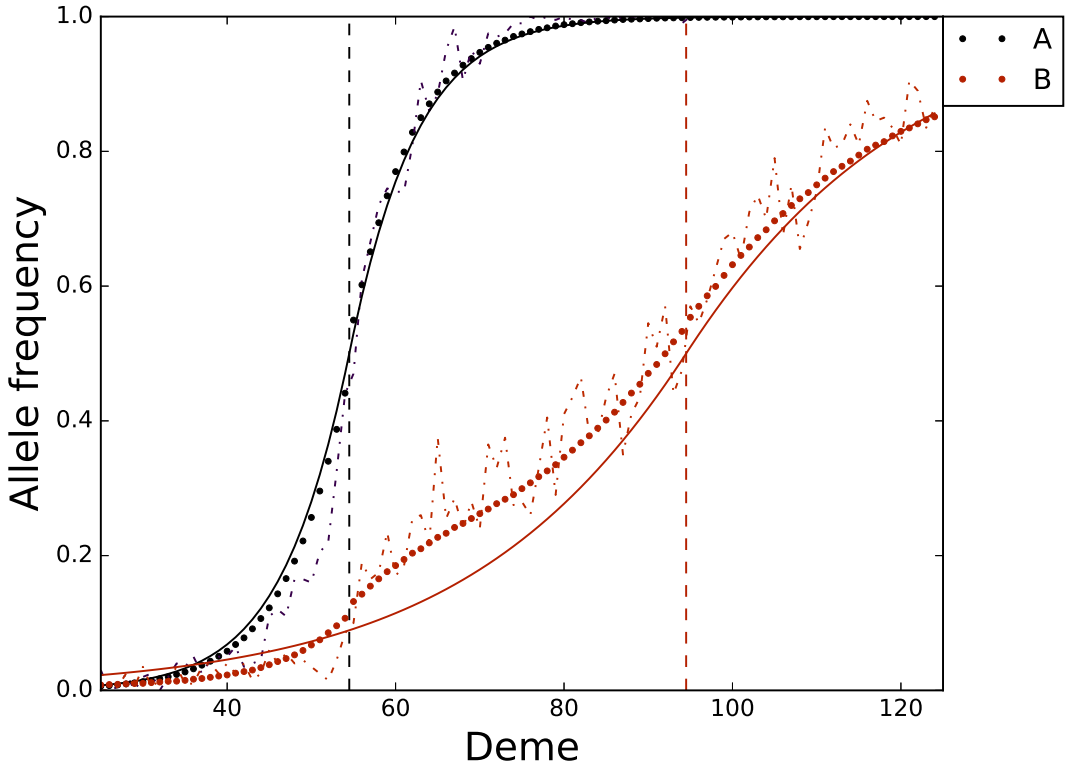
CHALMERS



Clines resulting from selection-dispersal balance in the presence of multiple environmental changes

Master's thesis in Complex Adaptive Systems

OSKAR FRIDELL

Department of Physics, CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2017

Master's thesis 2017

Clines resulting from selection-dispersal balance in the presence of multiple environmental changes

OSKAR FRIDELL

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2017

Clines resulting from selection-dispersal balance in the presence of multiple environmental changes.
OSKAR FRIDELL

Supervisor: Marina Rafajlovic, Department of Physics
Examiner: Bernhard Mehlig, Department of Physics

Master's Thesis 2017
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Allele frequencies of the alleles A (black) and B (red) at the loci $\mathcal{A}$ and $\mathcal{B}$ over the habitat. The habitat consists of 150 demes, with 100 diploid individuals in each deme. The solid lines are the theoretical predictions of the cline shapes assuming that only one locus is under selection. The dots are the simulated average allele frequencies, and the dash-dotted lines represent one snapshot (last iteration) of the allele frequencies. The dashed vertical lines indicate the environmental changes corresponding to the two loci. The selection parameters are $s_A = .1$ at locus $\mathcal{A}$ and $s_B = .01$ at locus $\mathcal{B}$. The dispersal distance is $\sigma_D = 3$. Recombination rate is $r = .01$ between the loci. The cline corresponding to the weaker selected locus $\mathcal{B}$ is being pulled towards the cline of the more strongly selected locus. For further information, see caption of figure 3.6.

Typeset in LaTeX
Gothenburg, Sweden 2017

Abstract

A hybrid zone is a spatial area in which two genetically and phenotypically different populations of the same species (so called divergent ecotypes) meet and produce viable and fertile offspring. Hybrid zones are commonly observed in nature. Examples include hybrid zones found in the sea snail *Littorina saxatilis* on the Swedish West coast. This species has formed two divergent ecotypes (a crab-exposed and a wave-exposed one) as a response to the adaptation to the local environmental conditions, primarily the presence and absence of crabs, respectively. However, other environmental conditions, such as the presence of additional species' interactions in a given ecosystem, may have contributed to the formation and maintenance of the hybrid zones observed. The spatial positions of the environmental transitions corresponding to different environmental factors may or may not coincide with each other. Yet, most theoretical studies on hybrid zones assume only a single environmental transition in the habitat in question (Gay et al. 2008; but see Slatkin, 1975). These studies have shown that allele frequencies at the locus under selection exhibit a sigmoid-like function of the spatial position in the habitat (so called genetic clines). However, when analyzing empirical data, it is of crucial interest to answer: Is the hybrid zone in question subject to multiple, spatially separated, environmental transitions? To answer this question, it is necessary to understand how the presence of multiple environmental transitions is reflected in spatial patterns of genetic variation at the loci under selection. Slatkin (1975) considered a model with two environmental transitions, assuming that two loci are under the selection pressures imposed by these environmental changes. His main finding concerns the relationship between the average cline centres at the two loci under selection in dependence of the recombination rate between the loci, and the distance between the two environmental transitions. However, these results are insufficient to answer under which conditions it becomes possible to detect multiple environmental transitions in the habitat based on empirical genetic data on allele frequencies. To answer this question it is necessary to estimate the probability distributions of the cline centers at the two loci, and the corresponding mixture distribution. Notably, in the case of a single environmental transition the distribution of cline centers is unimodal. By contrast, and as shown in this thesis, in the case of two environmental transitions the resulting mixture distribution may be bimodal. A bimodal mixture distribution is a direct indicator of the existence of two environmental changes. Using a test of bimodality, I estimate the critical value of the distance between the two environmental transitions above which a bimodal pattern of the mixture distribution of cline centres emerges. These results can be further improved by assessing the patterns of neutral loci linked to the loci under selection, and the bias these may introduce when inferring the presence of multiple environmental transitions in the hybrid zone in question.

Keywords: cline, *Littorina saxatilis*, hybrid zone, genetic drift, bimodality.

Acknowledgements

I would like to thank my supervisor Marina Rafajlovic for her support and thorough consideration.
I would also like to thank Roger Butlin, Kerstin Johannesson and Anja Westram for inspiration and input.

Oskar Fridell, Gothenburg, September, 2017

Contents

1

Introduction

Speciation is the separation process of one species into two or more distinct species (Howard & Berlocher, 1998). Speciation can occur in many different ways. Examples include two populations of one species becoming geographically separated, or a subset within a population developing some special adaptation to its environment. An important goal when studying speciation is to fully understand how new species form and how they are maintained. Also important is to understand why sometimes new species *do not* form, even though many necessary prefactors are present.

There are two common definitions of what a species is (Barton & Hewitt, 1985). They both originate in the definition that species are "groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups" (Mayr, 1942). The controversy lies in the definition of "reproductively isolated". One definition of reproductively isolated is based on the amount of gene flow[i]; if the gene flow between two groups of one species is limited so that the two groups do not fuse together into one, then the two groups are reproductively isolated (Mayr, 1982). Another definition is that if two groups are reproductively isolated, then no fertile offspring can be formed by individuals of the two groups (Key, 1981). The latter definition will be used in this Master's Thesis. The primary reason for this is that this definition is less ambiguous than the definition based on gene flow.

From the definition of a species above, one species has become two when the gene flow between two groups is completely gone and impossible to reestablish. The study of the state where gene flow is low could therefore inform about what is necessary for a new species to form. When gene flow between two groups is low the two groups could belong to different variations, genotypes, of the same species. The zones in which different genotypes of the same species meet and produce offspring are known as *hybrid zones* (Barton & Hewitt, 1989). Hybrid zones are found for many species, for example *Bombina bombina* (Szymura & Farana, 1978), *Campylorhynchus rufinucha* (Selander, 1964) and *Littorina saxatilis* (Janson, 1983). *L. saxatilis* is the inspirational species for this Master's Thesis.

Littorina saxatilis is a marine snail species that resides in the coastal areas near Gothenburg. They also reside in Greenland, Spain and the UK among other places (World Register of Marine Species, 2017). This species has been used as an example species in various studies of hybrid zones, both in simulation and in the field (Janson, 1983; Panova, Hollander, & Johannesson, 2006). What makes *L. saxatilis*

[i]Gene flow refers to the long-time dispersal of genes.

interesting is that it resides almost exclusively where water meets land (Janson, 1983). This makes the effective habitat nearly one dimensional. In some places, for example near Gothenburg, the frequencies of some alleles[ii] found in *L. saxatilis* exhibit cline[iii] patterns along some coastal areas where the snail resides in a hybrid zone (Johannesson, Rolan-Alvarez, & Ekendahl, 1995).

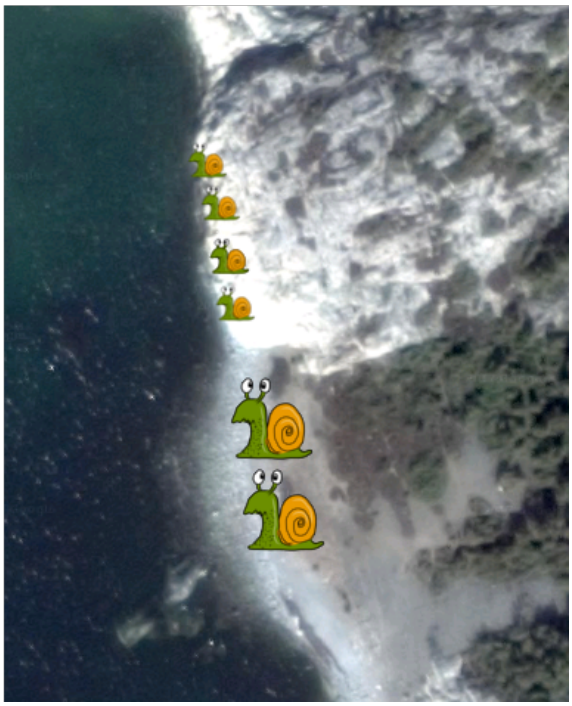


Figure 1.1: *L. saxatilis* can have different sizes depending on whether it lives among rocks or on a cliff. This is due to differences in natural selection in these different environments (Janson, 1983).

One concrete example of a cline that is formed is the one concerning the size of the snail (Janson, 1982). Individuals of *L. saxatilis* tend to have thicker shells, larger overall size and smaller aperture if they live among rocks. The contrary is true for individuals living on cliffs (Wilding, Butlin, & Grahame, 2001). When living on a cliff a smaller snail has a bigger chance of withstanding waves, while among rocks a bigger snail has a bigger chance of surviving a crab attack. See figure 1.1 for a schematic description. This thesis uses this species as a model to study the clines occuring through this habitat, using one locus[iv] in the genome for a trait under selection. While in reality there is generally more than one locus responsible for each biological trait of a species (Wilding et al., 2001), I make a simplification in this thesis and assume that only one locus underlies each biological trait of interest.

The clines observed in various species in nature are studied and explained by Barton

[ii] An allele is a variation of a gene.

[iii] A *cline* is defined as "a gradient of morphological or physiological change in a group of related organisms usually along a line of environmental or geographic transition" (Merriam Webster Online, 2017).

[iv] A locus (loci in plural) is a location on a chromosome.

and Hewitt (Barton & Hewitt, 1985). A typical cline shape can be seen in figure 1.2. In this figure, the mean allele frequencies of an allele, denoted by A, have been plotted as a function of distance from a given reference point in the habitat, averaged over time. The individuals of *L. saxatilis* are modelled as diploid[v], which they are in reality. In the simulations resulting in this figure they contain only information about one locus, denoted $\mathcal{A}$, which can contain one of the two alleles denoted a and A. The cline in that figure is the result of a balance between divergent selection and dispersal. A cline can also be formed due to selection against *hybrids* (Barton & Hewitt, 1985). The hybrid zones resulting from the clines formed due to selection against hybrids are also known as *tension zones* (Gay, Crochet, Bell, & Lenormand, 2008), and will not be considered in this thesis.

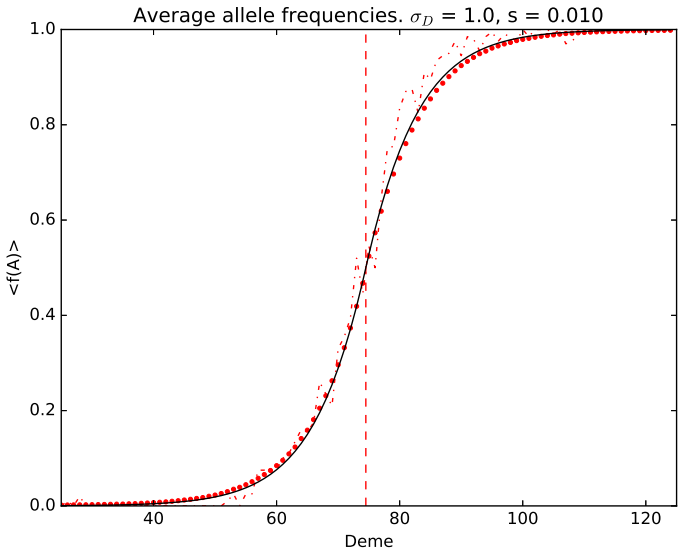


Figure 1.2: A typical cline shape. This cline is the result of a balance between dispersal of individuals and selection. In the model, an individual can have alleles A and a. It is assumed that selection favors allele A on the right side of deme 75, but favors a on the left side. Each deme contains 100 individuals and there are 150 demes in total. Further information about the model can be found in Appendix A.2.

The main aim of this Master's Thesis is to study how one can determine from the genome data of a species whether there is one environmental transition, or if there is in fact two or more environmental transitions. In the example described above there is only one environmental transition; the one between a cliffy environment and a rocky environment. Most studies have been carried out on models with one environmental transition. In reality, there could be a number of such transitions. These could be related to different kinds of changes in environment such as the absence or presence of predators, chemicals, other animals etc.

I carried out the simulations in this thesis using parameters that are reasonably close to those of *Littorina saxatilis*. I obtained the reasonable parameter values

[v]Diploidy means that each individual carries two copies of each chromosome.

from Roger Butlin and Kerstin Johannesson (personal communication, 2017), and they were in line with the ones suggested by Johannesson (Johannesson et al., 2010).

Simulations of the same kind as in this Master's Thesis were first studied by Slatkin (Slatkin, 1973, 1975). I repeated and improved some of Slatkin's results. This was possible primarily due to the large difference in computational power between 2017 and 1973-1975. However, the most important discovery in this Master's Thesis concerns the distributions of the cline centers when two loci are under selection. If the mixture distribution of the two distributions corresponding to the centers of the clines of the two traits is bimodal, then one can infer that there exist two environmental transitions.

In the next chapter, the models I created are presented. The details are discussed and motivated. Also, I discuss the methods and tests that I apply to the models.

In the Results chapter, plots of the results from the simulations are shown. In the Discussion and Conclusion chapter, the main conclusions from this thesis are presented.

2

Methods and Models

In this chapter, I present the models that I used. First, a model with only one locus under selection was used. This model was later generalized to account for multiple loci, with multiple environmental transitions. Although my model can handle a large number of loci and environmental transitions, I only went so far as to consider two loci. This is mostly due to the complexity of introducing even more loci.

The model with only one locus served as a verification that my modelling was correct. It also gave me understanding of the modelling, and helped me prepare for the development of the main model handling two loci.

2.1 Description of the model

This study concerns an implementation of a hybrid zone model, with a balance between dispersal and selection due to a discrete change in the environment. I will continuously refer to the species under consideration as *L. saxatilis*. This species was the inspirational species for this thesis.

I start by assuming that *L. saxatilis* lives in discrete patches, called *demes*, on a line. The demes are labelled by numbers from 1 to K as in figure 2.1.

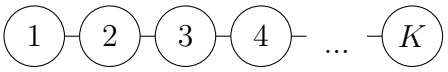


Figure 2.1: The K demes of the model.

In every deme, there exists in each generation N individuals. In almost all implementations in this Master's Thesis, I use $K = 150$ and $N = 100$. This value of N is derived from the observed number individuals being between 100 and 1000 individuals per square meter (Johannesson et al., 2010). The distance between every deme is assumed to be one meter. The individuals are diploid. I assume that there are two abrupt changes in environment (the model with only one change in environment is discussed in Appendix A.2). An abrupt change could be, for example, a transition between an environment with small rocks and an environment with large cliffs. I assume that for each such environmental transition, there is one locus under selection that alone underlies the trait relevant for that transition. This locus will have one allele that is advantageous on one side of the transition, and another allele that is advantageous on the other side.

The loci under selection are denoted $\mathcal{A}$ and $\mathcal{B}$, and can have alleles a or A and

a	B
A	B

Table 2.1: An example of two chromosomes in a diploid individual. In this thesis, each individual carries two copies of one chromosome that has length two (or length one in the one-locus case).

	aa	aA	AA
$x < Z_A/2$	1	$1 - s_A$	$(1 - s_A)^2$
$x \geq Z_A/2$	$(1 - s_A)^2$	$1 - s_A$	1

Table 2.2: Fitness values of individuals with the corresponding alleles.

b or B respectively. A diploid individual, then, could for example consist of the chromosomes as in table 2.1.

In the simpler case that was studied first, locus $\mathcal{B}$ was neglected so that only locus $\mathcal{A}$ was assumed to be under selection.

I assume that the life cycle of *L. saxatilis* consists of 2 phases:

1. Migration. I assume that the N individuals in every deme move independently on the one-dimensional line that they inhabit. The number of steps is drawn from a normal distribution with standard deviation σ_D and mean 0 (rounded to nearest integer since the demes are discrete). Typically, $\sigma_D = 3$ was used. The value suggested by Johannesson for *L. saxatilis* was $2 < \sigma_D < 10$ (Johannesson et al., 2010). If the movement exceeds the boundaries[i], the individual stays at the boundary. These type of boundaries I will call absorbing boundaries. All individuals are assumed to survive dispersal, and also there are no genetic differences in the dispersal patterns.
2. Breeding. In every deme a fitness value is calculated for each individual, which will determine the probability for that individual to generate offspring. The fitnesses are according to Table 2.2 for locus $\mathcal{A}$. The same respective table is used for locus $\mathcal{B}$ if we consider two loci, with a potentially different selection parameter s_B. Z_A and Z_B are the values on the x-axis where the environmental transitions take place for loci $\mathcal{A}$ and $\mathcal{B}$ respectively. Using the multiplied fitness values of loci $\mathcal{A}$ and $\mathcal{B}$ as weights, the individuals are sampled randomly with replacement to breed. The child receives one chromosome from the mother and one from the father. I assume that recombination (defined below) takes place with probability r per gamete.

The favored genotypes for different regions can be found in table 2.3. The size of the spatial distance $(Z_a - Z_b)$ between the environmental changes Z_A and Z_B is denoted as $2Z_0$. This is centered around $x = 0$, so that $Z_A = -Z_0$ and $Z_B = Z_0$. When only one locus is assumed to be under selection, I use $Z_A = \frac{K}{2}$.

[i]The boundaries occur at deme 1 and deme K, migration outside these limits is not allowed.

Region	$x < Z_A$	$Z_A \leq x < Z_B$	$Z_B \leq x$
Favored genotype	ab	Ab	AB

Table 2.3: Favored genotype as a function of x.

2.1.1 Recombination

In the breeding phase, an offspring receives one chromosome from the mother and one from the father. It could be that the chromosome the child receives from its mother (or father) is exactly the same as one of the chromosomes that parent had originally. However, this need not be the case if recombination occurs.

Recombination occurs with probability r per gamete. Say, for example, that a mother has genotype $ab|AB$. If $r = 0$, then one of these chromosomes, ab or AB, is simply selected randomly to be carried over to the child. But if $r > 0$, there is a probability r that we select the first allele from one chromosome and the second allele from the other chromosome (or the other way around). Say that we start copying from the second chromosome (in the simulations, there is a 50% chance that we start copying from either chromosome). We then have a $1 - r$ chance to take also the second allele from that chromosome, leaving us with AB to be carried over to the child. However, there is a chance equal to r that we instead take the second allele from the first chromosome, resulting in Ab being inherited by the offspring.

In reality, r depends greatly on the physical distance between the loci. There could be a number of neutral loci[ii] between the ones under selection. Since I consider at most two loci under selection in this thesis, r will serve as a way to artificially set a distance between the loci under selection on the chromosome without having to include the neutral loci that would lie in between in the model.

2.1.2 Mutation

Mutation occurs on each allele individually with probability μ, with $\mu \ll 1$. This takes place after breeding. The total number of alleles mutation can act on is

$$2 \cdot 2NK, \tag{2.1}$$

where the first 2 comes from the diploidy of the individuals and the second 2 is the number of loci under consideration per individual. This means that the total number of mutations per generation can be drawn from a Poisson distribution with parameter $4NK\mu$.

Mutation acts by changing one random allele from all the alleles in the habitat into the other possible allele for that locus. That is, if a random allele A is picked it is changed to a, and vice versa (or B would be changed into b and vice versa).

Mutation is necessary in this model, because without it we could reach extinction of an allele.

[ii]Neutral loci are loci that are not under direct selection.

2.1.3 Genetic drift

Genetic drift[iii] is implicitly included in the breeding phase since the parents generating offspring are drawn with replacement. Some previous studies have included genetic drift (Polechova & Barton, 2011; Durrett, Buttel, & Harrison, 2000) and others have not (Slatkin, 1973, 1975). In my Master's Thesis the inclusion of geneic drift is a desired attribute. This is mostly because of the discreteness of the model; it is impossible to know beforehand the total allele counts at the steady state of the model. So in order for the steady state to be more like the actual steady state one would observe in nature given the parameters, genetic drift is required.

2.2 Measurements

In this section, the measurements applied to the model are introduced. Most measurements were made every tenth generation during each simulation. The generations in which measures were made will be referred to as *sample generations*.

2.2.1 Width of a cline

When considering one locus, an important measurement is the *width* of a cline. The width is defined as one over the maximum slope of a cline. This is equal to (Polechova & Barton, 2011)

$$w = 4 \sum_{i \in Demes} p_i q_i, \tag{2.2}$$

where p_i is the frequency of allele A in deme i and q_i is the frequency of allele a (which is $1 - p_i$).

The average width of a cline will be denoted $w(\bar{p})$ and the width of the average cline[iv] will be denoted $w(\bar{p})$. These will, in general, be slightly different.

2.2.2 Cline fitting

I used curve fitting tools to fit cline shapes to simulated allele frequencies. The cline shapes were assumed to take one of four forms:

1. a sigmoid shape,
2. a sigmoid shape with the right tail transitioning into an exponential decay,
3. a sigmoid shape with the left tail transitioning into an exponential decay,
4. a sigmoid shape with both tails transitioning into exponential decays (Szymura & Barton, 1986).

[iii] Genetic drift is the change over time in the allele frequencies due to random sampling. Random sampling will over long time promote some alleles and depress others by chance.

[iv] The *average cline* refers to the average allele frequencies over time.

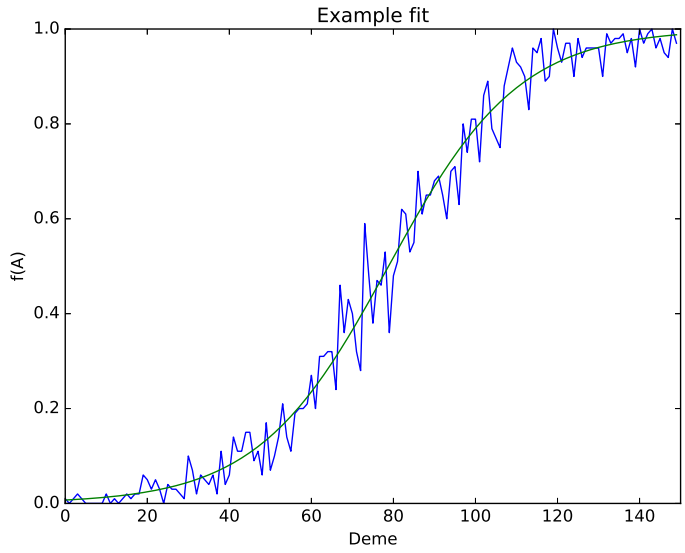


Figure 2.2: Example fit of a sigmoid function to some allele frequencies. The blue line are some simulated allele frequencies and the green line is the fit.

The actual theoretical shape of a cline is a *tanh* shape as derived in Appendix A.1. A sigmoid shape was used for fitting since it includes fewer parameters, and thereby is easier to fit. These two shapes are approximately the same (Guedj & Guillot, 2011). The fits were made either to the allele frequencies at a given generation, or to the temporal average of allele frequencies. An example fit to some allele frequencies can be seen in figure 2.2.

2.2.3 Linkage Disequilibrium

Linkage disequilibrium (LD) is the "nonrandom association of alleles at two or more loci" (Slatkin, 2008).

A measure of LD between two loci, denoted $\mathcal{A}$ and $\mathcal{B}$ below, that can have alleles a or A and b or B respectively, can be mathematically defined as

$$D = p_{AB} - p_A p_B, \tag{2.3}$$

where p_{AB} is the frequency of chromosomes that have both allele A in $\mathcal{A}$ and allele B in $\mathcal{B}$. p_A is the frequency of allele A and p_B is the frequency of allele B. It holds that $p_A + p_B = 1$ and $p_{ab} + p_{Ab} + p_{aB} + p_{AB} = 1$.

The measure of LD, D, in equation (2.3) can be calculated in every deme. Assuming there are multiple demes, which is the case in this Master's Thesis, and under the assumption that $s \ll r$, we have

$$D(z) = \frac{\sigma^2}{r} \left.\frac{dp_A}{dx}\right|_{x=z} \left.\frac{dp_B}{dx}\right|_{x=z}. \tag{2.4}$$

at deme z (Barton, 1986).

There are multiple ways of calculating $\frac{dp_A}{dx}$ (and $\frac{dp_B}{dx}$). One way is to calculate it directly from the average allele frequencies, as done by Durrett, Buttel and Harrison (Durrett et al., 2000). This method, however, is subject to large fluctuations. A more robust way to calculate the derivatives is to first fit a cline shape to the average allele frequencies, and then after that to take the derivative.

The maximum LD occurs when recombination between two loci is $r = 0$, since with no recombination there is maximal association between the two loci. Assume that these two loci are under selection with an environmental transition at the same place $Z_A = Z_B$. In the steady state at the transition, one will find that

$$p_A = p_B = 0.5 \tag{2.5}$$

and

$$p_{AB} = 0.5. \tag{2.6}$$

This steady state will always be reached since the model includes mutation. Mutation ensures that fixation is not a steady state. In the steady state that is reached, one can see from equation (2.3) that

$$D = p_{AB} - p_A p_B = 0.25. \tag{2.7}$$

2.2.4 Bimodality

As mentioned in the introduction, the bimodality in the mixture distribution of cline centers of two loci under selection is crucial since it could reveal that the corresponding environmental transitions occur at different places.

Bimodality of a distribution is easily estimated qualitatively by looking at a distribution; if it has two tops then it is bimodal. For the purpose of this Master's Thesis, a quantitative measure will be used instead. Using a quantitative measure, it is easier to estimate the level of bimodality.

The method used here is taken from Schilling *et al.* (Schilling, Watkins, & Watkins, 2002). It assumes that the possibly bimodal distribution consists of a mixture of two normal distributions and that one has information about the sample means and sample variances of these two distributions. Assume that μ_1 and μ_2 are the means and σ_1 and σ_2 are the corresponding standard deviations of the two normal distributions. Assume also that $\mu_1 < \mu_2$. Define the *separation factor*, $S(R)$, as

$$S(R) = \frac{\sqrt{-2 + 3R + 3R^2 - 2R^3 + 2(1 - R + R^2)^{\frac{3}{2}}}}{\sqrt{R}(1 + \sqrt{R})}, \tag{2.8}$$

where

$$R \equiv \frac{\sigma_1^2}{\sigma_2^2}. \tag{2.9}$$

Then, the mixed distribution is bimodal if and only if

$$\frac{\mu_2 - \mu_1}{\sigma_2 + \sigma_1} > S(R). \tag{2.10}$$

(Schilling et al., 2002).

Note that if $\sigma_1 = \sigma_2$ then $S(R) = 1$. In this case the means have to be separated by exactly two standard deviations for the distribution to be bimodal.

In the next chapter I present results obtained using the model and the methods explained in this chapter.

3

Results

In this chapter, the results from the simulations are presented. The results from the case with two loci under selection are presented here. For the results concerning only one locus, refer to Appendix A.2.

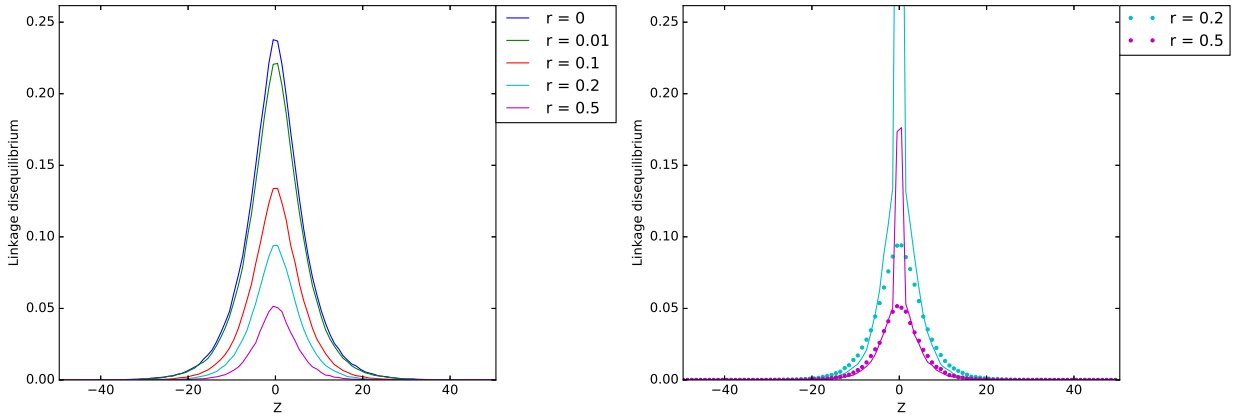
3.1 Linkage Disequilibrium

The Linkage Disequilibrium (LD) as a function of r can be seen in figure 3.1a. In figure 3.1b the LD has been compared to the theoretical value calculated according to Barton (1986), as stated in equation (2.4). In these plots, the environmental transitions take place at the same location $Z_A = Z_B$.

The LD reaches its maximum around the environmental change. When there is no recombination, the maximum LD is 0.25. The deviation from 0.25 in the plot is due to the following:

1. genetic drift,
2. the discreteness of the model; LD cannot be measured exactly on the environmental change which effectively takes place in between deme Z_A and deme $Z_A - 1$,
3. mutation occurs just before calculations of D, meaning that it sometimes will affect the demes closest to Z_A and thereby decrease the average LD.

Since the environmental change is between deme 74 and 75, we cannot get closer than 0.5 to where the maximum should actually be. The theoretical values seem to agree with simulations except near the environmental change, where the theory greatly overestimates the LD. The theory is expected to work when $s \ll r$ (Barton, 1986), which is not the case here since $r = 0.2$ and 0.5 and we have $s = 0.1$.



(a) **(b)**

Figure 3.1: Linkage disequilibrium over the habitat for different values of r obtained from the simulations. Panel (a) shows a comparison between results using different values of the recombination rate r. Panel (b) shows the simulated values as dots in comparison with equation (2.4). The same selection parameter $s_A = s_B = 0.1$ is used for both loci. The dispersal distance is $\sigma_D = 3$ and $Z_0 = 0$.

Instead of using equation (2.4) to calculate D directly on the allele frequencies, an attempt was made to fit a cline to the frequencies and use this fit to calculate $\frac{df(A)}{dx}$ and $\frac{df(B)}{dx}$. Some plots of this can be seen in figure 3.2. We see that using a fit cline gives a better agreement with simulations. Theory agrees better overall here since we use $s = 0.01$, fulfilling the requirement of $s \ll r$.

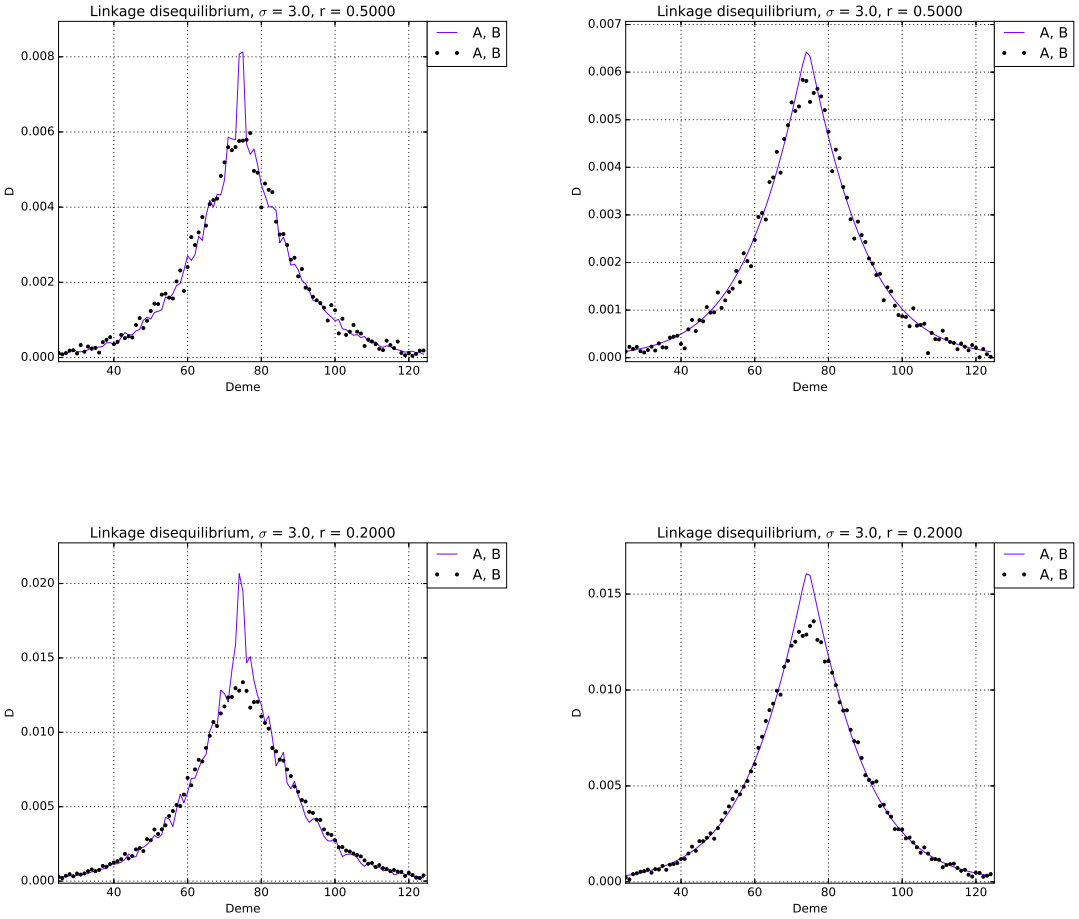


Figure 3.2: Linkage Disequilibium. The plots to the right have used a fit cline to calculate $\frac{df(A)}{dx}$ and $\frac{df(B)}{dx}$, while the plots on the left have used the mean allele frequencies directly. These simulations have been run for 50000 generations and $s_A = s_B = 0.01$ was the value of the selection parameter. $Z_0 = 0$. The dots are the average values of D calculated and the lines are the theoretical expectations. Remaining parameter values are listed on top of each panel.

Figure 3.3 shows an example of two fitted clines and their corresponding LD plot. The green line, which is the fit, agrees well with the average allele frequencies. Interestingly, unlike in cases where the two environmental transitions coincide, as in figures 3.1b and 3.2, in figure 3.3 we see that the LD exhibits a wide peak. This is an indication of the exsistance of multiple environmental transitions.

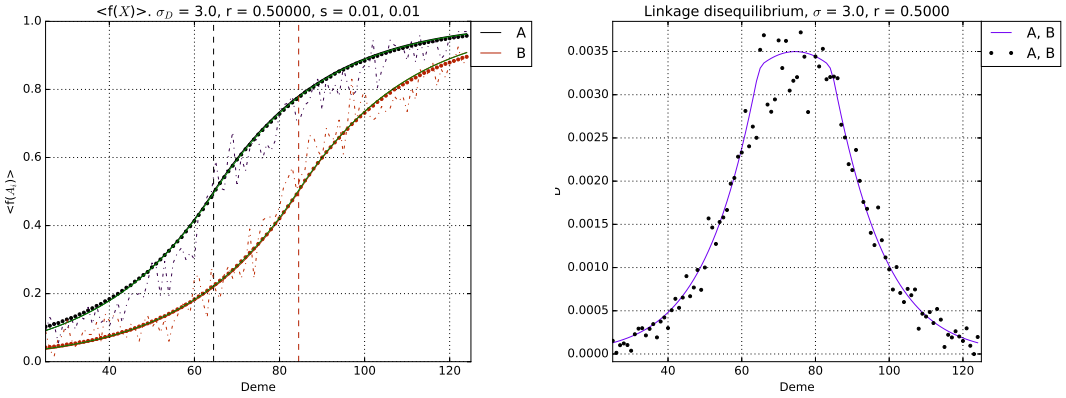


Figure 3.3: An example of two fitted clines (green lines) and the corresponding LD plot. $s_A = s_B = 0.01$ and $Z_0 = 10$.

In figure 3.4, the cline width w is plotted as a function of r. The environmental change occurs at the same location for both loci. B is more weakly selected for than A, and the cline of the B locus is drawn towards that of A as recombination decreases. This is due to the effect of indirect selection on the B locus as a results of the selection on the A locus. This was also observed by Slatkin (1975).

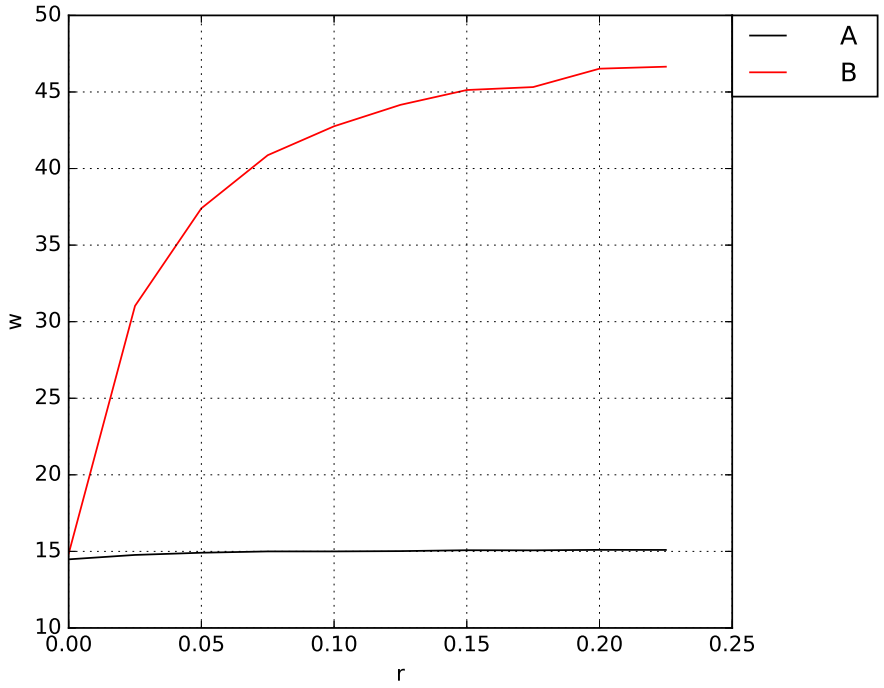


Figure 3.4: w as a function of r. For every value of r, the simulation ran for 50000 generations and every 10th generation has been used for averaging. The B trait is more weakly selected for, $s_B = 0.01$ compared to $s_A = 0.1$. $\sigma_D = 3$. The theoretical values of w, assuming only one locus is under selection, are 16.4 and 52.0 respectively.

In figure 3.5 we see that if the distance is too large between the environmental changes for the two traits, then linkage has no appearant effect on the locations of the clines even with $r = 0$. This was also a result of Slatkin, where the conclusion was that for $\sigma/\sqrt{s} < 2Z_0$, linkage will have a smaller effect and eventually no effect at all. By contrast, when Z_0 is sufficiently small the clines will be pulled together.

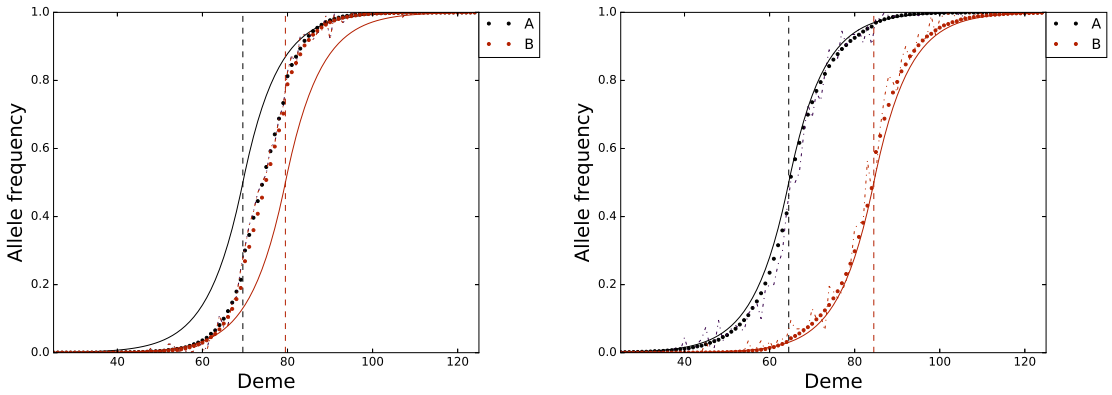


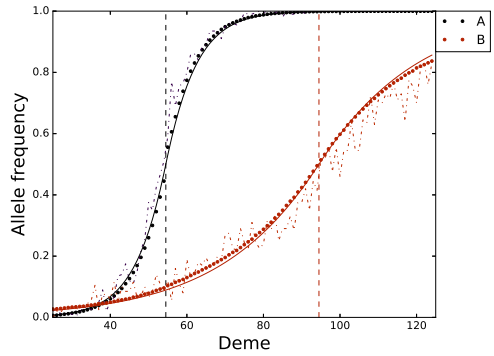
Figure 3.5: Clines resulting from using different values of the distance $2Z_0$ between the environmental changes. The values of the frequencies have been averaged over 10000 iterations. Thick dots represent the mean of the frequencies of one of the alleles in each locus. Solid lines represent the theoretical cline shape as in the one-locus model. Dashed lines represent one snapshot of the frequencies, taken at the last iteration of the realization. In the left panel, $Z_0 = 5$. In the right panel, $Z_0 = 10$. $\sigma/\sqrt{s} = 9.49$ in this case, so the behavior is as expected since this is less than $2Z_0$ in the left panel but larger by a big margin in the right panel.

In figure 3.6 are shown the resulting clines from simulation using different values of the selection parameter for each locus. We can see that for higher values of r, the shapes and locations of the clines are less correlated. Observe also that it is the cline of the trait with weaker selection that is drawn to the cline of stronger selection when recombination is low. The weaker cline gets a distorted shape, and a two-tailed sigmod gives the best fit to this curve.

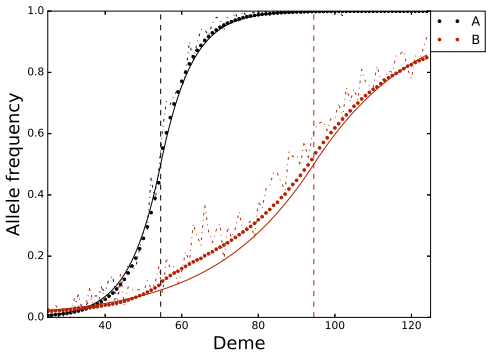
3.2 Midpoint of the clines

The midpoint of a cline for an allele B is defined (in this thesis) as the point on the x-axis for which the frequency of allele B, $f(B)$, is equal to 0.5. To estimate this, I fit cline shapes to the allele frequencies for several time-points and extract the midpoint data from these fits. This I found to be a more accurate way of finding the midpoint, rather than directly trying to estimate it based on the allele frequencies for every sample generation. An example fit can be seen in figure 2.2 for a cline with no tails and another example in figure 3.7 for a cline with tails at both ends.

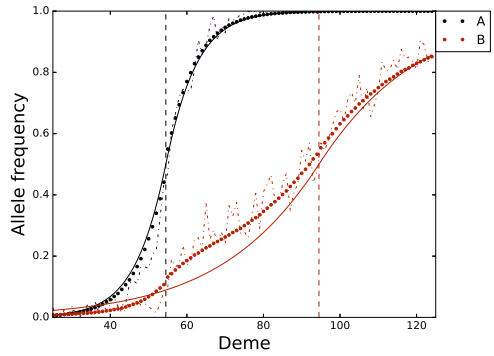
In figure 3.8 are shown the midpoints in the x-axis, denoted $Z_{0.5}$, of clines as a function of r for some different values of Z_0. Again, the effect of linkage is greatest for $\sigma/\sqrt{s} \approx 2Z_0$. This is in agreement with Slatkin (1975). However, the plot in subfigure 3.8b does not fully agree with the results of Slatkin in the limit of $r \to 0$. In the results of Slatkin, the clines were joined in the middle between the two environmental transitions. This does not agree either with his other results or my results; that the cline corresponding to the locus under stronger selection pressure will pull



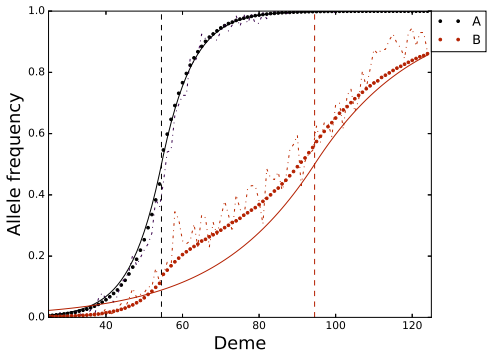
(a) $r = 0.50$.



(b) $r = 0.05$.



(c) $r = 0.01$.



(d) $r = 0$.

Figure 3.6: Thick dots represent the mean of the frequencies of one of the alleles in each locus. Solid lines represent the theoretical cline shape as in the one-locus model. Dashed lines represent one snapshot of the frequencies, taken at the last iteration of the realization. The cline corresponding to the trait more weakly selected for is pulled more than the other cline for decreasing r. Simulation generating these plots have run for 20000 generations, and every 10th generation have been used for averaging. $s_A = 0.1$ and $s_B = 0.01$. $Z_0 = 0$. The value of r is given below each panel.

more than the one corresponding to the locus under weaker selection pressure. This could be because Slatkin did not include drift in his model.

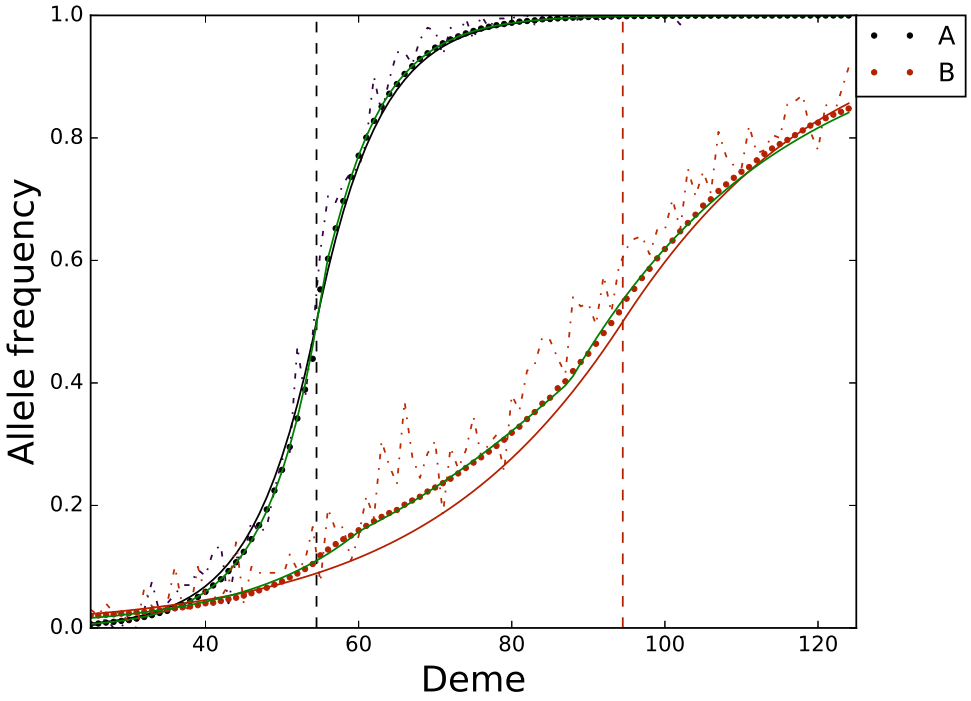


Figure 3.7: An example fit for a cline with tails at both ends.

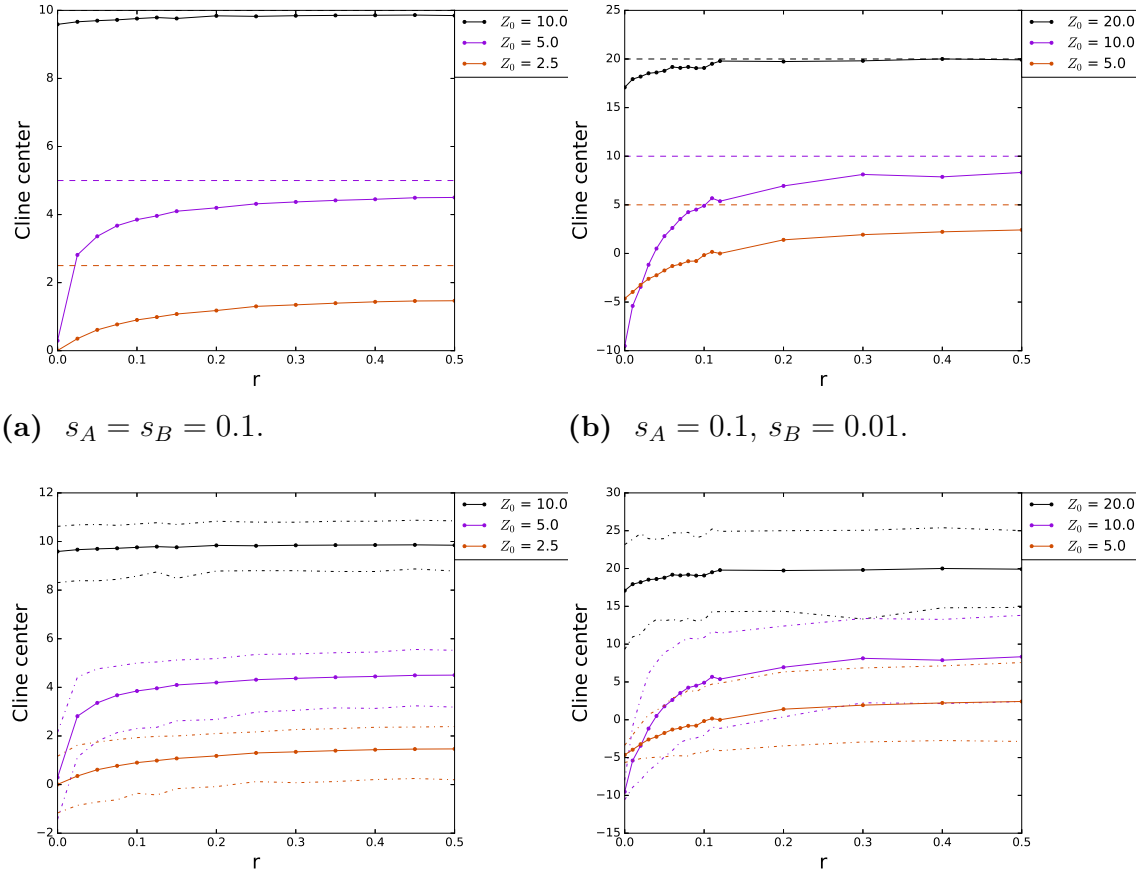
3.3 Critical value of r

The critical value of r, r_{crit}, is defined as the lowest value of r for which Z_0 is within the 95% CI of the estimated midpoint of the cline for allele B. In other words, for $r < r_{crit}$, the environmental transition for locus is outside of the 95% CI of the midpoint of the cline for allele B.

A heat map of r_{crit} can be seen in figure 3.9. One can see that the critical value of r increases for increasing s_B, and decreases for decreasing Z_0. This was expected; decreasing Z_0 means that the pulling together of the clines does not need to be as strong in order to join them together since they are close even without recombination. Also, increasing s_B means that the pulling together of the clines will be stronger since there is more overall selection pressure.

3.4 Bimodality in distribution of centers

When the two environmental changes occur close to each other, the cline centers are pulled together. This is supposed to happen for low values of r (Slatkin, 1975). I plotted the distributions of the estimations for the centers of the clines along with their mixture distribution. This can be seen in figures 3.10 and 3.11. Above certain critical values of Z_0, the resulting mixture distribution is bimodal. This critical value of Z_0, which I will call Z_{crit}, depends on s_A, s_B and r. Plots visualizing this



(a) $s_A = s_B = 0.1$.

(b) $s_A = 0.1$, $s_B = 0.01$.

Figure 3.8: Midpoints of the clines as a function of r for different values of Z_0. The dotted lines are the 95% confidence intervals of the midpoints. These data have been averaged over 5000 sample generations, with a space of 10 generations between each sample. In the plots to the left, $s_A = 0.1$ and $s_B = 0.01$. In the plots to the right, $s_A = s_B = 0.1$. The environmental changes occur $2Z_0$ from each other. The values of Z_0 can be found to the right of each panel.

can be found in figures 3.12, 3.13 and 3.14. Values of Z_{crit} obtained for various parameter combinations can be found in table 3.1. The general trend is clear; Z_{crit} increases for decreasing s, and it increases for decreasing r.

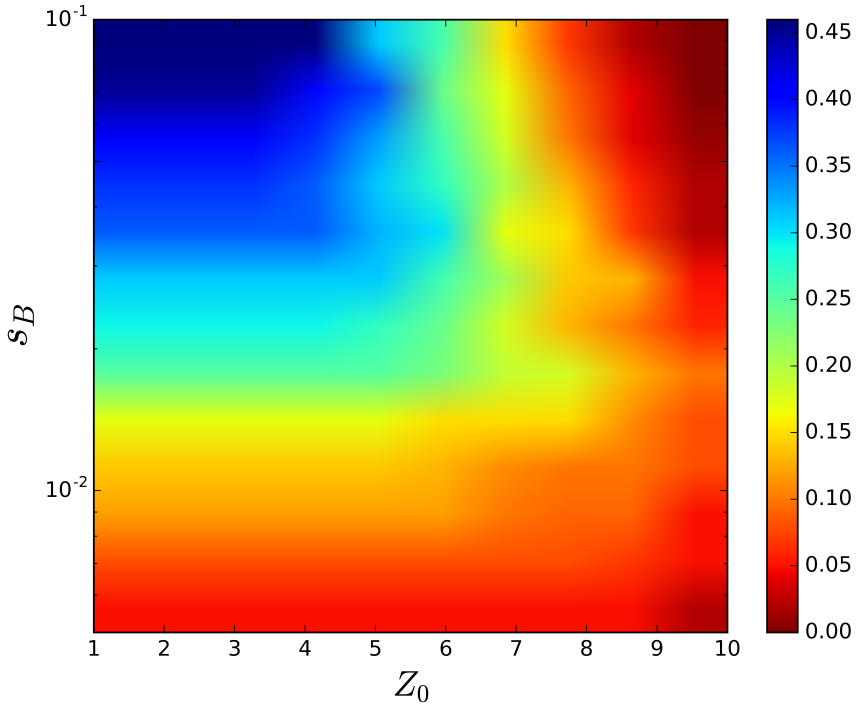


Figure 3.9: r_{crit} for different values of s_B and z_0. $s_A = 0.1$ in these simulations. The values of r_{crit} are color coded (see the color bar). Simulation time for every data point used to generate this figure was 50000 generations.

	$r = 0.5$	$r = 0.05$	$r = 0.005$
$s_A = s_B = 0.1$	1	2	4
$s_A = 0.1$, $s_B = 0.01$	1	2	7
$s_A = s_B = 0.01$	3	4	7
$s_A = s_B = 0.001$	14	16	17

Table 3.1: Z_{crit} for different value of s_A, s_B and r.

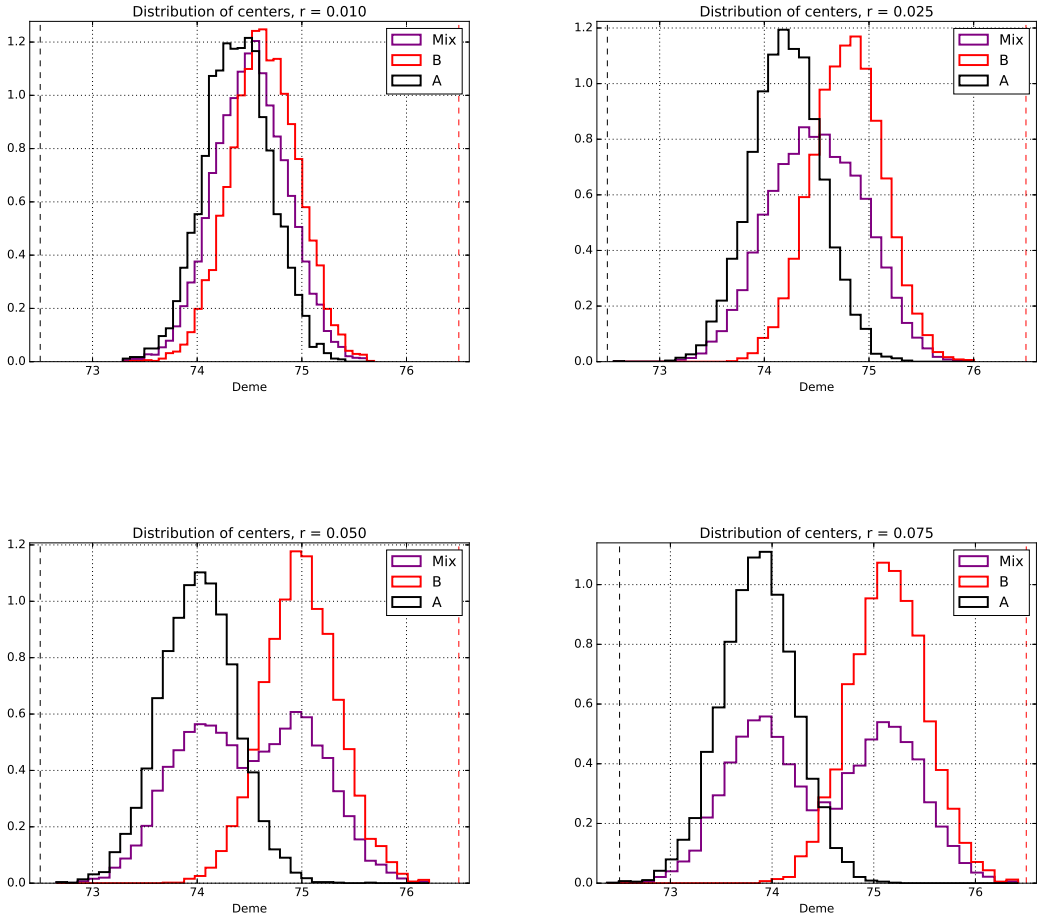


Figure 3.10: Distribution of cline centers for alleles A and B at the two loci under selection, along with their mixture distribution. Data have been gathered over 5000 sample generations. The dashed lines are the environmental changes for loci $\mathcal{A}$ (black) and $\mathcal{B}$ (red). Remaining parameter values: $z_0 = 2$, $s_A = s_B = 0.1$ and r is indicated above each panel.

The results presented in this chapter are further discussed and summarized in the following chapter.

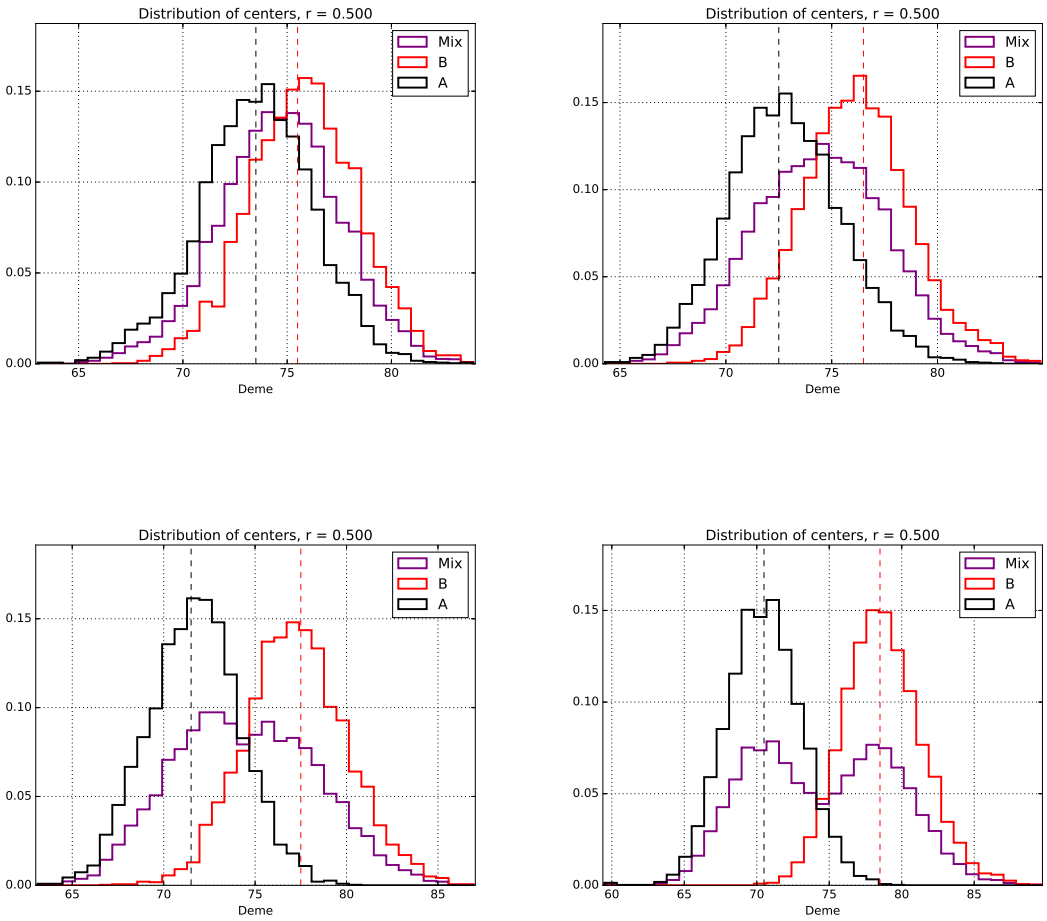


Figure 3.11: Same as in figure 3.10, but with $s_A = s_B = 0.01$.

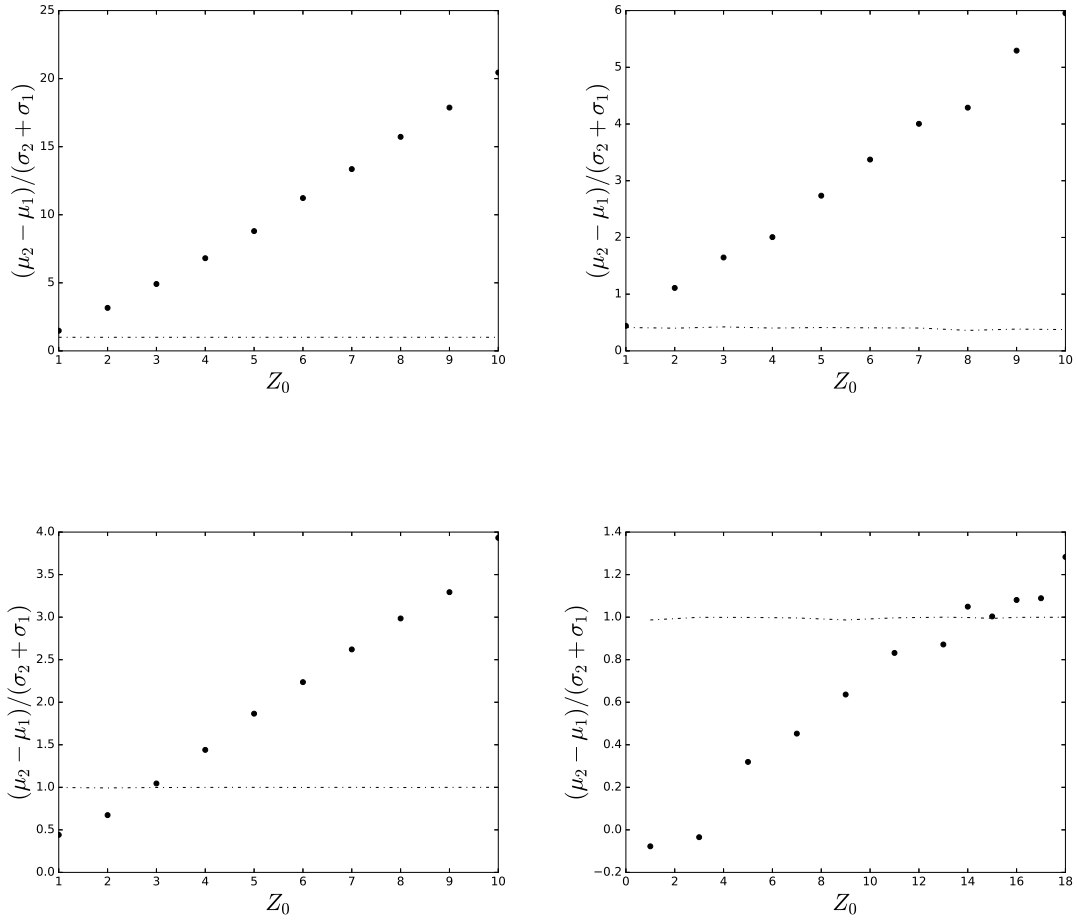


Figure 3.12: Bimodality compared to $S(R)$, as defined in equation (2.10), for different values of Z_0. The dots represent $\frac{\mu_2-\mu_1}{\sigma_2+\sigma_1}$ estimated from the distributions of the cline centers for alleles A and B. The dashed line is $S(R)$. When $\frac{\mu_2-\mu_1}{\sigma_2+\sigma_1} > S(R)$, the distribution is bimodal. The values of s_A and s_B vary as follows: $s_A = s_B = .1$ in the top left, $s_A = 0.1$ and $s_B = 0.01$ in the top right, $s_A = s_B = 0.01$ in the bottom left and $s_A = s_B = 0.001$ in the bottom right. $r = 0.5$ in all figures. Data were gathered every 10th iteration and the total simulation time was 50000 iterations.

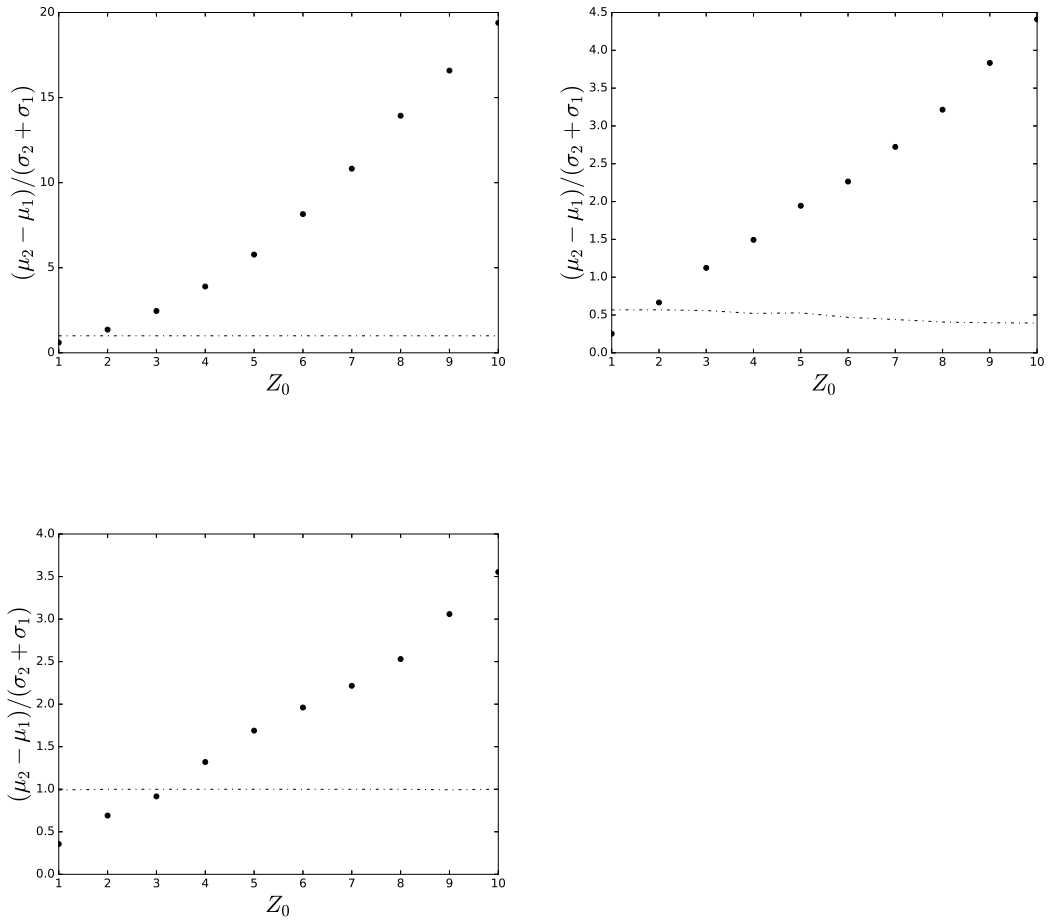


Figure 3.13: Same as in figure 3.12, but with $r = 0.05$.

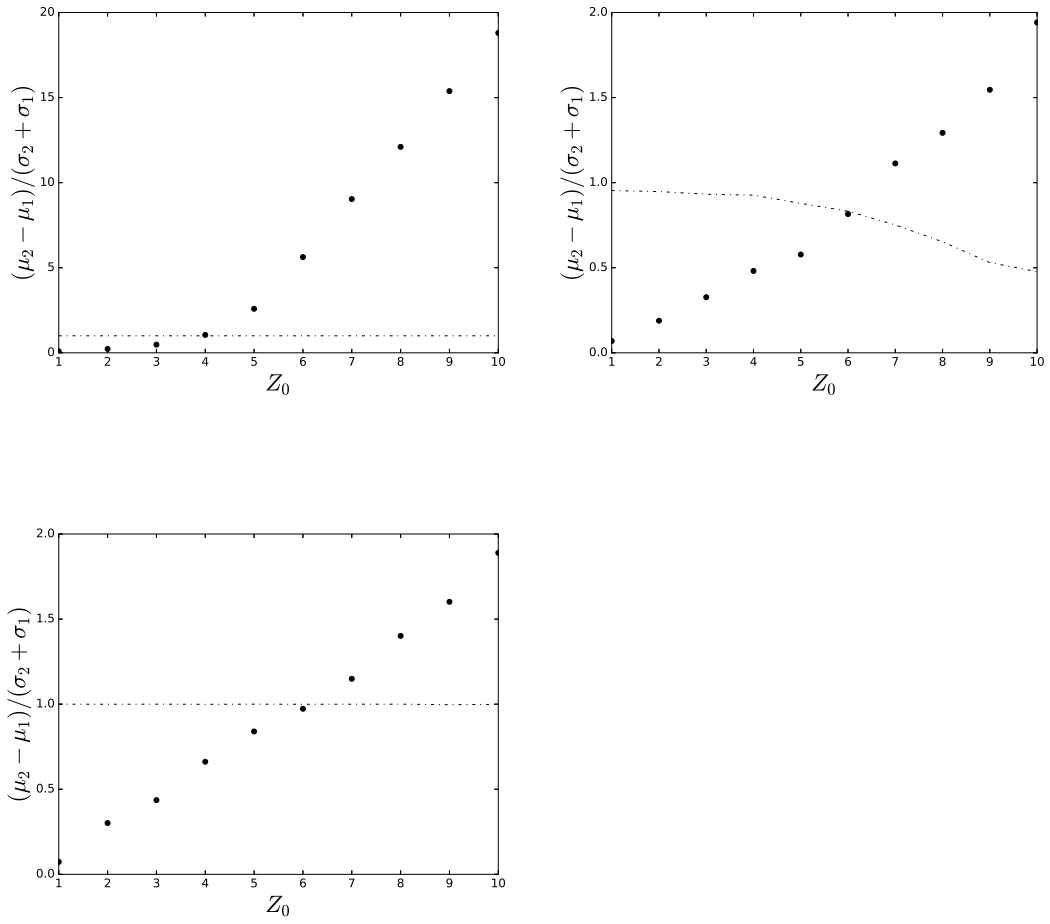
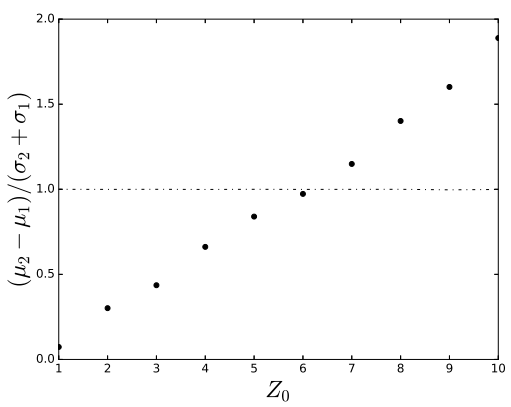
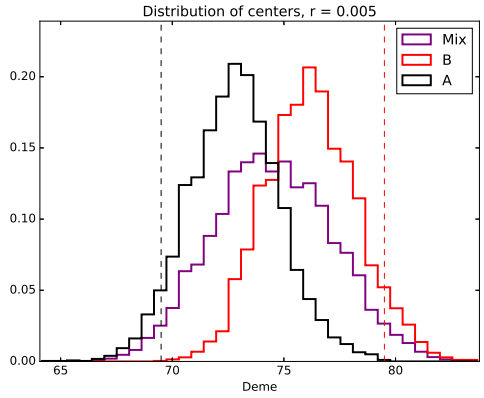


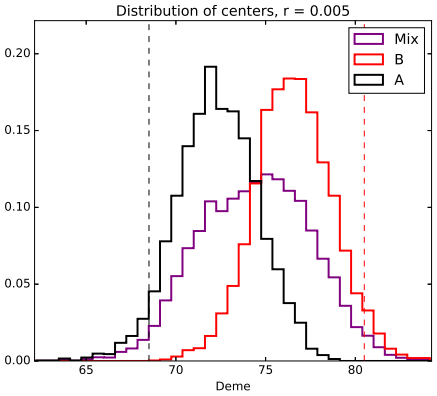
Figure 3.14: Same as in figure 3.12, but with $r = 0.005$.



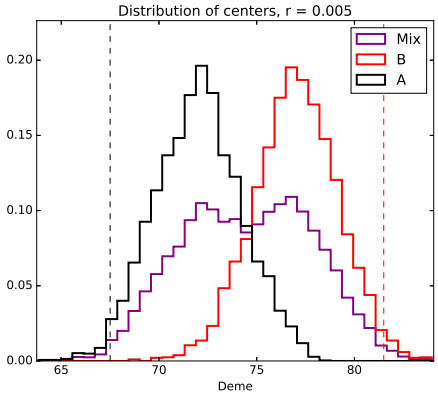
(a) Bimodality plot.



(b) $Z_0 = 5$

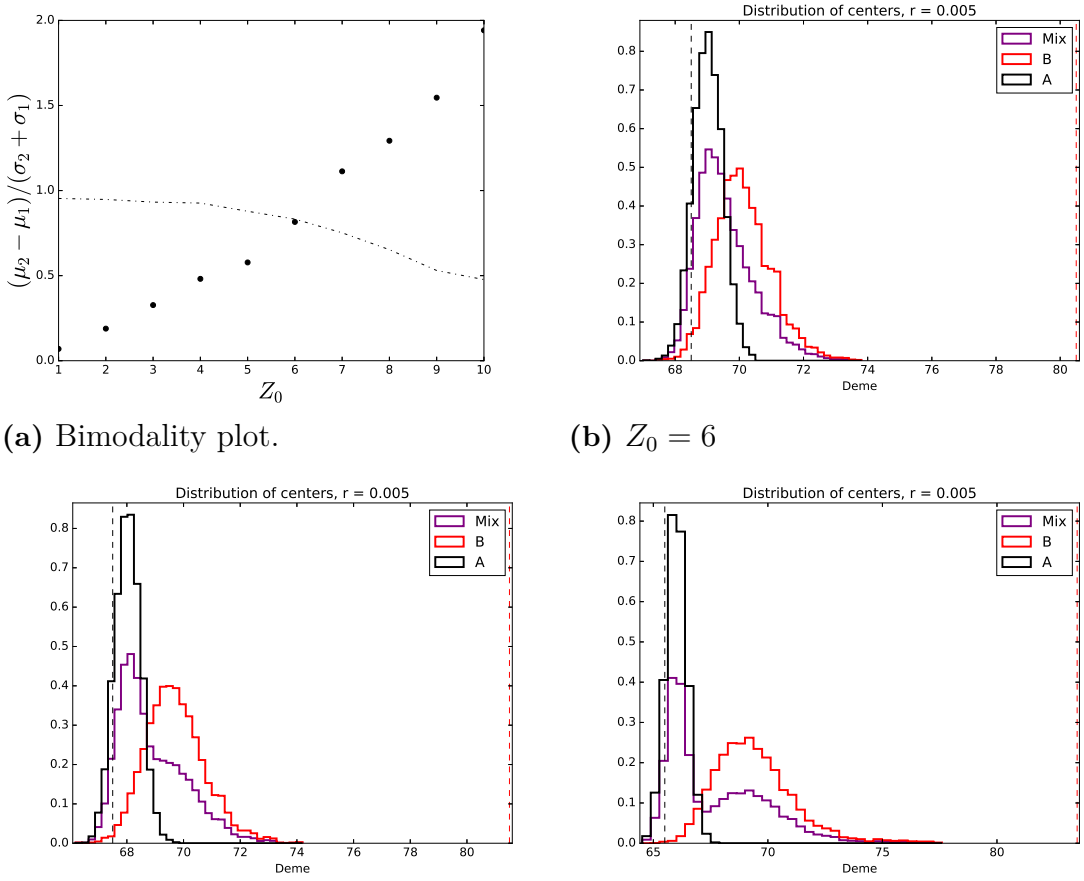


(c) $Z_0 = 6$



(d) $Z_0 = 7$

Figure 3.15: Top left: The same as in figure 3.14, bottom left. The other plots are the distributions of the centers and the resulting mixture distribution for $Z_0 = 5$, 6 and 7. These are the values closest to $S(R)$, as can be seen in panel (a). Data were gathered every 10th iteration and the total simulation time was 50000 iterations for every data point. Panel (d) shows a bimodal shape of the mixture distribution, which according to panel (a) is what we would expect since for $Z_0 = 7$ we have $(\mu_2 - \mu_1)/(\sigma_2 + \sigma_1) > S(R)$.



(a) Bimodality plot.

(b) $Z_0 = 6$

(c) $Z_0 = 7$

(d) $Z_0 = 9$

Figure 3.16: Top left: The same as in figure 3.14, top right. The other plots are the distributions of the centers and the resulting mixture distribution for $Z_0 = 6$, 7 and 9. Data were gathered every 10th iteration and the total simulation time was 50000 iterations for every data point. Panel (c) shows a mixture distribution that is bimodal according to the test, but the bimodality is not appearant to the eye. In panel (d), however, the bimodality of the mixture distribution is unquestionable.

4

Discussion and Conclusion

For the parts of my simulations which were run with the same parameters as those of Slatkin (Slatkin, 1975), my results mostly agreed completely with his. A difference was found in the plot of the midpoints of the clines as a function of r in figure 3.8. In Slatkins figure (figure 9 in Slatkin (1975)), the midpoint of the cline for allele B went to 0 for $r = 0$ when $s_A = 0.1$ and $s_B = 0.01$. The midpoint of the B-cline in my case went instead to the midpoint of the A-cline. This controversy is probably due to the fact that Slatkin did not include genetic drift in his simulations. I would say that my result is more intuitive; the clines are completely joined together for $r = 0$ when Z_0 is low enough to allow it. The point on the x-axis where they are joined are in my case much closer to the environmental change corresponding to the trait under stronger selection pressure than the one with weaker selection pressure. This makes sense considering the results in figure 3.6, where we have seen that it is almost solely the weaker cline that is being pulled towards the stronger cline. In Slatkin (1975), for $r = 0$ the clines are joined at $x = 0$. This means they are joined in the middle between the environmental changes with no regards to which trait is under stronger selection pressure. In my view, this does not make sense.

The mixture distribution of the centers of the clines need higher values of Z_0 to be bimodal when selection is weaker. Some concrete values have been presented in the results (figures 3.12, 3.13 and 3.14). A bimodal pattern directly points towards the existence of multiple environmental transitions in the habitat in question. However, for $Z_0 < Z_{crit}$ such an existence cannot be detected based on the mixture distribution of cline centers. An interesting plot to make in the future would be to map out the critical values of Z_0 for a grid of values of s_A and s_B. This grid would tell us which distances between the environmental selection we would need in order to be able to detect that there are actually two environmental transitions and not just one, for a large span of parameters.

One thing to point out is that when inferring about the bimodality of the mixture distribution of cline centers in my simulations, I already have a clustering of the data used to make up the mixture distribution. That is, I can estimate the values of μ_1, μ_2, σ_1 and σ_2 since I know which data of the mixture distribution comes from which original normal distribution. When using empirical data, no such clustering of cline center data is given. Clustering would first have to be made of the data, separating it into two, or more, groups for which we want to assess whether or not the clines are the result of two environmental transitions. This could be done in a trial-and-error fashion; if we assume that bimodality directly implies multiple

environmental transitions then we cloud try all sorts of clusterings to see if we even once get bimodality in the resulting mixture distribution. This can also be seen directly by inspection, if the total distribution does not look bimodal then there will be no clustering of the data which would make it bimodal according to the formula in equation 2.10.

An extension of my model would be to include neutral loci. These could tell us about whether some observations really are a consequence of the loci being under selection, or if some observations simply are due to linkage between the loci. Neutral loci would allow me to make this distinction. Another extension would be to handle fragmentation of the population in my model, where the number of individuals in some demes is lowered or equal to zero. Such an extension would be particularly relevant for the analysis of genetic data from *L. saxatilis* sampled from the hybrid zones in Sweden (K. Johannesson, R. K. Butlin, personal communication, 2017).

I will now present a reasonable approach to successfully inferring whether or not there are two environmental changes in a habitat.

Empirical data consists of the allele frequencies of many loci of *L. saxatilis* as a function of distance. Let us say it also contains information about local population sizes, which is a fair assumption. From this, one can estimate selection strength for each cline shape. Let us also say we have a genetic map of the genome, so that we can estimate r between all the loci. The problem now is to determine whether these allele frequencies are the result of one or two environmental transitions. The null hypothesis is that there is only one environmental transition.

1. First, one would look at the shape of the clines. From my results, if the distance between the environmental transitions is large enough, but not too large, then the cline(s) corresponding to the trait more weakly selected for could exhibit a two-tailed shape. This could be quantitatively tested by fitting both a no-tail cline and a two-tailed cline and compare some test of goodness of fit, for example using the Akaike information. If the fit of a two-tailed shape is significantly better then we could reject the null hypothesis that there is only one environmental transition.
2. If the fit of a two-tailed shape is not significantly better, one would have to look at the distribution of centers as discussed above. If it yields a significantly bimodal distribution then we could reject the null hypothesis.
3. Finally, one could look at the shape of the LD plots between the two groups of loci derived in 2). A wider peak in the LD plot could point to there being more than one environmental transition.

If none of there three tests point towards there being more than one environmental transition, the null hypothesis that there is only one transition would be accepted.

References

Barton, N. H. (1986, Dec). The effects of linkage and density-dependent regulation on gene flow. *Heredity (Edinb)*, *57 (Pt 3)*, 415–426.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, *16*, 113-148. Retrieved from http://www.jstor.org/stable/2097045

Barton, N. H., & Hewitt, G. M. (1989, Oct 12). Adaptation, speciation and hybrid zones. *Nature*, *341*(6242), 497-503. Retrieved from http://dx.doi.org/10.1038/341497a0 doi: 10.1038/341497a0

Bazykin, A. (1969). Hypothetical mechanism of speciaton. *Evolution*, *23*(4), 685–687.

Durrett, R., Buttel, L., & Harrison, R. (2000, 01). Spatial models for hybrid zones. *Heredity*, *84*, 9-19. Retrieved from http://proxy.lib.chalmers.se/login?url=http://search.proquest.com/docview/229999192?accountid=10041 (Copyright - Copyright Blackwell Science Ltd. Jan 2000; Last updated - 2013-01-25; CODEN - HDTYAT)

Gay, L., Crochet, P.-A., Bell, D. A., & Lenormand, T. (2008). Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, *62*(11), 2789–2806.

Guedj, B., & Guillot, G. (2011). Estimating the location and shape of hybrid zones. *Molecular Ecology Resources*, *11*(6), 1119–1123.

Howard, D., & Berlocher, S. (1998). *Endless forms: Species and speciation*. Oxford University Press. Retrieved from https://books.google.se/books?id=pLzY1-wyOKwC

Janson, K. (1982). Genetic and environmental effects on the growth rate of *Littorina saxatilis*. *Marine Biology*, *69*(1), 73–78.

Janson, K. (1983, Aug). Selection and migration in two distinct phenotypes of *Littorina saxatilis* in Sweden. *Oecologia*, *59*(1), 58–61.

Johannesson, K., Panova, M., Kemppainen, P., André, C., Rolán-Alvarez, E., & Butlin, R. K. (2010). Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1547), 1735–1747. Retrieved from http://rstb.royalsocietypublishing.org/content/365/1547/1735 doi: 10.1098/rstb.2009.0256

Johannesson, K., Rolan-Alvarez, E., & Ekendahl, A. (1995). Incipient reproductive isolation between two sympatric morphs of the intertidal snail *Littorina saxatilis*. *Evolution*, *49*(6), 1180-1190. Retrieved from http://www.jstor.org/

stable/2410443

Key, K. H. L. (1981). Species, parapatry, and the morabine grasshoppers. *Systematic Biology*, *30*(4), 425. Retrieved from +http://dx.doi.org/10.1093/sysbio/30.4.425 doi: 10.1093/sysbio/30.4.425

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press. Retrieved from https://books.google.se/books?id=mAIjnLp6r_MC

Mayr, E. (1982). *The growth of biological thought: Diversity, evolution, and inheritance*. Belknap Press. Retrieved from https://books.google.se/books?id=pHThtE2R0UQC

Merriam Webster Online. (2017). *Definition of cline*. Retrieved 2017-04-27, from https://www.merriam-webster.com/dictionary/cline

Panova, M., Hollander, J., & Johannesson, K. (2006, Nov). Site-specific genetic divergence in parallel hybrid zones suggests nonallopatric evolution of reproductive barriers. *Mol. Ecol.*, *15*(13), 4021–4031.

Polechova, J., & Barton, N. (2011, Sep). Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, *189*(1), 227–235.

Schilling, M. F., Watkins, A. E., & Watkins, W. (2002). Is human height bimodal? *The American Statistician*, *56*(3), 223-229. Retrieved from http://dx.doi.org/10.1198/00031300265 doi: 10.1198/00031300265

Selander, R. (1964). Speciation in wrens of the genus canlpylorhynchus. *Univ. Cahf Pubis Zool*, *74*, l–224.

Slatkin, M. (1973, Dec). Gene flow and selection in a cline. *Genetics*, *75*(4), 733–756.

Slatkin, M. (1975, Dec). Gene flow and selection in a two-locus system. *Genetics*, *81*(4), 787–802.

Slatkin, M. (2008, Jun). Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, *9*(6), 477-485. Retrieved from http://dx.doi.org/10.1038/nrg2361 doi: 10.1038/nrg2361

Ståhl, N. (2016). *The Effect of Sexual Selection on Cline Patterns in Biological Traits* (Tech. Rep.). Chalmers University of Technology.

Szymura, J. M., & Barton, N. H. (1986). Genetic analysis of a hybrid zone between the fire-bellied toads, bombina bombina and b. variegata, near cracow in southern poland. *Evolution*, *40*(6), 1141–1159.

Szymura, J. M., & Farana, I. (1978). Inheritance and linkage analysis of five enzyme loci in interspecific hybrids of toadlets, genus bombina. *Biochemical genetics*, *16*(3), 307–319.

Wilding, C. S., Butlin, R. K., & Grahame, J. (2001). Differential gene exchange between parapatric morphs of littorina saxatilis detected using aflp markers. *Journal of Evolutionary Biology*, *14*(4), 611–619. Retrieved from http://dx.doi.org/10.1046/j.1420-9101.2001.00304.x doi: 10.1046/j.1420-9101.2001.00304.x

World Register of Marine Species. (2017). *Littorina saxatilis*. Retrieved 2017-08-23, from http://marinespecies.org/aphia.php?p=taxdetails&id=140264

A

Appendix

A.1 Theoretical shape of a cline

In this section, the formula for the deterministic shape of a cline is derived. This is done in more or less the same way as in Ståhl (Ståhl, 2016), but with some steps being described in more detail.

This formula is valid in the following model:

- One locus is under selection.
- The environment is a step change around $x = 0$ with selection parameters according to Table A.1.

The following limits are considered:

- local population size is infinitely large, i.e. $N \to \infty$,
- the distance between demes $\to 0$,
- the number of patches is infinitely large, i.e. $K \to \infty$.

This means we are considering the entire x-axis as our habitat with infinitely many individuals on each point. This approximation will work better for larger K and smaller s in the original discrete model. Larger σ_D will also make the approximation work better as long as K is large enough to support the entire cline.

Assume $p(x,t)$ is the frequency of allele A after migration, and $q(x,t) \equiv 1-p(x,t)$ is the frequency of allele a. Let $p'(x,t)$ and $q'(x,t)$ denote the frequencies after selection and reproduction, but before migration.

This derivation of the cline shape will consist of first writing p in terms of p', and then p' in terms of p. Insertion of one of these formulas into the other will reduce the model to a differential equation, which can be solved.

Let $M(x,x')$ be the probability that an individual migrates from x' to x in one time

	aa	aA	AA
$x < 0$	$1-s$	1	$1+s$
$x \geq 0$	$1+s$	1	$1-s$

Table A.1: The selection parameters as used in the derivation of the theoretical cline.

step. One time step will be denoted as δt. Since M is a probability distribution,

$$\int_{-\infty}^{\infty} M(x,x')\,dx = \int_{-\infty}^{\infty} M(x,x')\,dx' = 1.$$

We will assume that M is independent of direction, i.e. an even function in both x and x' centered around $x = x'$.

Now, we can write p in terms of p' as

$$p(x,t) = \int_{-\infty}^{\infty} M(x,x')p'(x',t-\delta t)\,d(x'-x) \tag{A.1}$$

By Taylor expanding p' around (x,t) we find

$$p'(x',t-\delta t) = p'(x,t) + (x'-x)\frac{\partial p'(x,t)}{\partial x} - \delta t\frac{\partial p'(x,t)}{\partial t} + \frac{(x'-x)^2}{2}\frac{\partial^2 p'(x,t)}{\partial x^2} - \tag{A.2}$$

$$- (x'-x)\delta t\frac{\partial^2 p'(x,t)}{\partial x \partial t} + \frac{\delta t^2}{2}\frac{\partial^2 p'(x,t)}{\partial t^2} + \dots \tag{A.3}$$

By inserting equation (A.2) into equation (A.1) we find

$$p = p' - \delta t\frac{\partial p'}{\partial t} + \frac{\delta t^2}{2}\frac{\partial^2 p'}{\partial t^2} + \int_{-\infty}^{\infty} M(x,x')\frac{(x'-x)^2}{2}\frac{\partial^2 p'}{\partial x^2}\,d(x'-x), \tag{A.4}$$

where $p = p(x,t)$ and $p' = p'(x,t)$. The linear part of the Taylor expansion vanishes since M is even.

If $M(x,x') \sim Normal(x'-x, \sigma_D)$ we have

$$p = p' - \delta t\frac{\partial p'}{\partial t} + \frac{\delta t^2}{2}\frac{\partial^2 p'}{\partial t^2} + \frac{\partial^2 p'}{\partial x^2}\frac{\sigma_D^2}{2}. \tag{A.5}$$

To write p' in terms of p we will look more in detail on the frequencies of each allele.

Since the frequency of allele A is p, the frequencies of each possible genotype are according to the following table:

aa	aA	AA
$(1-p)^2$	$2p(1-p)$	p^2

These frequencies sum to 1.

The contribution to the total frequency of allele A for each genotype is therefore

aa	aA	AA
0	$p(1-p)$	p^2

.

These frequencies sum to p.

Multiplying these frequencies with the relative fitness values for these genotypes, according to Table A.1, will be the exact effect of selection when $x > 0$. The resulting formula is

$$p' = \frac{p(1-p)\cdot 1 + p^2(1+s)}{(1-p)^2(1-s) + 2p(1-p)\cdot 1 + p^2(1+s)}. \tag{A.6}$$

Here, we have divided by the total fitness to get the relative fitness.

This equation can be simplified to

$$p' = \frac{p + sp^2}{1 + s(2p-1)}. \tag{A.7}$$

By Taylor expanding around $s = 0$ and neglecting terms of order s^2 and higher we find

$$p' = p + s(p - p^2) + \dots \tag{A.8}$$

We now assume that we are in the steady state, meaning $p(x,t) = p(x,t')\forall t, t'$. Inserting equation (A.7) into equation (A.5) we get

$$p = p + s(p-p^2) + \frac{\partial^2(p + s(p-p^2))}{\partial x^2}\frac{\sigma_D^2}{2}. \tag{A.9}$$

Neglecting terms of order $s\sigma_D^2$ we get

$$p = p + s(p-p^2) + \frac{d^2p}{dx^2}\frac{\sigma_D^2}{2}. \tag{A.10}$$

Now it remains to find the solution to the differential equation

$$\frac{d^2p}{dx^2} = -\frac{2s}{\sigma_D^2}(p-p^2). \tag{A.11}$$

Using $g = \frac{dp}{dx}$ yields

$$\begin{aligned} g\frac{dg}{dp} &= -\frac{2s}{\sigma_D^2}(p-p^2) \Rightarrow \\ \int g\,dg &= -\frac{2s}{\sigma_D^2}\int p - p^2\,dp \Rightarrow \\ \frac{1}{2}g^2 &= -\frac{2s}{\sigma_D^2}\left(\frac{p^2}{2} - \frac{p^3}{3}\right) + C. \end{aligned} \tag{A.12}$$

In the limit of $x \to \infty$, we must have $g \to 0$ and $p \to 1$. This yields

$$C = \frac{s}{3\sigma_D^2}. \tag{A.13}$$

So now we have

$$\frac{dp}{dx} = \sqrt{-\frac{4}{\sigma_D^2} s \left(\frac{p^2}{2} - \frac{p^3}{3} - \frac{1}{6} \right)}. \tag{A.14}$$

This differential equation was solved by using Mathematica. The final equation for the slope of the cline is

$$p(x) = -\frac{1}{2} + \frac{3}{2} \tanh \left(\frac{\sqrt{s/2}}{\sigma_D} x + \operatorname{atanh} \left(\sqrt{2/3} \right)^2 \right). \tag{A.15}$$

Differentiating this and evaluating at $x = 0$ gives

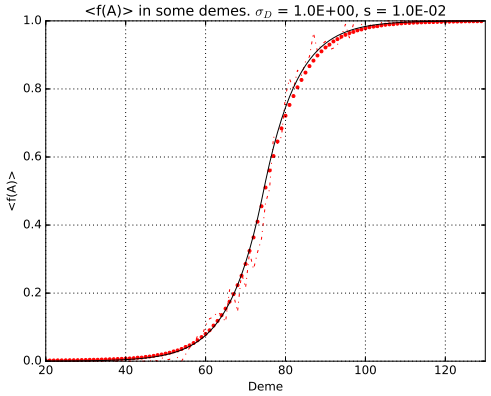
$$\frac{dp}{dx}(0) = \frac{1}{\sigma} \sqrt{\frac{s}{3}}, \tag{A.16}$$

which is the maximum slope of the cline. The width of a cline, w, is defined as $w = \frac{1}{\text{maximum slope}}$ (Bazykin, 1969).

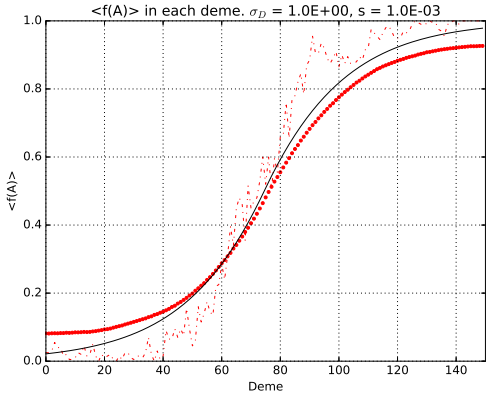
A.2 Results using one trait and one locus

In this section, results from running the reduced model with one trait and one locus are presented.

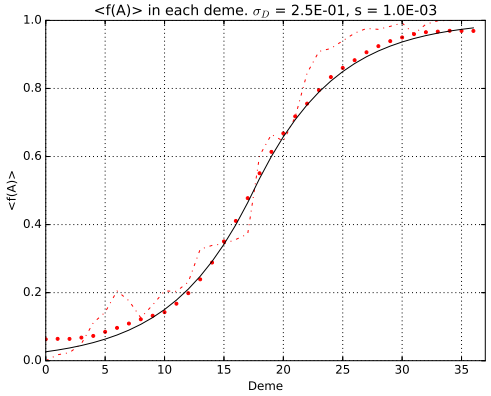
First is the exact shape of the cline resulting from averaging the allele frequencies of allele A. The shape has been derived in the theoretical limit in Appendix A.1. Some plots can be seen in figure A.1. The frequencies were initialized according to the deterministic limit in equation (A.15). I let the simulations run for 500 generations before I begun averaging. This was to ensure that the model had reached its equilibrium state in terms of fluctuations. $f(A)$ are the frequencies of allele A across the demes, also denoted as p in the main text of this Master's Thesis.



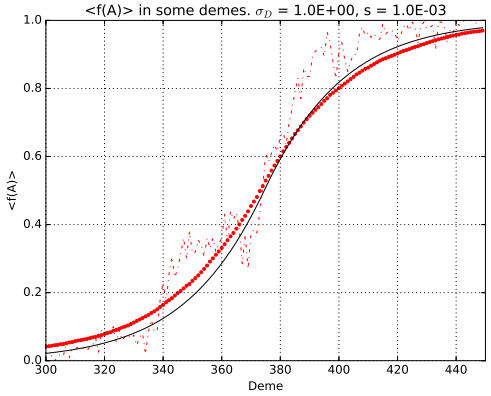
(a) K = 150, N = 100, $\sigma_D = 1$, $s = 10^{-2}$.



(b) K = 150, N = 100, $\sigma_D = 1$, $s = 10^{-3}$.



(c) K = 37, N = 400, $\sigma_D = 0.25$, $s = 10^{-3}$.



(d) K = 750, N = 100, $\sigma_D = 1$, $s = 10^{-3}$.

Figure A.1: Average frequencies of allele A. Black: The theoretical cline shape according to equation (A.15). Red dots: Average frequency of allele A. Dashed red line: One snapshot of $f(A)$. In total the simulations were run for 10000 generations, and every 10th of these generations were used for averaging.

The simulated values agree with the theory for certain values of s and σ_D. For some values, the theory fails. More specifically, the theory fails when s is too low or σ_D is too high. This is because the width scales as $\frac{1}{\sigma}\sqrt{\frac{s}{3}}$ (see Appendix A.1). When the

width is too high the boundaries start having an effect on the cline shape, which is undesirable.

Indeed, as can be seen in figure A.1b the width is high in relation to the amount of demes; cline width is more than half of the total habitat length. The boundaries then have a high impact on the shape of the cline. Two workarounds can be considered to deal with this, if these values of the width are to be studied. One way is to rescale the parameters according to

$$\begin{aligned} K &\to \frac{K}{R} \\ N &\to N \cdot R \\ \sigma_D &\to \frac{\sigma_D}{R}, \end{aligned}$$

where R is a rescale factor. One result from using the rescaling can be seen in Figure A.1c. The fit to the cline is better here than in A.1b, where the unscaled parameters were used.

Another way to avoid the effect of the boundaries is simply to use a higher K. The tradeoff is that simulation time grows linearly with K. One result from running the simulation with $K = 750$ can be seen in figure A.1d. Compare to figure A.1b where the same σ_D and s have been used but with a lower K.

Another reason why the fit of the theoretical slope is not perfect is due to the discreteness of the model. The theoretical cline shape has been derived in the limit of continuous space. Using high values of both K and σ_D, the effect of discreteness can be lowered.

And lastly, one more reason why the fit is not perfect is due to the neglection of terms of order $s \cdot \sigma_D$, s^2, σ_D^4 and higher in the derivation of the cline shape (see Appendix A.1).

The path to convergence of the width of the cline has been plotted in figure A.2 for some values of s and σ_D. For lower values of s, and higher values of σ_D, it takes longer to reach equilibrium. F is defined as

$$F(x,t) \equiv \mathrm{var}(p(x,t))/(p(\bar{x},t)q(\bar{x},t)), \tag{A.17}$$

where $q = 1-p$. Averages are taken over time. $(1-\overline{F})w(\overline{p})$ is equal to $\overline{w(p)}$ according to theory (Polechova & Barton, 2011), and this agrees with my simulations presented in this figure. $w(\overline{p})$ is the width of the average cline and $\overline{w(p)}$ is the average width of a cline. My results regarding the average width of a cline and the width of the average cline also agree with Polechova, J. and Barton, N. (2011).

In Figure A.3 is presented the width as a function of s compared to the theoretical width for one value of σ_D. The corresponding plots for other values of σ_D can be seen in Appendix A.3. We are supposed to see $\overline{w(p)}$ equal to the theoretical values of w. According to Polechova, J. and Barton, N. (2011), we are also supposed to see $(1 - \overline{F})w(\overline{p}) = \overline{w(p)}$. In figure A.4 is the relative difference between the theoretical width and the width from the simulations. For low values of $1/\sqrt{s}$, theory deviates from simulation. This is most easily seen in figure A.4. The deviation is due the

discreteness of the model; higher values of s mean that the cline will be steeper and hence discreteness has a larger effect. For high values of $1/\sqrt{s}$, theory fails since the width becomes too large in relation to the habitat size. When the cline is close to being as wide as the habitat, the boundaries can have a larger effect on the outcome of the simulation.

A.3 Additional plots

In this section additional interesting plots are presented.

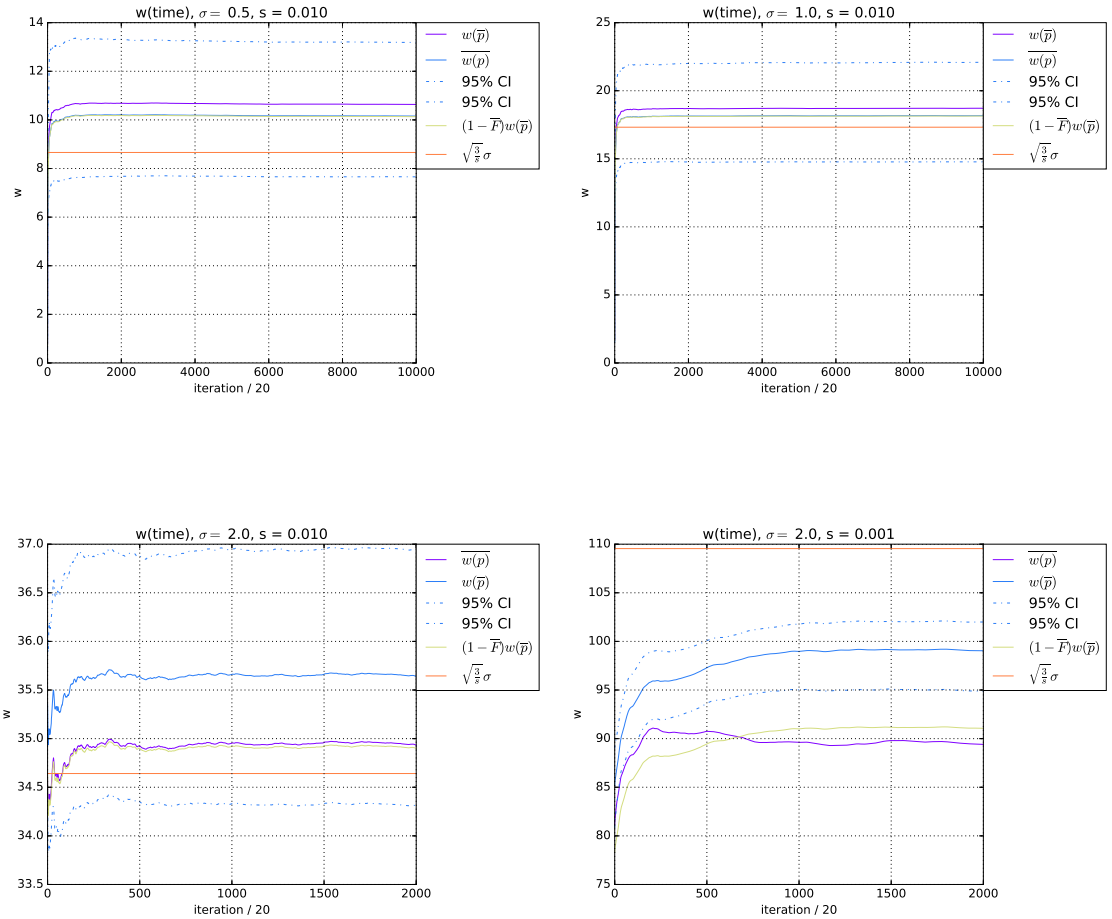


Figure A.2: Travelling mean of w as a function of time for different values of s and σ_D. These values have been averaged over 20 realizations of the model. Note the time required to reach equilibrium and the amplitude of the fluctuations in equilibrium. $\overline{w(p)}$ is the average of the widths measured each realization. $w(\overline{p})$ is the width measured from the average gene frequencies across the realizations. The dashed lines are the 95% CI for $w(\overline{p})$. F is the standardized local variance in allele frequency defined therein. $\sqrt{\frac{3}{s}}\sigma$ is the theoretical cline width derived in Appendix A.1. $K = 150$.

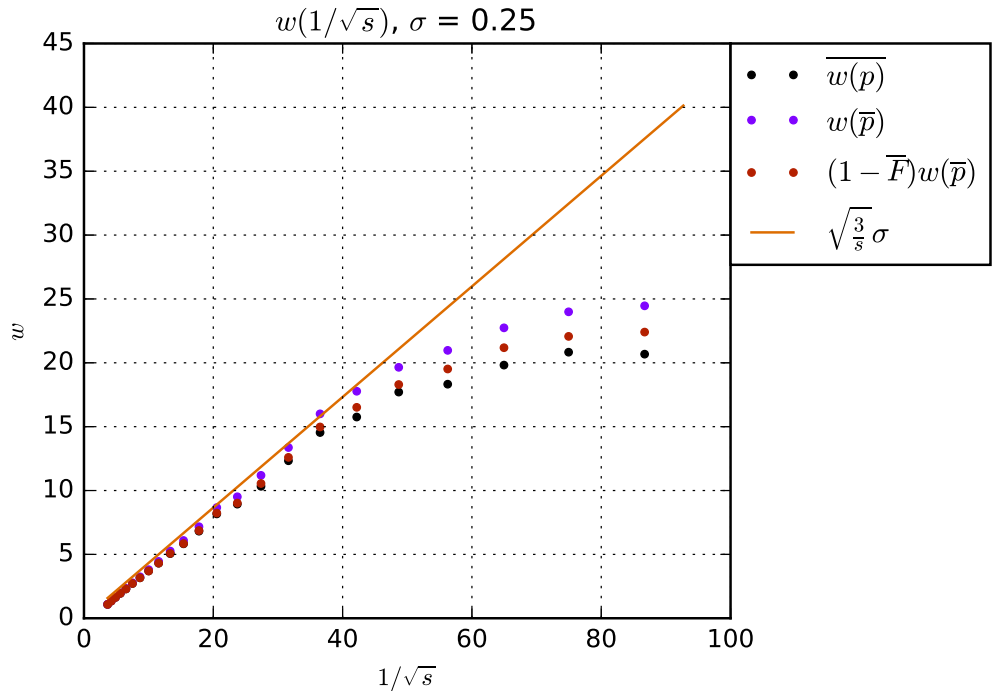


Figure A.3: w as a function of $\frac{1}{\sqrt{s}}$. The theoretical width is defined as $w = \sqrt{\frac{3}{s}}\sigma_D$.

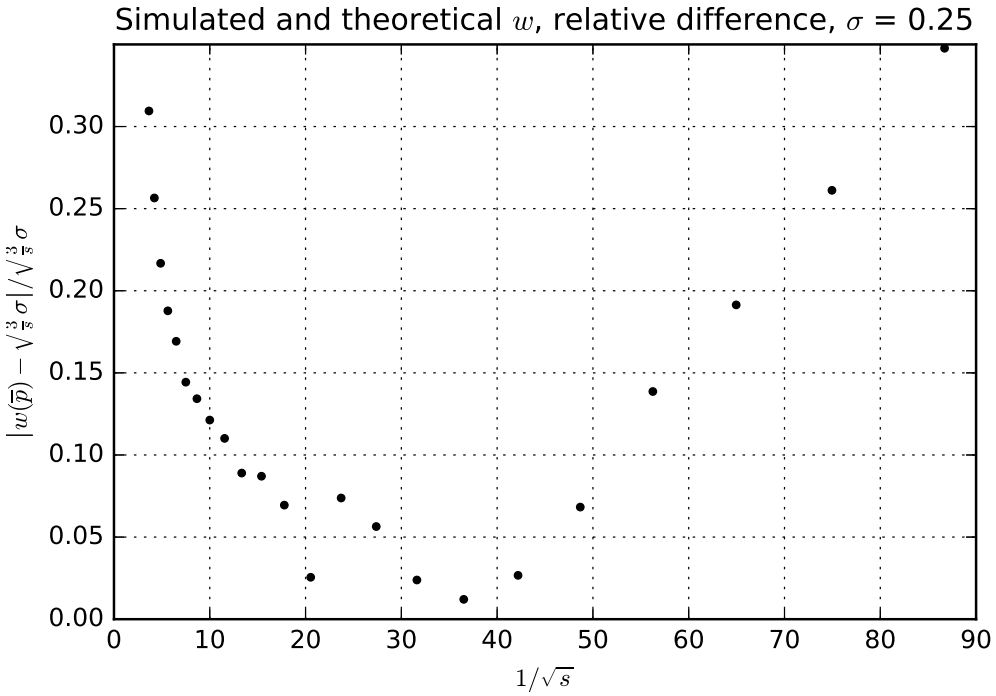


Figure A.4: The absolute value of the relative difference between the simulated and the theoretical width of the cline for different values of σ_D. The theoretical width is defined as $w = \sqrt{\frac{3}{s}}\sigma$. On the x-axis is $\frac{1}{\sqrt{s}}$.

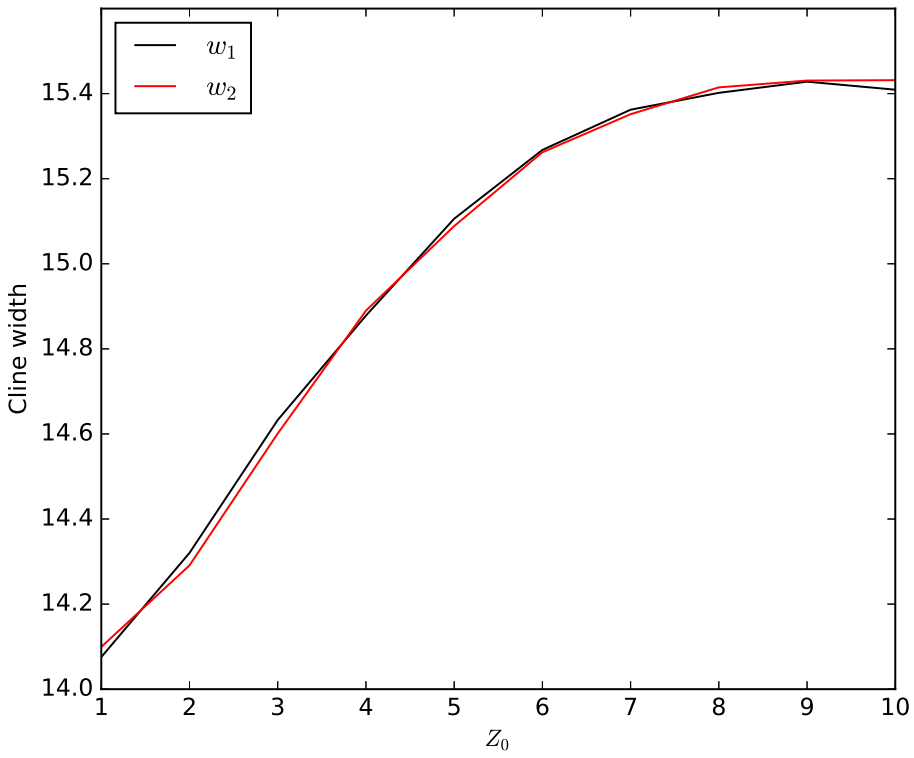


Figure A.5: Width of the two clines for increasing Z_0. $s_1 = s_2 = .1$, $r = .5$. For these values of s_1 and s_2, the width is clearly increasing as I separate the environmental transitions. As I decrease s, the trend becomes weaker.

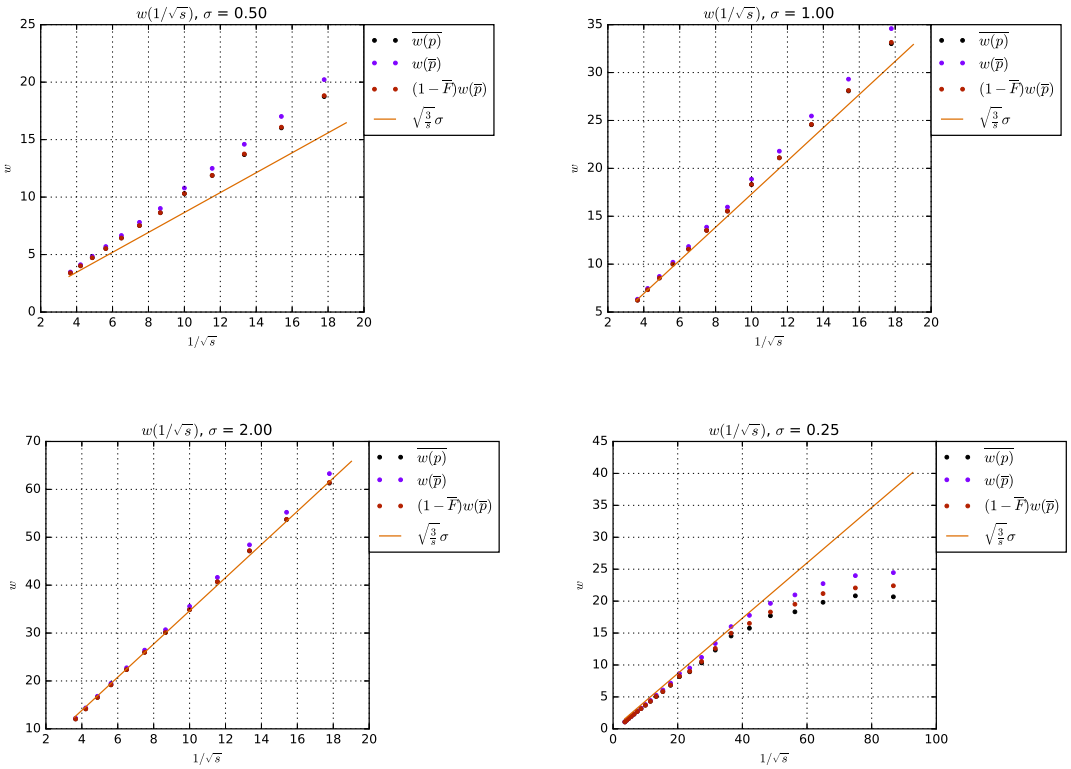


Figure A.6: w as a function of $\frac{1}{\sqrt{s}}$. We are supposed to see $\overline{w(p)}$ equal to the theoretical values, but because of the discreteness of the model this isn't the case. According to Polechova, J. and Barton, N. (2011), we are also supposed to see $(1-\overline{F})w(\overline{p}) = \overline{w(p)}$. The theoretical width is defined as $w = \sqrt{\frac{3}{s}}\sigma$

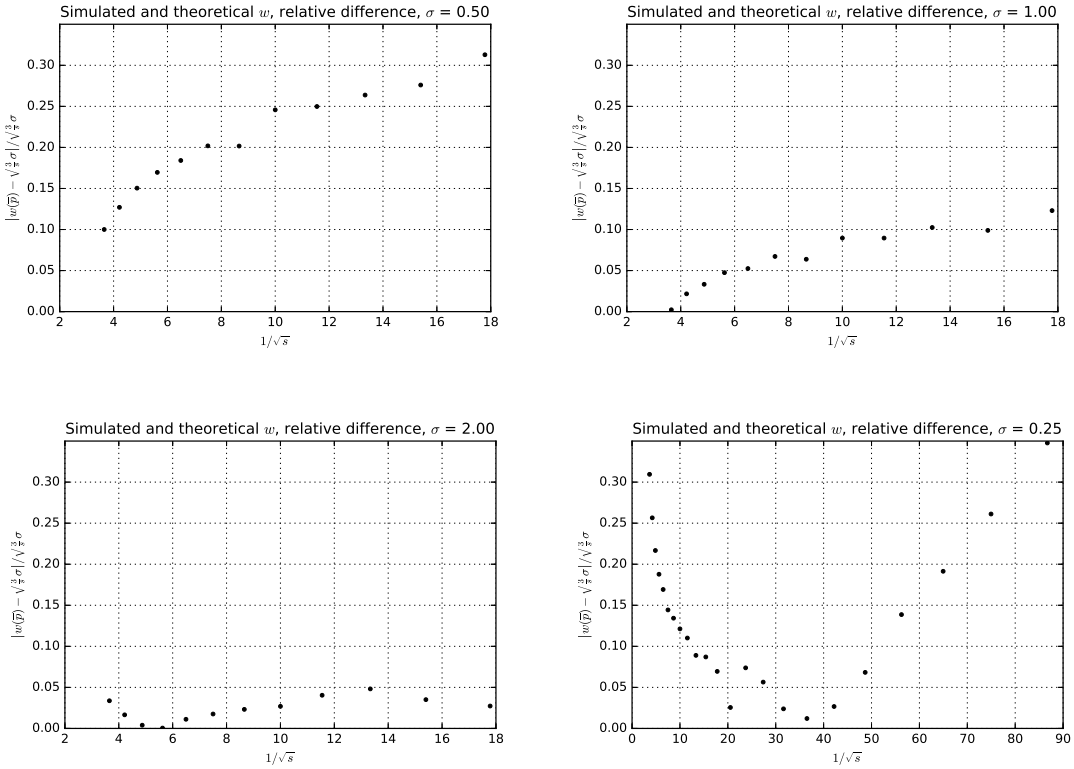


Figure A.7: The absolute value of the relative difference between the simulated and the theoretical width of the cline for different values of σ_D. The theoretical width is defined as $w = \sqrt{\frac{3}{s}}\sigma$. On the x-axis is $\frac{1}{\sqrt{s}}$.