: When the deme number d shrinks, the three lines get closer and closer together. This shows the fact that the IM and the STST are more similar for small deme numbers. Conversely, the IM shows a much smaller mean coalescence time than the STST for larger deme numbers leading to the red and the blue line drifting apart for increasing deme numbers. The explanation for this is analogous to the one for plot (4.4).

The following plot D.4 shows the GMM, the IM code and formula and the STST code and formula combined where $m = 1$.

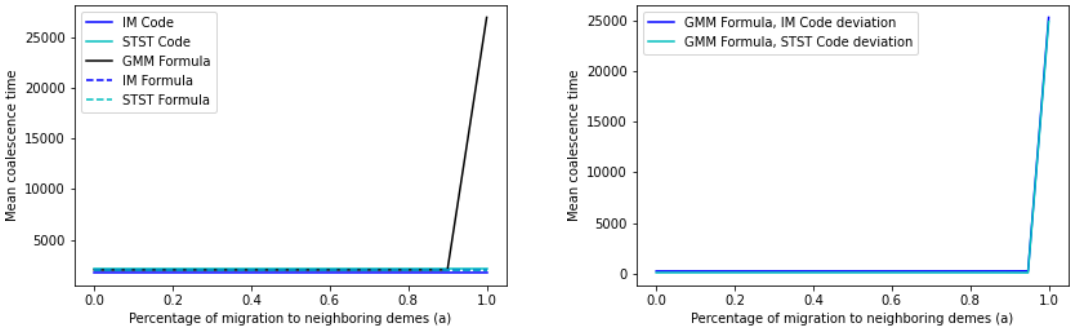

Figure D.4.: Python results for d=8 and m=1.

One can see that there is less deviation between the lines in this case, they almost overlap.

Remark D.2.0.1. *Note the sudden change of the green line for a close to* 1. *This comes along with an error message from sympy. For* $m = 0.99$ *this error is much smaller, see 4.7.*