UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PROMOTING SPECIATION THROUGH VARIABLE DOMINANCE IN GENETIC ALGORITHMS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

BY

ROBERT J. DARO
Norman, Oklahoma
2020

PROMOTING SPECIATION THROUGH VARIABLE DOMINANCE IN GENETIC ALGORITHMS

A MASTER'S THESIS APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Dean Frederick Hougen

Dr. Ingo Schlupp

Dr. Andrew H. Fagg

Acknowledgements

I'd like to thank my parents for helping me through college and always encouraging me.

I'd also like to thank my sister for keeping in touch even though she's having all kinds of adventures in New York.

And finally, I'd like to thank my dog, Cookie, whose incessant barking was the driving force behind me leaving home for school.

Abstract

Genetic algorithms are a class of search algorithms that have been around since the 1970s. Despite their age, genetic algorithms still see a great deal of use in various applications and so many efforts have gone into addressing some of their limitations. Chief among the genetic algorithm's limitations is its tendency to converge to singular answers even when working with problem spaces that feature multiple optima. Furthermore, genetic algorithms will often converge to answers that are only local optima and not globally optimum. To that end, various researchers have developed methods to both encourage exploration of a problem space as well as promote diversification of genes to avoid global convergence. The most promising method is speciation, a term and concept borrowed from biology. New research shows that genetic structures using diploid chromosome models, meiotic inheritance, and gene dominance encourage speciation without the need for expensive selection modifications. Current implementations of dominance are hypothesized to skew exploration of search spaces. This work modifies the diploid approach to encourage exploration more generally by allowing dominance to evolve alongside the genes affecting the phenotype of the chromosome. These changes encourage speciation by allowing more symmetric exploration of the search space. While the modifications are not perfect, they serve as an excellent proof of concept for future work.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Nature has guided human ingenuity and engineering for time immemorial. It is no surprise, then, that the field of computer science has also looked to biology for inspiration. One of the most well known inventions inspired by the natural world in computer science is the *genetic algorithm*, a heuristic search method based on the theory of natural selection and evolutionary biology.

Genetic algorithms are effective and easy to implement, but are not without their shortcomings. Of interest to this thesis is their tendency to converge to singular solutions in a search space, which presents several problems when applying genetic algorithms to multi-modal problems and can reduce the effectiveness of a genetic algorithm approach. A common approach to tackling this issue is to introduce some method of encouraging speciation, a term and concept borrowed from biology. Several well-developed methods of encouraging speciation exist, but they introduce additional computational complexity to genetic algorithms.

This thesis expands on previous work that used a diploid model of gene expression capable of speciating without the need for a more expensive form of selection (Booker and Hougen, 2018). This thesis introduces a new genetic model archetype that not only speciates without using a $O(N^2)$ speciation method, but also does so regardless of where the initial population starts in the search space. In doing so, this work serves as a proof of concept for a powerful new genetic model with promising applications to future technologies.

In creating this new model of gene expression, this thesis spearheads the concept of genes being affected by *dominance scalars*, numerical representations of the effect genes have on the phenotype of the organism they belong to. Furthermore, this thesis proposes a new form of mutation in which dominance scalars evolve alongside the chromosomes they modify. This new form of mutation is called *variable dominance*.

It is hypothesized that variable dominance will perform better than the dominance in Booker and Hougen (2018) in regards to exploration of the search space. This is because the key hypotheses of this thesis are that the non-variable dominance present in Booker and Hougen (2018) will lead to asymmetrical exploration of the search space, variable dominance will feature symmetrical exploration, and that the methods will both speciate at the same rate.

The experiments for this thesis are empirical studies based on the work done in both Woehrer et al. (2012) and Booker and Hougen (2018), but with some modifications to accommodate the new model and test its effectiveness.

Ultimately, this thesis broadens the understanding of how dominance and meiotic inheritance interact with speciation, and how artificial evolution might be directed to better perform in multi-modal environments. It is important for the reader to understand that the speciation method introduced in this thesis is inspired by biology, but does not strictly adhere to the mechanisms of real-world evolution. While the results of this study may be less informative of the natural world because of this, the potential applications to problem solving from a computer science perspective are believed to offset this penalty.

Chapter 2 explains how genetic algorithms work and examines some of the problems with genetic algorithms that stem from their tendency to converge early. Additionally, the next chapter describes speciation in more detail, and how computer scientists have applied speciation as a way to avoid the pitfalls of early convergence. The last two experiments discussed in chapter two are of particular relevance to this thesis. The first experiment demonstrates how genetic models featuring haploid (single chromosome) genetic models necessitate assortative mating strategies in order to speciate (Woehrer et al., 2012). The second experiment introduces a novel diploid (two chromosome) genetic model that removes the need for assortative mating to produce speciation (Booker and Hougen, 2018).

In Chapter 3, this thesis discusses some hypothesized drawbacks to the dominance model employed in Booker and Hougen (2018) and a theoretical method of addressing these drawbacks. New terms are designed to discuss dominance as a spectrum rather than in binary terms. Furthermore, a new method of gene expression is proposed that leverages dominance to encourage a broader range to speciation.

Chapter 4 focuses on the implementation of the proposed changes as well as the experimental design used to test the effectiveness of the changes.

Chapter 5 explores the results of the experiments. It is found that the introduction of variable dominance has mixed results on speciation. On one hand, variably dominant models exhibit a reduced rate of speciation compared to the model of Booker and Hougen (2018), in which dominance is tied to phenotypic value. On the other hand, the variably dominant models developed in this thesis demonstrate an ability to speciate into a wider range of niches when compared to the models employed by Booker and Hougen (2018).

In Chapter 6, the hypotheses are revisited to determine how well they hold given the results of the experiments, conclusions are drawn, and future work is discussed.

Chapter 2

Related Work

This chapter provides an overview of what a genetic algorithm is. It also covers some key limitations of genetic algorithms. After discussing the limitations of genetic algorithms, this chapter gives an overview of speciation, how it is applied in genetic algorithm research, and some well known speciation methods. Finally, this chapter discusses in detail two experiments that are the predecessors to this thesis.

2.1 The Genetic Algorithm

As early as the 1950s, scientists and mathematicians looked to evolution as an inspiration for developing new algorithm paradigms (Mitchell and Forrest, 1994). By the end of 1975, John Holland of the University of Michigan had cemented the fundamental structure of genetic algorithms (Holland, 1975; Mitchell and Forrest, 1994). While genetic algorithms have seen many variations in form and application since Holland's time, they follow this general procedure:

1. Generate an initial population of *chromosomes*, with each chromosome rep-

resenting a solution to the problem at hand.

2. Assign a fitness to each chromosome. *Fitness* is a value representing the likelihood of a particular chromosome to pass on its genes.

3. Mate the solutions together by applying selection, crossover, and mutation (each defined later in this chapter). The offspring of this mating produces the next generation of candidate solutions.

4. Repeat steps 2 and 3 until a stopping criterion is reached.

To understand how these steps can lead to optimal answers, first we need to understand the basic data model underlying genetic algorithms: the chromosome. A chromosome in the genetic algorithm has this basic structure:

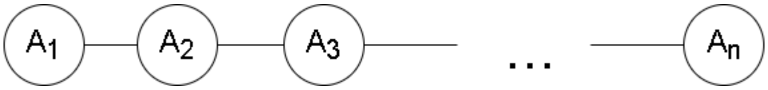

Figure 2.1: A depiction of a chromosome in a genetic algorithm.

As can be seen in Figure 2.1, $A_1, A_2, etc.$ are arguments to a function of some kind; this function is called the *objective function.* The objective function is the function representing the problem we aim to solve. The domain of this function is called the *search space.* This string of arguments is conceptually thought of as a chromosome. The goal is to eventually evolve a chromosome whose arguments maximize the problem function we are evaluating them on. We do this through a process reminiscent of natural selection.

Natural selection, in the genetic algorithm sense, has three basic components to it. The first part of the process is assigning fitness to a chromosome. In the traditional genetic algorithm model, fitness is assigned explicitly to each chromosome every generation by a *fitness function.* Exactly what the fitness function evaluates depends on the application. In some applications, the ideal answer is known, but the steps to get there are not. In this case, the fitness function could very well measure a given chromosome's performance compared to the ideal answer and assign higher fitness to the better performing chromosomes. In other applications, the ideal answer is unknown; in this case, fitness must be assigned to chromosomes relative to the performance of their peers. It is important to note that fitness and performance are not necessarily interchangeable. For instance, sometimes the fitness function rewards novelty, rather than performance (Conti et al., 2017). Sometimes fitness is implicitly assigned rather than explicitly; this is particularly apparent in simulations of biological evolution in which the "fitness function" is actually just low performance individuals failing to thrive rather than an explicit call to a well defined function (Woehrer et al., 2012). Regardless of the mechanism, fitness is a key part of the genetic algorithm's ability to make incremental improvement towards an optimal solution.

The next part of the natural selection process is the selection of genes to be passed on to the next generation. The key part of selection, regardless of the particular mechanism employed, is that more fit individuals should be more likely to pass on their genes than less fit individuals. Examples include:

- *Roulette wheel selection*, in which an individual's chance of being selected to reproduce is proportional to its fitness relative to its peers.

- *Tournament selection*, in which a sample of the population is taken and ranked. With this ranking, winners are chosen with a probability related to their positions in the ranking, often with the highest fitness individual in the tournament selected with probability 1 (higher rank means a higher probability of mating).

- *Reward-based selection*, in which the cumulative reward obtained by an individual determines its probability of reproducing. This model also allows a child to inherit some of its parent's fitness (Loshchilov et al., 2011).

Crossover, in genetic algorithms, is the mixing of chromosomes to produce new individuals. Often this process is stochastic. For example, the chromosome produced by the mating of two individuals may have a 50% chance to inherit each of its genes from either parent.

The final ingredient to a successful genetic algorithm is *mutation*, in which chromosomes are changed in some stochastic way to increase genetic diversity in the mating pool. Genetic diversity is important as it helps to avoid early convergence to an non-optimal solution. Generally speaking, the offspring of two different chromosomes has a small chance to mutate upon creation. Under *clonal reproduction*, the copying of single individuals without mating, mutation may be applied to clones as well. Mutation may involve changing only part of a gene, or the entirety of the chromosome, but the important part of mutation is to diversify the gene pool.

Algorithm 1 shows a generalized genetic algorithm approach. The algorithm distinguishes between an organism's fitness (obtained from the fitness function) and its score (obtained from the objective function). The function INITIALIZE_RANDOM_POPULATION is assumed to produce a properly diverse starting population. The function SELECT_BASED_ON_FITNESS returns a subset of a given population based on the fitness values of the individuals and the selection strategy involved in mating. The subset returned by SELECT_BASED_ON_FITNESS will be sized according to a given fraction of the population. Similarly, the function SELECT_PARTNERS determines how many mating partners a particular individual will have and which partners they will be. MUTATE has a random chance to mutate an individual. Finally, SELECT_MAX_BASED_ON_SCORE is used to determine what individual scores the highest using the objective function.

2.2 Limitations of Genetic Algorithms

Genetic algorithms are not without limitations. Genetic algorithms tend to converge to singular, local optima in any given search space (Bäck, 1996). This creates several problems:

- Many problems are *deceptive*, in that the search space encourages convergence early on to a local optimum that is not competitive with other, global optima (Lehman and Stanley, 2010). This is thought to arise whenever lower-order schemas give misleading information about the probable average fitness about the higher-order schemas contained within them (Forrest and Mitchell, 1994).
- With some problems, such as travelling salesman problems, the knowledge of multiple optimal solutions would be beneficial. Traditional genetic algorithms fail to reliably capture more than one solution at a time.
- It can be difficult to maintain genetic diversity even with mutation, and as such, it is possible to not sufficiently explore a search space with traditional genetic algorithms.

Different solutions have been proposed and implemented to correct these limitations to varying degrees of success. For example, in many cases of deception, a better fitness evaluation metric can overcome early convergence. A good example of this would be the experiment shown in Conti et al. (2017), in which an agent is trained to walk. This paper features a problem designed to be intentionally deceptive, and then demonstrates that changing the fitness function to reward novelty, rather than performance, allows a genetic algorithm to overcome the deception. While in this case the limitations of a genetic algorithm approach could be mitigated, at least partially, by a change in the fitness function alone, in other cases we require a more radical approach to encouraging diversity while avoiding early convergence. One promising solution is to implement some form of speciation.

2.3 Speciation: A Promising Solution

Speciation, in brief, is the process that causes one species to split and become two new species (Berlocher, 1998). Generally speaking, speciation is thought to occur when subpopulations of a species are physically divided from one another, thus causing them to continue to evolve independent of one another. Over time, the differences in the mating pools bring about enough change that they are no longer the same species. However, not all forms of speciation occur through a physical separation of subpopulations. *Sympatric speciation* is the emergence of a new species from an existing species' population in the same region. Under sympatric speciation both species continue to inhabit the same environment (Smith, 1966). While there is some controversy as to whether or not sympatric speciation occurs in the wild, from a computer science standpoint sympatric speciation has the potential to overcome some of the most glaring constraints of genetic algorithms. It is important to understand that evolution in the natural world is an unguided process, and that computer scientists use objective fitness in an effort to turn evolution into a guided process (Shah and Hougen, 2019). With this in mind, it is no surprise that much of the speciation research in the field of computer science focuses on controlling speciation rather than explaining it.

Combining speciation and genetic algorithms is not a new idea, but it is still a growing field of study. Some of the more basic attempts at this are through *niching* methods. The idea behind niching methods is to actively choose what chromosomes will reproduce in order to maximize genetic diversity; these methods can be somewhat expensive, and, in general, do not find optimal solutions quickly (Yu and Suganthan, 2010). Because of some of the issues niching methods present, some researchers moved to modifications of the genetic algorithm structure (Yu and Suganthan, 2010). Such modifications include the introduction of taxon strings and other tags to indicate how similar a particular chromosome was to another one. The end goal of these mechanisms is to encourage similar chromosomes to mate with each other; however, they still require more computation time in order to compare individuals during selection (Bäck et al., 1997).

2.3.1 Example Niching Methods

NSGA-II is a good example of a niching algorithm based on distances in the search space (Deb et al., 2002). While the paper introducing the algorithm notes how NSGA-II is remarkably more performative than other popular niching algorithms, it is still limited by the need to compare every solution to every other solution.

PRISM is another example of a distance-based speciation algorithm. However, unlike NSGA-II, it takes into account common ancestors as well (Grouchy et al., 2009). An interesting part of PRISM is that as an island model it actually enforces a strict mating barrier between subpopulations with only periodic migrations of individuals between each mating pool providing genetic flow between them. Once more, this system requires every solution to be compared to one another and can become computationally expensive.

Both NSGA-II and PRISM require the programmer to specify the distance at which a species is considered distinct (Deb et al., 2002; Grouchy et al., 2009). In an effort to avoid parameters such as this, Della Cioppa et al. (2011) presents a niching method called Adaptive Species Discovery (ASD). In practice, ASD often avoids the worst case complexity of N^2 by having species defined by key individuals, not large collections. However, in order to do this ASD employs a more complicated distance function that relies on finding hills and valleys in the search space. This function requires a good set of sampling points to function correctly, and can be computationally expensive (Della Cioppa et al., 2011).

A recurring theme in the algorithms discussed is the need to compare each individual in a population to every other individual. The next section discusses a paper of particular importance to this thesis. In this paper, evidence is collected which suggests that the traditional model for genetic algorithms requires modified selection in order to exhibit speciation.

2.4 Sexual Selection Leads to Speciation

One of the predecessors to this work is Woehrer et al. (2012), which aims to explore how speciation can be induced through the actions of individual agents in the environment as well as a preference for mating with similar individuals (Woehrer et al., 2012). In order to accomplish this, Woehrer et al. (2012) develops a simulation of an island populated by finches. These finches' beak sizes determined what resources they could exploit. The experiment is briefly detailed in this section.

Woehrer et al. (2012) features five, key hypotheses. The hypotheses, in order, are:

1. Speciation will occur in the case of a bimodal resource distribution and assortative mating.
2. In the case of a bimodal resource distribution and random mating, only one resource will be exploited and speciation will not occur.
3. With a uniform resource distribution and assortative mating, speciation may occur from the assortative mating alone.
4. With a uniform resource distribution and random mating, no speciation should occur.
5. Increasing the availability of resources will lead to larger population sizes and more stability, but not affect the behavior of any speciation.

2.4.1 Experimental Design

The experiment consisted of an artificial island, 100 × 100 units in size, populated with birds and seeds. Birds have the following properties: age, beak size, energy level, and gender; beak size is determined by a bird's genetics. Birds have a maximum energy capacity of two units as well as a lifespan of four years. Seeds have an energy value, size, and location.

A single run of an experiment consists of up to 1000 generations of birds; a run can end before 1000 generations if all birds die out. At each generation the age, beak size, energy, gender, and mating count are recorded for all birds.

Two different seed distributions are used. The first is a bimodal seed distribution. Under this distribution, seed sizes are drawn from two Gaussian distributions with means of 3 and 8 and variances of 0.5. Both of the Gaussian distributions have an equal chance of being selected each time a seed is generated. The second distribution is a uniform distribution from 1 to 10. The number of seeds depends on the resource availability; the small population condition calls for 5000 seeds each generation, but the large population condition calls for ten times that number.

Each generation is split into two phases, a dry season and a mating season. Dry seasons consist of 100 consecutive days. Each day during the dry season, each bird will attempt to feed. Individuals first pick a random 10×10 region of the island to search, then search within that region for an acceptable seed. Birds can only eat a seed whose size falls within one unit of their beak size.[1] Seeds contain a uniform random amount of energy between zero and two. A feeding attempt itself costs 0.1 units of energy to perform. After feeding, if a bird's energy level falls below zero, it is removed from the population and considered dead. At the end of a dry season, mating takes place. During a mating season, food is considered abundant. Because of this, the only part of the season simulated is the mating attempt as all birds are assumed to survive.

Two forms of sexual selection are allowed: assortative mating or random mating. Under assortative mating rules, females choose a mate within one unit of their beak size. Males are randomly chosen if multiple qualify. Under random mating, a mate is selected at random from the whole population. In either case, males are restricted to only mating five times per generation in order to ensure genetic diversity. Offspring have a beak size that is the average of the sizes of

---

[1] For example, a bird whose beak size is 4 must eat a seed of size [3, 5].

the parents' beaks plus some mutation in the form of Gaussian noise (mean 0, variance 0.2). Gender is assigned with an equal chance of being either male or female, and the energy level starts at zero. Finally, after reproduction is finished, the age of all birds is incremented by one and birds with an age over four are removed. After the removal of the old birds is complete, the simulation removes all uneaten seeds and begins the next dry season.

For the cases with 5000 seeds, 48 trials were run for each combination of seed distribution and mating strategy. For the cases with 50000 seeds, 24 trials were run.

2.4.2 Results

The results of the experiments largely line up with the hypotheses put forth. In the case of bimodal seeds and assortative mating, 31 out of 48 of the trials resulted in a speciation event that created a stable population that persisted throughout the trial. In the case of bimodal seed distribution and random mating, 41 of the trials resulted in only one branch being present, with the remaining 7 yielding complete extinctions. With a uniform seed distribution and random mating, constant branching took place, but these branches were highly unstable and tended to disappear in a few generations. With a uniform distribution and random mating, the norm was for there to be one population of birds, with the only exceptions being 13 trials that ended by extinction.

In the case of the 10× seeds condition, the only tangible difference came in the case with uniform seed distribution and assortative mating. In this case the unstable branching was replaced with four distinct populations featuring frequent interbreeding.

Woehrer et al. (2012) illustrates that without modified selection parameters, the standard genetic algorithm fails to speciate. In order to encourage speciation without relying on expensive comparisons during selection, a new genetic model needed to be designed. The next paper discussed introduces a new genetic model that overcomes the need for modified selection.

2.5 Meiotic Inheritance: A New Model for the Genetic Algorithm

Booker and Hougen (2018) builds on what Woehrer et al. (2012) demonstrates; using the same environment, Booker and Hougen (2018) investigates a very novel approach to speciation by changing the fundamental model of a genetic algorithm. The model used by Woehrer et al. (2012), like most models used in genetic algorithms, features a genetic structure with only one chromosome. However, many species of animals inherit two sets chromosomes from their parents, not just one. Booker and Hougen (2018) develops a new framework based on a two-chromosome model, hypothesizing that this change in combination with a concept of gene dominance could accomplish speciation. Booker and Hougen (2018) calls this new form of inheritance *meiotic inheritance*.

The two relevant hypotheses put forth in the paper are as follows:

1. Meiotic inheritance will allow for speciation without assortative mating.
2. Complete dominance, which will be explained below, will amplify the effect that meiotic inheritance has on speciation.

2.5.1 Experimental Design

The experimental design Booker and Hougen (2018) uses is similar to the one employed in Woehrer et al. (2012). The largest changes were designing a new genetic structure, new forms of inheritance, and a modified system of gene expression. Furthermore, only 5,000 seeds were used.

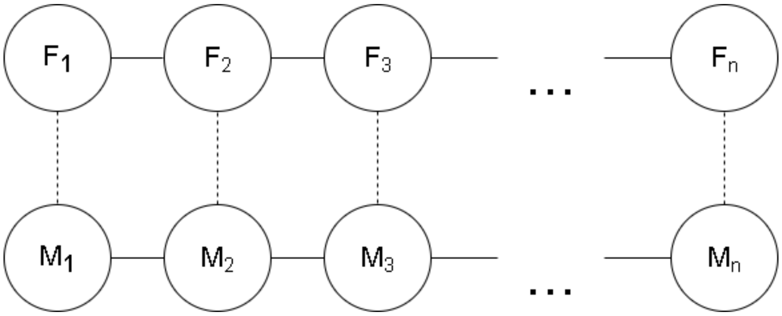

Figure 2.2: A depiction of a diploid genetic algorithm.

In the new model, an individual has two chromosomes, not one, as shown in Figure 2.2. The genes denoted as $F_1, F_2, ..., F_n$ represent the genes received from the father of an organism, and the genes labeled $M_1, M_2, ..., M_n$ represent the genes inherited from the organism's mother. For the simulation, chromosomes contain one-hundred genes. The starting population's chromosomes are initialized randomly, with each gene represented by a floating-point value chosen from a uniform distribution with a mean of 0.055 and a range of 0.245.[2] When two birds mate, one chromosome is randomly selected from each parent to construct the offspring. During the mating process, each gene copied from a parent has a 10% chance to be mutated. When mutated, a gene has a random value taken from a

---

[2] Booker and Hougen (2018) explains these values were chosen to mimic the behavior of the single-gene model employed in Woehrer et al. (2012), with the final beak size being the average of two sums of 100 genes: $\frac{1}{2}\sum_{i=1}^{200} X_i$. The distribution of the beak sizes is considered normal by the central limit theorem. The mean and standard deviation are thus $\frac{1}{2}(200 * 0.055) = 5.5$ and $\frac{1}{2}\sqrt{\sum_{i=1}^{200} \text{Var}(X_i)} = 0.5$, respectively. X_i denotes the value for gene i in chromosome X.

uniform distribution with a range of 0.310 and a mean of 0 added to it.[3]

The experiment featured two forms of inheritance for the diploid model: blending and discrete inheritance. Under blending inheritance, an offspring's genes are the average of its inherited genes, while in a discrete model its genes are kept distinct from one another. The blending model designed in Booker and Hougen (2018) mimics the performance of the Woehrer et al. (2012) model, and is more similar to the traditional genetic algorithm chromosome. The model of inheritance of interest to this thesis is the discrete version, and thus it will be the form of inheritance discussed in this section. This is because discrete inheritance produced speciation under random mating, as explained later in this section.

One of the largest consequences of switching to a diploid model is the need for a new form of expression. Booker and Hougen (2018) considers two different forms of expression: incomplete dominance and complete dominance. In both forms of expression, the phenotypic value of a chromosome is determined by summing up the values of the genes making up that chromosome. Where the models differ is how these values are used. In the incomplete dominance case, the beak size of a bird was the average of its two chromosome values. For complete dominance, the larger of the chromosome values was used as the beak size, with the smaller chromosome value not affecting the phenotype of a bird but still having a chance to be passed on to offspring.

2.5.2 Results

The introduction of a two-chromosome genetic structure did, in fact, encourage speciation. Out of 48 trials with random mating and a bimodal seed distribution,

---

[3] These values were determined empirically by the researchers in order to mimic the mutation rate of the single-gene model used in Woehrer et al. (2012).

33 of them featured speciation using complete dominance. All this shows that by switching to a diploid model with some concept of dominance, speciation is attainable without the need for some kind of assortative selection process. This concept could radically change how genetic algorithms operate in the future, as more costly methods of speciation may be avoided entirely by simply adopting a more expressive genetic structure.

In particular, complete dominance shows great potential in the experiments of Booker and Hougen (2018). However, complete dominance depends on the user to determine what phenotype is more dominant than another, and as such introduces some user bias in how the model operates. This thesis aims to improve on complete dominance by creating a model that is devoid of this bias.

Booker and Hougen (2018) opened a doorway to new methods of encouraging speciation, and research needed to be done to determine how best to utilize this new method to effectively cover search spaces more generally. This thesis aims to do just that, and the details of this endeavor are detailed in the following chapters.

Algorithm 1: Genetic Algorithm Example

```
Input: $MaxGenerations$ // The maximum number of generations until solution is returned
$CrossoverRate$ //The fraction of the population that is generated by crossover [0, 1]
Output: $BestIndividual$ // The best performing individual found during the last generation of the search
$GenerationNumber \leftarrow 0$;
$Population \leftarrow$ INITIALIZE_RANDOM_POPULATION();
while $GenerationNumber < MaxGenerations$ do
    foreach $P \in Population$ do
        $P.Fitness \leftarrow$ FITNESS_FUNCTION($P$);
    end
    $MatingPopulation \leftarrow$ SELECT_BASED_ON_FITNESS($Population$, $CrossoverRate$);
    $CloningPopulation \leftarrow$ SELECT_BASED_ON_FITNESS($Population$, $1 - CrossoverRate$);
    $NewPopulation \leftarrow \emptyset$;
    foreach $P \in MatingPopulation$ do
        $MatingPartners \leftarrow$ SELECT_PARTNERS($MatingPopulation$);
        foreach $Q \in MatingPartners$ do
            $M \leftarrow$ CROSSOVER($P$, $Q \in MatingPopulation$);
            MUTATE($M$);
            $NewPopulation$.add($M$);
        end
    end
    foreach $P \in CloningPopulation$ do
        MUTATE($P$);
        $NewPopulation$.add($P$);
    end
    $Population \leftarrow NewPopulation$;
    $GenerationNumber \leftarrow GenerationNumber + 1$;
end
foreach $P \in Population$ do
    $P.Score \leftarrow$ OBJECTIVE_FUNCTION($P$);
end
$BestIndividual \leftarrow$ SELECT_MAX_BASED_ON_SCORE($Population$);
return $BestIndividual$;
```

Chapter 3

Hypotheses

In order to capture a deeper understanding of the effects that a dominance schema might have on speciation, a new concept is proposed. A *dominance scalar* is a numerical representation of how dominant one chromosome is over another, or in other words a measurement of how much weight a chromosome or gene has on the expression of a organism's genes. In Booker and Hougen (2018), dominance was either directly tied to the phenotypic value a chromosome possessed, or it was equal for all chromosomes and independent of genetics entirely. While this led to fascinating and field-changing discoveries, it also opened up research into how manipulations of the dominance scalars present in a population might affect speciation, or even alter its course. This thesis delves into this topic in order to demonstrate a new dominance model that treats dominance scalars as pieces of a genetic algorithm open to evolution themselves. Previous research with diploid models indicates that variable dominance schemes tend to outperform schemes where dominance is considered fixed (Smith and Goldberg, 1992). However, unlike in Smith and Goldberg (1992), this thesis proposes a dominance model that blends incomplete dominance and complete dominance. Under the proposed model, chromosomes will affect the phenotype of an individual proportionally to their dominance scalars, as opposed to the more “all or nothing” implementation in Smith and Goldberg (1992). This paradigm of dominance scalars changing with the genes they affect is dubbed *variable dominance*.

3.1 Hypothesis H_1: Trait-Based Dominance is Directional

The first hypothesis addresses a theorized limitation of dominance being directly related to expression. It is proposed that with larger beaks being more dominant, exploration of the search space becomes feasible in the direction of smaller beaks, but stunted in the direction of larger beak. To illustrate this, consider a bird with one large-beak chromosome and one small-beak chromosome. The small-beak genes could very well be nonviable, but they could still persist in the gene pool as they have no effect on the bird’s survival chances. However, if the large-beak genes would leave the bird to starve, the bird starves; the large-beak genes dictate beak size and thus decide whether or not the bird can live to pass on its genes. Intuitively, we can see that large beaks being dominant leaves smaller-beak genes more likely to survive, as not every bird carrying them expresses them. It also follows that a higher survival rate allows for a greater degree of exploration. This conjecture leads to the first hypothesis of this thesis:

H_1: Under the model with large beaks being dominant, exploration will be asymmetrically skewed toward niches that require small beaks.

3.2 Hypothesis H_2: Variable Dominance is Symmetric

The second hypothesis addresses how variable dominance will encourage symmetric exploration. Consider again the case of a bird with one large-beak chromosome and one small-beak chromosome. However, this time dominance is evolved as part of a chromosome rather than tied to beak size. It could be that for this bird the large-beak chromosome has a lower dominance scalar than the small-beak chromosome; in that case the bird's genes are free to explore in the direction of larger beaks rather than smaller ones. Alternatively, the dominance scalars could be close to or identical to what they would have been under the model in which large beaks are dominant, in which case the genes are free to explore in the direction of smaller sizes. While freedom of exploration in either direction for an individual is not guaranteed, as a population it becomes very likely that at least a few low-dominance chromosomes will exist on both sides of the size spectrum, thus facilitating more symmetric exploration of the search space. Thus, the second hypothesis of this thesis becomes:

H_2: Under the variable gene dominance model, exploration will be symmetric. Both smaller and larger beak sizes will be explored.

3.3 Hypothesis H_3: Trait-Based Dominance and Variable Dominance are Comparable

The third hypothesis addresses a possible shortcoming to variable dominance. Under the model used by Booker and Hougen (2018), speciation was a common occurrence when using meiotic inheritance, but not guaranteed. The biggest concern with variable dominance is if it will reduce the rate of speciation as compared to dominance based on expression. It is possible that with the bimodal seed distribution, too many pairings between large-beak and small-beak chromosomes with similar dominance scalars could produce many individuals who die out due to the lack of seeds in the middle ground between the modes, and these die outs could limit genetic diversity too much to speciate. However, it can also be reasoned that variable dominance will actually encourage fewer extinctions to occur. Take, for instance, a scenario in which large beaks are dominant, but all large beaked individuals have died. In this scenario, it becomes very hard for the large beak niche to become filled again, as no genes dictating large beak sizes are present in the population. If there had been any, the birds that had them would have had large beaks and thus died. However, in a variable dominance model, some of the small beaked birds are likely to have large-beak chromosomes that are not expressed because of their low dominance scalars. All it takes is pairings between several of these individuals to produce a viable population of large-beaked individuals again. Considering the competing factors, the third hypothesis of this thesis states:

H_3: Variable gene dominance will allow for speciation at a comparable rate to a model in which size determines dominance.

3.4 Hypothesis Testing

All hypotheses listed above are defined in such a way that two sample hypothesis testing should be used. For H_1 and H_2, an experiment must be formed that can test each model on its ability finding a larger seed size and smaller seed size.

For H_1 to hold, a significant difference between the performance of finding the smaller seed niche and finding the larger seed niche will need to be detected. In contrast, for H_2 to hold, no significant difference should be detected. In order to test a model's ability to locate a niche compared to another model, several runs will be performed. The number of times the models located the additional niche in each collection of runs will be recorded and compared to each other to determine if there is a statistically significant difference in performance.

For H_3, the variable dominance model will need to be compared to the size-based dominance model to determine if there is a significant difference in the rate of speciation; if no difference is detected, the hypothesis holds. In order to test the speciation rate, the process will be similar to how H_1 and H_2 are tested. The key difference is the total number of speciation events, regardless of whether or not any additional niche was located, will be recorded. This will determine if there is a significant difference in the likelihood of speciation to occur at all between models.

Chapter 4

Experimental Design

The experimental design of this project is similar to the one employed in Booker and Hougen (2018). The experiment features a 100 × 100 unit sized island, a dry season of one-hundred days, and a mating season that follows. However, there are some key changes made to test the hypotheses put forth in the previous chapter. These changes include two new genetic structures, a new form of expression, and two more seed distributions. This chapter will detail the changes made to the environment used in Booker and Hougen (2018) as well as the new speciation mechanisms developed as part of this thesis. First, this chapter will briefly discuss the unchanged characteristics of the environment. Next, it will discuss the modifications made to test the hypotheses.

4.1 Unchanged Environmental Properties

The island used in this experiment is exactly the same as the one used by Booker and Hougen (2018) and Woehrer et al. (2012): 100 × 100 units and populated with birds and seeds. No changes were made to these properties as it was decided not to yield any useful information about speciation.

Birds again have the following properties: age, beak size, energy level, and gender. Their lifespans are four years, and they have a maximum energy capacity of two units. Seeds remain unchanged as well, being defined by their energy value, size, and location as in the previous experiments.

As before, a single run of an experiment consists of up to 1000 generations of birds. Runs can end before 1000 generations if all birds die out. At each generation the beak sizes of all birds are recorded.

The mating strategy employed is random mating. Female birds select a mate each year at random and males are restricted to five mates per year. Since the hypotheses of this thesis concerned random mating only there was no need to include the assortative mating from the previous experiments.

4.2 New Genetic Structures

In order to capture the idea of a variable dominance, a new genetic structure needed to be designed that could allow for the evolution of dominance scalars. Two such structures were tested as part of this experiment.

The first model developed for exploring variable dominance is a *poly-weighted* approach. Under this model, each gene in a chromosome carries a weight that determines the effect it will have in expression. The initial weights of the chromosomes are integers randomly assigned from a uniform distribution ranging from 1 to 100. Other than these weights, the model is identical to the one used by Booker and Hougen (2018).

The second model is similar to the first; however, under this *mono-weighted* model, all the genes in a chromosome share the same weight rather than having independent weights. This weight still ranges from 1 to 100, and is assigned at random for the initial population.

In addition to two different structures being used, two forms of mutation are tested on the weights. Under the *high variance* model, when a weight is mutated it is changed to another value drawn from a uniform distribution ranging from 1 to 100. Under the *low variance* model, when a weight is mutated the weight's value has a 50% chance to be shifted 5 points up and a 50% chance to be shifted down 5 points. If the shift up would cause a weight to become greater than 100, it is capped at 100 instead. Likewise, a weight will never fall below 1.

For poly-weighted chromosomes, each weight independently has an 8% chance of being mutated during reproduction. For mono-weighted chromosomes, the single weight that all genes share has an 8% chance of being mutated during reproduction.

4.3 New Form of Expression

The new genetic structures demand a new form of expression. For poly-weighted chromosomes, a bird's beak size is determined using the formula

$$\sum_{i=0}^{100} \frac{X_i * M_i + Y_i * F_i}{X_i + Y_i} \tag{4.1}$$

where X_i indicates the weight for gene i in the chromosome originating from a bird's mother while M_i indicates the value of that gene. Similarly, Y_i indicates the weight for gene i in the chromosome originating from a bird's father, and F_i indicates the value of that gene. The goal of this equation is to ensure that genes with greater weight than their counterpart have more control over the beak size of the individual they belong to.

For mono-weighted chromosomes, the equation becomes

$$\sum_{i=0}^{100} \frac{X * M_i + Y * F_i}{X + Y} \tag{4.2}$$

Here, only one weight is used for each chromosome, so no indices are needed for them. This can also be seen as a special case of the first chromosome where $X_1 = X_2 = ... = X_{100}$ and likewise for the father's chromosome.

4.4 New Seed Distributions

Seeds are once again the resource that the birds will evolve to exploit. In this experiment, the uniform seed distribution is not explored as the new genetic structures are more aimed towards finding peaks in the problem space than encouraging speciation for its sake alone. Consequently, the bimodal seed distribution makes a return as one of the seed distributions to be tested. A bimodal distribution is made by drawing 2500 seeds from a Gaussian distribution with a mean of 3 and standard deviation of 0.5 and another 2500 seeds from a Gaussian distribution with a mean of 8 and a standard deviation of 0.5.

Two more distributions are employed in this study. The first, called *trimodal large*, adds an additional niche to the bimodal distribution centered at 13. This is accomplished by drawing 2500 more seeds from a Gaussian distribution with a mean of 13 and a standard deviation of 0.5. The second, called *trimodal small*, adds a niche centered at -2 and is again formed by drawing 2500 more seeds from a Gaussian distribution similar to the other niches. Note that seed and beak sizes are arbitrary integers with little meaning beyond what resources a bird has access to, so negative numbers work perfectly fine. These additional distributions will be used to test how well a given method can explore in a particular direction in the problem space. A model's success rate in finding and exploiting these additional niches will allow us to answer hypotheses H_1 and H_2. Note that the starting population of birds will have chromosomes specifying beak sizes identical to the starting population of the experiments in Booker and Hougen (2018). Considering that the chromosomes will have a mean value of 5.5 and a variance of 0.5, it is very unlikely a significant portion of the starting population will begin with beaks sized to exploit these additional niches. Because of this, these distributions will require the model to explore the search space far more than the bimodal distributions did in order for all resources to be exploited.

Algorithm 2 gives an overview of the island simulation. The inputs to this algorithm are the seed distribution to use (bimodal, trimodal large, trimodal small) and the genetic model (complete dominance, incomplete dominance, etc.). Note that both Woehrer et al. (2012) and Booker and Hougen (2018) used additional seed distributions, such as uniform seed distributions or larger numbers of seeds. The function INITIALIZE_RANDOM_POPULATION produces the initial population of birds randomly as described in Booker and Hougen (2018). The function GENERATE_SEEDS_FOR_ISLAND produces a two-dimensional array of locations, with each location having a list of seeds. The function FIND_SEEDS_IN_RANDOM_AREA takes this array and returns a list of the seeds found in a random area as defined in Booker and Hougen (2018).

Algorithm 2: Island Simulation

```
Input: $SeedDistribution$ // The seed distribution for this run
       $GeneticModel$ // The genetic model to use
$Year \leftarrow 0$;
$Population \leftarrow$ INITIALIZE_RANDOM_POPULATION();
$PopulationSize \leftarrow Population$.size();
$OutputLogs \leftarrow \emptyset$;
while $Year < 1000 \wedge PopulationSize > 0$ do
    $Island \leftarrow$ GENERATE_SEEDS_FOR_ISLAND($SeedDistribution$);
    for $Day = 0; Day < 100; Day \leftarrow Day + 1$ do
        foreach $Bird \in Population$ do
            $Bird.energy \leftarrow Bird.energy - 0.1$;
            $FoundSeeds \leftarrow$ FIND_SEEDS_IN_RANDOM_AREA($Island$);
            SHUFFLE($FoundSeeds$);
            foreach $Seed \in FoundSeeds$ do
                if $Bird.beakSize - 1 \leq Seed.size \leq Bird.beakSize + 1$
                 then
                    $Bird.energy \leftarrow$ MIN(2, $Bird.energy + Seed.energy$);
                    $Island.remove(Seed)$;
                    BREAK;
                end
            end
        end
    end
    // recording the beak sizes and energy levels
    // of this generation of birds and removing dead birds
    foreach $Bird \in Population$ do
        $OutputLogs.add$(TUPLE($Year$, $Bird.energy$, $Bird.beakSize$));
        $Bird.age \leftarrow Bird.age + 1$;
        if $Bird.energy \leq 0 \vee Bird.age \geq 4$ then
            $Population.remove(Bird)$;
        end
    end
    $NewMembers \leftarrow$ MATE($Population$);
    $Population \leftarrow Population \cup NewMembers$;
    $Year \leftarrow Year + 1$;
    $PopulationSize \leftarrow Population$.size();
end
$FileHandle \leftarrow$ a file to write to;
WRITE($FileHandle$, $OutputLogs$);
```

Chapter 5

Results and Discussion

This chapter details the methods used for analysis of the results, specifically with regards to how speciation is measured and what determines statistical significance. Furthermore, this chapter discusses the results of the experiments and how well the models performed relative to each other.

5.1 Methods Used

Several methods of analysis are used in this study. For clarity, this section outlines what methods are used to determine population counts as well as statistically significant differences.

5.1.1 Population Counts

To determine whether or not speciation occurs, this thesis employs a method using histograms. The beak sizes of all birds that survived for each generation are binned. In the case of the bimodal seed distribution, beak sizes in the interval [2, 4] were binned together and beak sizes in the interval [7, 9] were binned together. Speciation is said to have occurred if both bins contain at least 15% of the total number of birds which survived a year for that run. Any run that failed to complete 1000 years before extinction was considered to have featured no speciation.

In the case of the trimodal large seed distribution, the bins were similar to the bimodal case except beak sizes in the interval [12, 14] are also binned together. The model is said to have found this additional niche if at least 15% of the total number of birds that survived fall into the larger bin. Similarly, for the trimodal small seed distribution, beak sizes in the interval [−3, −1] are binned together. The model is said to have found the additional niche if 15% or more of surviving birds fell within the smaller bin.

For each model, a total of 96 runs are performed for each seed distribution. Runs are independent of one another.

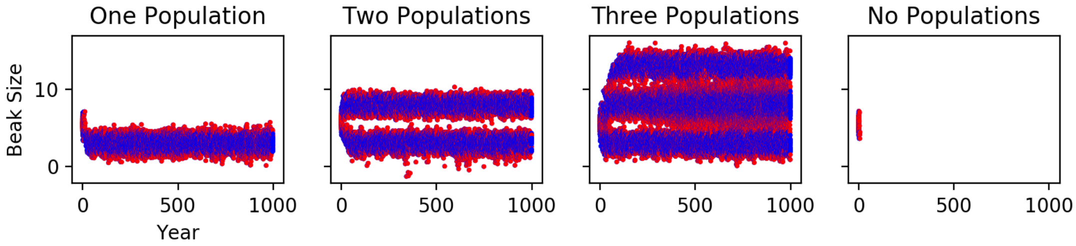

Figure 5.1: Temporal plots of select runs. Blue dots indicate birds that survived that year. Red dots are birds that died that year.

Figure 5.1 shows the results of four runs temporally, with the horizontal axis representing years and the vertical axis the phenotypic beak size. Each bird is represented by a single dot for each year it of its existence, with blue indicating surviving that year and red indicating death. The leftmost plot in Figure 5.1 shows a typical run in which the initial population of birds, with mean beak size approximately 5.5, finds the small seed niche (with a mean seed size of 3.0) and the population survives for all 1000 years. The surviving birds mostly have beak sizes between 2 and 4 (central blue band) whereas birds outside this range starve (flanking red bands). The plot second from the left shows a run in which the initial population finds both the small seed niche (with a mean seed size of 3.0) and the larger seed niche (with a mean seed size of 8.0). Two distinct populations are seen, with one population having beak sizes between 2 and 4 and the other population having beak sizes between 7 and 9. The plot second from the right in Figure 5.1 shows a run in which three seed size niches are found, having mean seed sizes of 3.0, 8.0, and 13.0. Here three distinct populations of birds are formed, whose beak sizes range from 2 to 4, 7 to 9, and 12 to 14, respectively. Finally, the rightmost plot in Figure 5.1 shows a run in which an early extinction occurred, and the simulation did not reach 1000 years.

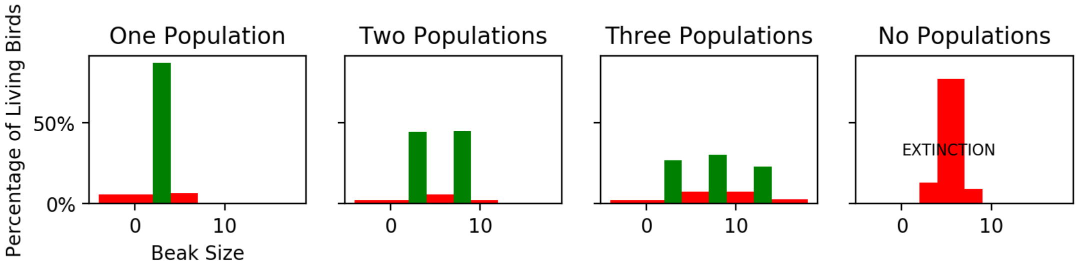

Figure 5.2: Example histograms used for population counting. Green bars indicate that the bin contains at least 15% of the birds present in the population. Red bars indicate that the 15% cutoff was not reached.

Figure 5.2 shows how the temporal results translate to the histogram method. Note that distinct populations are reflected in the histograms created from them. In the leftmost plot of Figure 5.2, created from the same data that the leftmost plot in Figure 5.1 was created from, the majority of surviving birds fall into one bin. Similarly, the second plot from the left in 5.2 shows that surviving birds generally fell into two bins, and in the second plot from the right, the surviving birds generally fell into three bins. The rightmost plot in Figure 5.2 depicts the case in Figure 5.1 in which early extinction occurred and is marked as such.

5.1.2 Statistical Significance

A key part of evaluating the hypotheses lies in the ability to detect if a perceived difference in the methods is statistically significant. For H_1 and H_2, success is defined as whether the population populated the additional niche introduced in either of the trimodal distributions. In the case of H_3, success is whether the population speciated at all. Since it is common for a model to feature a success rate of 0, chi-square and G-tests can be misleading and give p-values that are too low (McDonald, 2014). Instead, this thesis will use Fisher's exact test as calculated in R. The main drawback to this method is that the unfixed rates of speciation will cause the test to become more conservative, in which case the p-value reported becomes the upper bound for the actual p-value (McDonald, 2014). This was decided to be preferable to using chi-square as a statistically significant p-value reported by Fisher's exact test will more accurately reflect whether a difference between the models exists. For each comparison, the p-value of the observed counts reported using a two-tailed hypothesis test. A p-value less than 0.05 will be considered evidence to reject the null hypothesis that there exists no difference in speciation between the models in question.

5.2 Preliminary Comparisons of New Models

This section presents preliminary results of the new models. These comparisons were made to show which of the new models seem to speciate the most.

As seen in Table 5.1 and Table 5.2, both the mono-weighted and poly-weighted models display similar performance under both forms of mutation. Also, as tables 5.3 and 5.4 show, the mono-weighted models clearly outperform the poly-weighted models. As such, it is the mono-weighted models that will be used for hypothesis testing.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted Low Variance	49	47
Mono-Weighted High Variance	36	60

Table 5.1: A comparison of the rates of speciation using the mono-weighted models. Here the key difference is whether low variance or high variance dominance mutation was used.

Model	**Times Speciated**	**Times Did Not**
Poly-Weighted Low Variance	4	92
Poly-Weighted High Variance	6	90

Table 5.2: A comparison of the rates of speciation using the poly-weighted models. Here the key difference is whether low variance or high variance dominance mutation was used.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted Low Variance	49	47
Poly-Weighted Low Variance	4	92

Table 5.3: A comparison of the rates of speciation using the mono-weighted model versus the poly-weighted model. Here both models use low-variance dominance mutation.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted High Variance	36	60
Poly-Weighted High Variance	6	90

Table 5.4: A comparison of the rates of speciation using the mono-weighted model versus the poly-weighted model. Here both models use high-variance dominance mutation.

5.3 Test of Hypothesis 1

To test hypothesis H_1, we must examine how skewed complete dominance is in its exploration of the the search space. To do this, the complete dominance model's ability to locate the larger beak size niche versus its ability to find the smaller beak size niche. Table 5.5 shows the results of this comparison, there exists a detectable difference in the complete dominance model's ability to locate the two niches. With a p-value < 0.001, there exists enough evidence to reject the null hypothesis, and as such it can be reasonably concluded that complete dominance does lead to a skew in exploration. As such, hypothesis H_1 is supported.

Distribution	**Times Found Niche**	**Times Did Not**
Trimodal Large	0	96
Trimodal Small	14	82

Table 5.5: A comparison of the rates of finding the additional niche for complete dominance in the trimodal large case versus the trimodal small case. Statistical test returns p-value of $7.328e-05$.

5.4 Test of Hypothesis 2

To test hypothesis H_2, we must see if variable dominance is skewed in its exploration of the search space. Similarly to the last section, the variable dominance models' ability to locate the larger beak size niche versus their ability to find the smaller beak size niche will be investigated for evidence of any skew. Table 5.6 and Table 5.7 both show how there is no discernible difference in exploration for any of the mono-weighted models. As such, it is fair to conclude that H_2 is supported by these results.

Distribution	**Times Found Niche**	**Times Did Not**
Trimodal Large	9	87
Trimodal Small	8	88

Table 5.6: A comparison of the rates of finding the additional niche for the mono-weighted model with high variance in the trimodal large case versus the trimodal small case. Statistical test returns p-value of 1.

Distribution	**Times Found Niche**	**Times Did Not**
Trimodal Large	1	95
Trimodal Small	2	94

Table 5.7: A comparison of the rates of finding the additional niche for the mono-weighted model with low variance in the trimodal large case versus the trimodal small case. Statistical test returns p-value of 1.

5.5 Tests of Hypothesis 3

To test hypothesis H_3, we must compare how often speciation occurs for the mono-weighted models versus the complete dominance model. Note that because of the preliminary results, the poly-weighted models are not compared to the complete dominance model here.

5.5.1 Hypothesis 3: Bimodal Distribution

For the bimodal distribution, the mono-weighted distribution with low-variance mutation is found to be statistically similar to complete dominance, as seen in

Table 5.8. However, the p-value obtained in Table 5.9 shows a statistically significant difference between the two models.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted Low Variance	49	47
Complete Dominance	56	40

Table 5.8: A comparison of the rates of speciation between the mono-weighted model using low-variance dominance and complete dominance in the bimodal seed distribution. Statistical test returns p-value of 0.3844.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted High Variance	36	60
Complete Dominance	56	40

Table 5.9: A comparison of the rates of speciation between the mono-weighted model using high-variance dominance and complete dominance in the bimodal seed distribution. Statistical test returns p-value of 0.00591.

5.5.2 Hypothesis 3: Trimodal Distribution

In the trimodal distributions, there is greater evidence for a difference in speciation rates. Table 5.10 and Table 5.11 both show that, regardless of the mutation method, the mono-weighted model is statistically outperformed by complete dominance in exhibiting any speciation in the trimodal large case. As Table 5.12 and Table 5.13 both show, the mono-weighted model is also statistically outperformed in the trimodal small case.

5.5.3 Hypothesis 3: Summary

Considering how the mono-weighted model only performed comparably to complete dominance in one comparison, it is reasonable to conclude that variable dominance does negatively impact the speciation rate. As such, there is evidence to reject hypothesis H_3.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted Low Variance	41	55
Complete Dominance	57	39

Table 5.10: A comparison of the rates of speciation between the mono-weighted model using low-variance dominance and complete dominance in the trimodal large environment. Statistical test returns p-value of 0.03008.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted High Variance	35	61
Complete Dominance	57	39

Table 5.11: A comparison of the rates of speciation between the mono-weighted model using high-variance dominance and complete dominance in the trimodal large environment. Statistical test returns p-value of 0.0023.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted Low Variance	31	65
Complete Dominance	55	41

Table 5.12: A comparison of the rates of speciation between the mono-weighted model using low-variance dominance and complete dominance in the trimodal small environment. Statistical test returns p-value of 0.0008.

Model	**Times Speciated**	**Times Did Not**
Mono-Weighted High Variance	33	63
Complete Dominance	55	41

Table 5.13: A comparison of the rates of speciation between the mono-weighted model using high-variance dominance and complete dominance in the trimodal small environment. Statistical test returns p-value of 0.0023.

Chapter 6

Conclusions

This chapter addresses the hypotheses laid out earlier in the thesis, discusses the contributions of this work in the field of computer science, and proposes some promising avenues for future work.

6.1 Hypotheses

This section will examine each of the hypotheses of this thesis in turn, and determine whether or not each hypothesis is supported by the evidence collected.

H_1: Under the model with large beaks being dominant, exploration will be asymmetrically skewed towards niches that require small beaks.

The results of the experiments with the trimodal seed distributions support this hypothesis. There was a significant difference in the complete dominance model's ability to find the smaller seed size compared to its ability to find the larger one. This confirms a skew towards exploration in one direction within the search space.

H_2: Under the variable gene dominance model, exploration will be symmetrical.

This hypothesis is also supported by the results of the experiment. Under both models using high variance mutation, exploration in one direction was not significantly different than exploration in the other. This indicates a lack of a skew in the exploration of a problem space when using variable dominance.

H_3: Variable gene dominance will allow for speciation at a comparable rate to a model in which size determines dominance.

This hypothesis, unlike the others, is not supported by the results. While speciation did still occur at an appreciable rate, it was markedly less often than under the complete dominance model. When making this hypothesis, several competing factors were taken into consideration; namely, it was assumed that the increase in exploratory ability and re-population potential would offset the added randomness of variable dominance. However, it appears that these factors were not balanced. Most likely, it is the higher rates of extinction that hurt the model. As can be seen in Figure 6.1, mono-weighted chromosomes with high-variance mutation featured more genetic diversity in their beak sizes. Unfortunately, this may have led to them dying more often than birds whose genetics used complete dominance, as more individuals would have beaks not sized to exploit existing resources in the environment. If the higher rates of extinction are to blame, then fixing the speciation decrease would be as simple as using an environment in which organisms cannot die off. Such a fix would be easy to implement in future work.

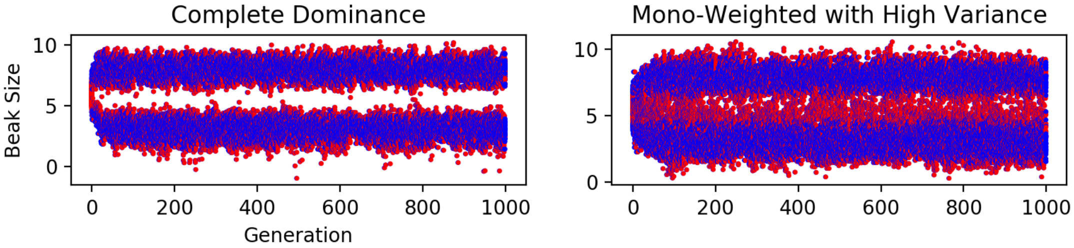

Figure 6.1: Population over time for complete dominance verses mono-weighted dominance with high-variance mutation. Note that blue dots represent birds that survived while red dots represent birds that died.

6.2 Contributions

This thesis introduces a new concept for genetic algorithm research in the form of dominance scalars. While the concept of dominance is not new to genetic algorithms, previous research tends to treat it as a categorical value instead of a numerical value (Smith and Goldberg, 1992). This thesis proposes a new line of thought when it comes to dominance by treating dominance as spectrum.

This thesis also introduced the concept of variable dominance to utilize and evolve dominance scalars. This allows for the implementation of a dominance schema without prior knowledge of a given search space, thus eliminating the bias introduced in complete dominance.

Multiple approaches to variable dominance were devised, and each approach was investigated to determine its relative effectiveness. Due to the demonstrated effectiveness of mono-weighted dominance with high-variance mutation, future endeavors can use that model as a starting block for implementing variable dominance.

While not successful in all its goals, this project does serve as a proof of concept for variable dominance in future work. If the negative impact on the speciation rate could be addressed, more genetic algorithm applications could adopt a variably dominant, diploid based genetic structure as a means to encourage exploration while only minimally affecting exploitation of known solutions. This could greatly increase the effectiveness and applicability of genetic algorithms as a whole without a significant performance trade-off. By eliminating the need for a modified selection process and instead encouraging speciation through changes in the genetic structure alone, variable dominance turns what was an $O(N^2)$ process into an $O(N)$ one. This is because mating can be accomplished without every solution needing to be compared to its peers.

In addition to performance increases, variable dominance will allow for more open exploration than complete dominance. This is because it explores in all directions in a problem space, thus allowing for sufficient exploration without prior knowledge of the search space to begin with.

6.3 Future Work

Considering the potential of variable dominance, several ideas for future work come to mind.

Developing a more multidimensional problem space to test how well variable dominance scales may go a long way to understanding how effective it could be in other environments. Considering that variable dominance outperformed complete dominance in finding distant niches despite its lower rate of speciation, variable dominance may shine in a multimodal environment with a large number of peaks.

Another endeavor worth taking would be to test how well variable dominance performs when compared to older diploid models. While many of these older models employ binary dominance, variable dominance employs a scaling dominance that could more easily adapt to subtleties in a search space that cause difficulty for the more rigid notions of dominance (Smith and Goldberg, 1992).

While the biological simulation proves to be an excellent environment for speciation research, applying variable dominance to a industry related problem could accomplish two things. First, it would demonstrate that variable dominance is not constrained to biological simulations. Second, it may show that variable dominance performs even better in an environment in which extinction is not supported. Special care must be taken when designing the genetic algorithm so that dominance has a meaningful impact on speciation.

Building on the idea presented in this thesis that recessive traits are more free to explore than dominant ones, it may be worth exploring mutation methods that increase mutability in chromosomes inversely proportional to their dominance scalars. This could allow for a greater degree of exploration while not sacrificing much in terms of exploitation. As long as there existed a stable population of dominant chromosomes, the genetic algorithm could evolve incremental changes improving a found peak while still encouraging enough diversity to find other niches in the search space.

Finally, it is believed that changing the representation of dominance scalars may lead to better results. While different forms of mutation were tested on the dominance scalars in this thesis, the basic structure of weights remained the same. Possible alternatives to integers ranging from 1 to 100 include dominance functions, dominance maps, or simply integers drawn from a different distribution.

Bibliography

Thomas Bäck. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms.* Oxford University Press, Inc., USA, 1996. ISBN 0195099710.

Thomas Bäck, David B. Fogel, and Zbigniew Michalewicz. *Handbook of Evolutionary Computation.* IOP Publishing Ltd., GBR, 1st edition, 1997. ISBN 0750303921.

Stewart H. Berlocher. *Endless forms species and speciation.* Oxford University Press, New York, 1998.

William Booker and Dean Frederick Hougen. Meiotic inheritance and gene dominance in synthetic sympatric speciation. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, July 2018. doi: 10.1109/CEC.2018.8477761.

Edoardo Conti, Vashisht Madhavan, Felipe Petroski Such, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Improving exploration in evolution strategies for deep reinforcement learning via a population of novelty-seeking agents. *Computing Research Repository*, abs/1712.06560, 2017. URL `http://arxiv.org/abs/1712.06560`.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

Antonio Della Cioppa, Angelo Marcelli, and Prisco Napoli. Speciation in evolutionary algorithms: Adaptive species discovery. In *Genetic and Evolutionary Computation Conference, GECCO'11*, pages 1053–1060, 01 2011. doi: 10.1145/2001576.2001719.

Stephanie Forrest and Melanie Mitchell. What makes a problem hard for a genetic algorithm? Some anomalous results and their explanation. *Machine Learning*, 13, 10 1994. doi: 10.1007/BF00993046.

Paul Grouchy, Jekanthan Thangavelautham, and Gabriele M.T. D'Eleuterio. An island model for high-dimensional genomes using phylogenetic speciation and species barcoding. In *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, GECCO '09, pages 1355–1362, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605583259. doi: 10.1145/1569901.1570083. URL `https://doi.org/10.1145/1569901.1570083`.

John H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, 1975.

Joel Lehman and Kenneth O. Stanley. Efficiently evolving programs through the search for novelty. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, GECCO '10, pages 837–844, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300728. doi: 10.1145/1830483.1830638. URL `https://doi.org/10.1145/1830483.1830638`.

Ilya Loshchilov, Marc Schoenauer, and Michèle Sebag. Not all parents are equal for MO-CMA-ES. In Ricardo H. C. Takahashi, Kalyanmoy Deb, Elizabeth F. Wanner, and Salvatore Greco, editors, *Evolutionary Multi-Criterion Optimization*, pages 31–45, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19893-9.

John H. McDonald. *Handbook of Biological Statistics.* Sparky House Publishing, Baltimore, Maryland, 3rd edition, 2014.

Melanie Mitchell and Stephanie Forrest. Genetic algorithms and artificial life. *Artificial Life*, 1(3):267–289, 1994. doi: 10.1162/artl.1994.1.3.267. URL `https://doi.org/10.1162/artl.1994.1.3.267`.

Syed Naveed Hussain Shah and Dean Frederick Hougen. Rethinking stochasticity in neural networks for reinforcement learning with continuous actions. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 488–496, 2019.

John Maynard Smith. Sympatric speciation. *The American Naturalist*, 100 (916):637–650, 1966. doi: 10.1086/282457. URL `https://doi.org/10.1086/282457`.

Robert Elliot Smith and David E Goldberg. Diploidy and dominance in artificial genetic search. *Complex Systems*, 6(3):251–285, 1992.

Mark Woehrer, Dean Hougen, and Ingo Schlupp. Sexual selection, resource distribution, and population size in synthetic sympatric speciation. *Artificial Life Conference Proceedings*, (24):137–144, 2012. doi: 10.1162/

978-0-262-31050-5-ch020. URL `https://www.mitpressjournals.org/doi/abs/10.1162/978-0-262-31050-5-ch020`.

En Lin Yu and Ponnuthurai Nagaratnam Suganthan. Ensemble of niching algorithms. *Information Sciences*, 180:2815–2833, 2010.

Appendix A

Population Bands

The following are the results of each of the 96 trials for all combinations of genetic models and seed distributions. Blue dots represent birds that survived while red dots represent birds that died. Due to the number of trials only half of the trials in a run can fit on one page. As such, each run is split into two parts, labeled "Part 1" and "Part 2" respectively.

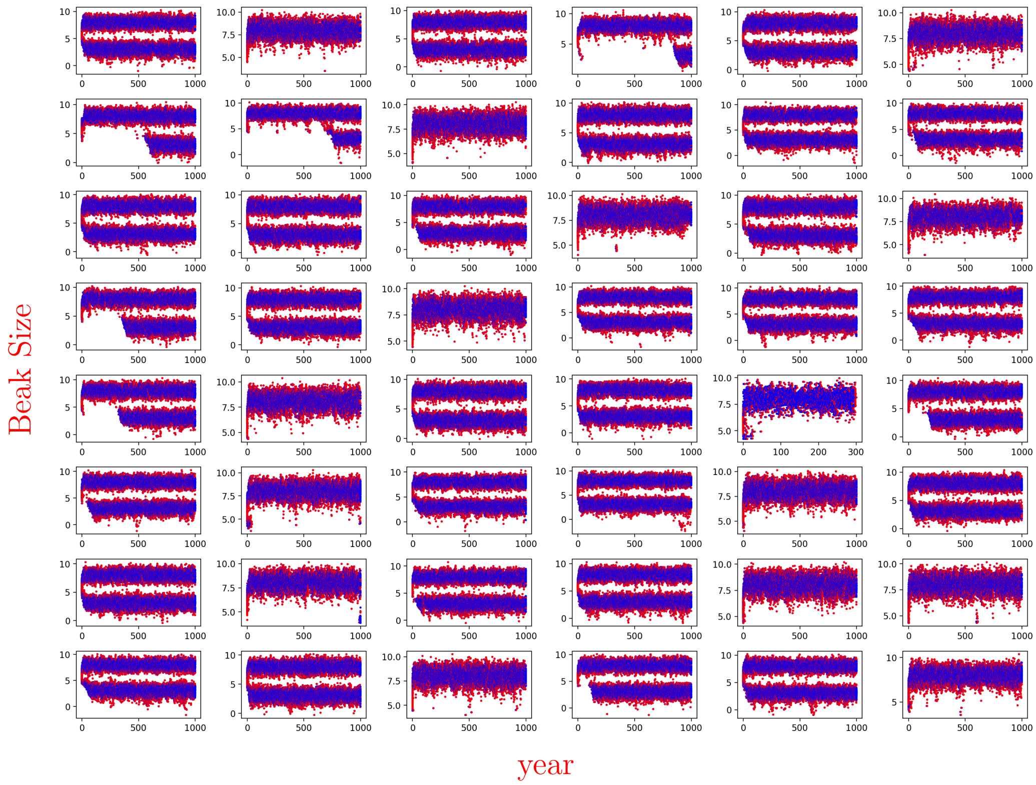

Figure A.1: Performance of the complete dominance model under the bimodal seed distribution. Part 1.

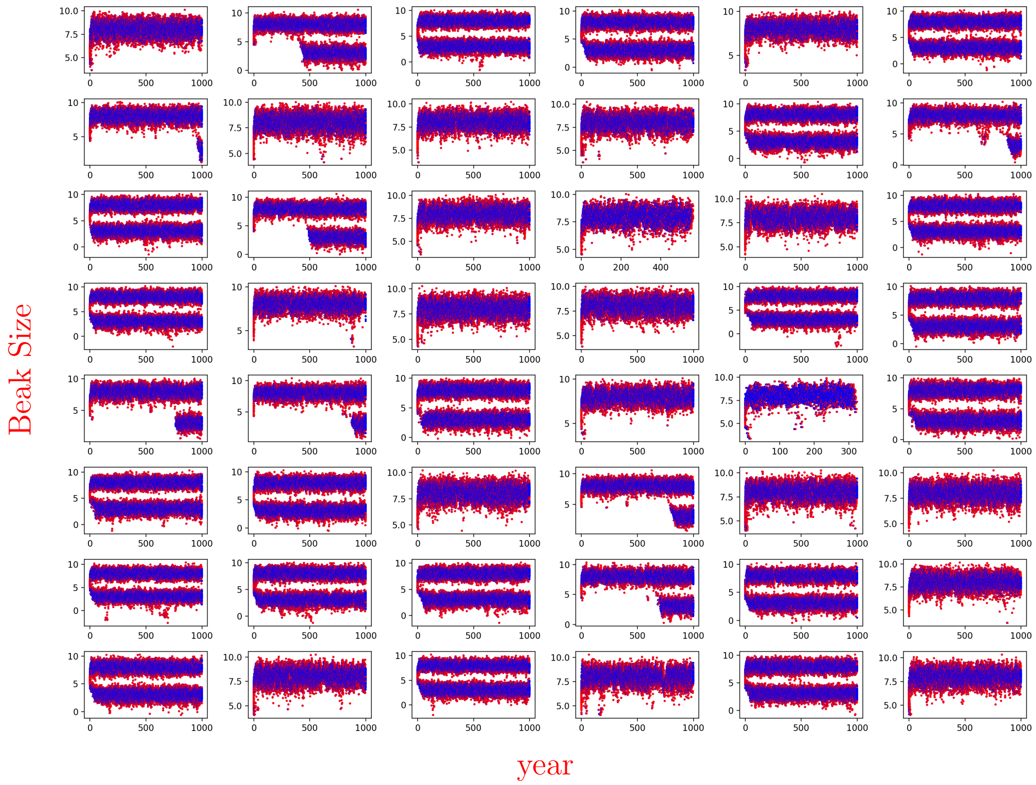

Figure A.2: Performance of the complete dominance model under the bimodal seed distribution. Part 2.

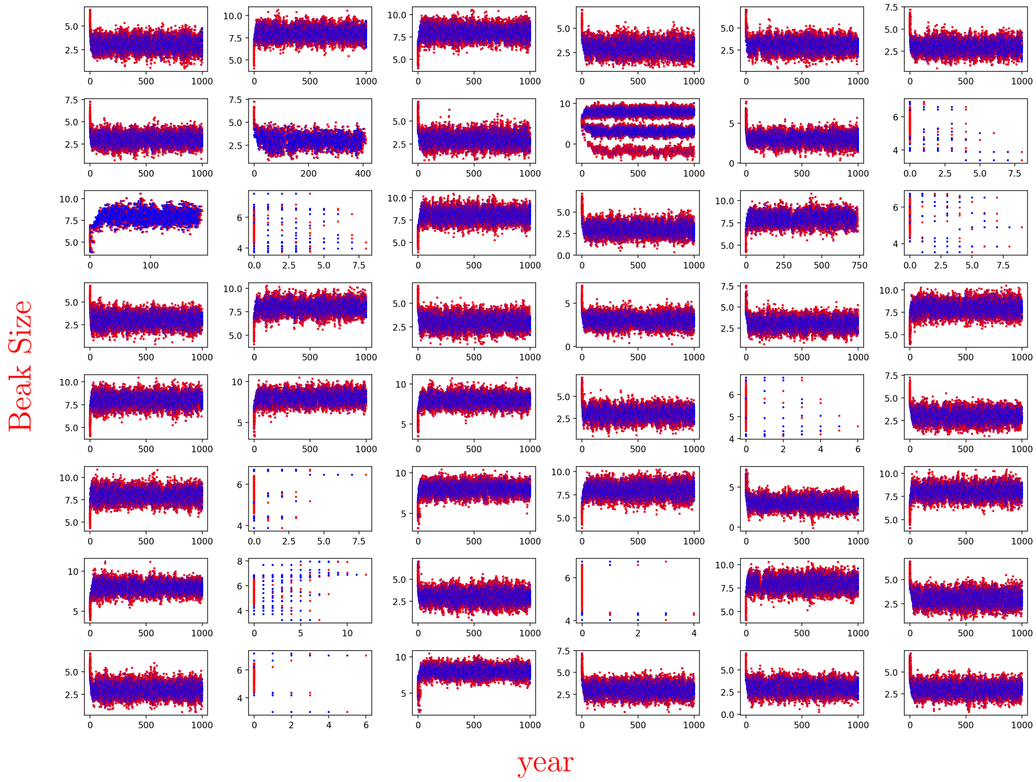

Figure A.3: Performance of the incomplete dominance model under the bimodal seed distribution. Part 1.

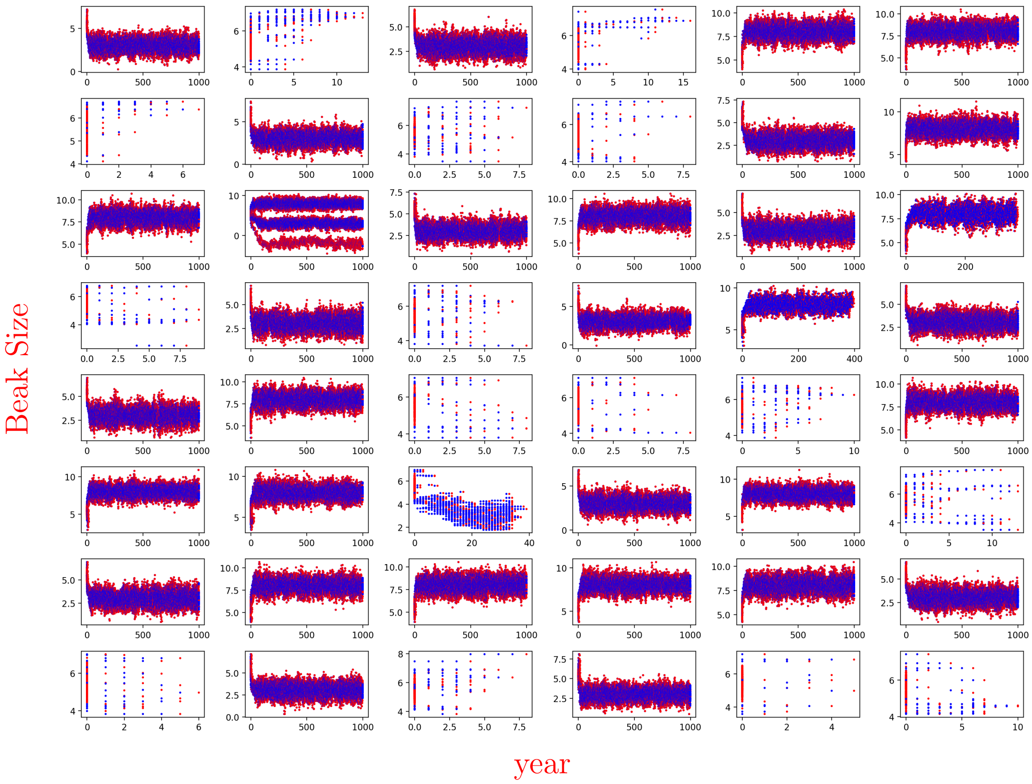

Figure A.4: Performance of the incomplete dominance model under the bimodal seed distribution. Part 2.

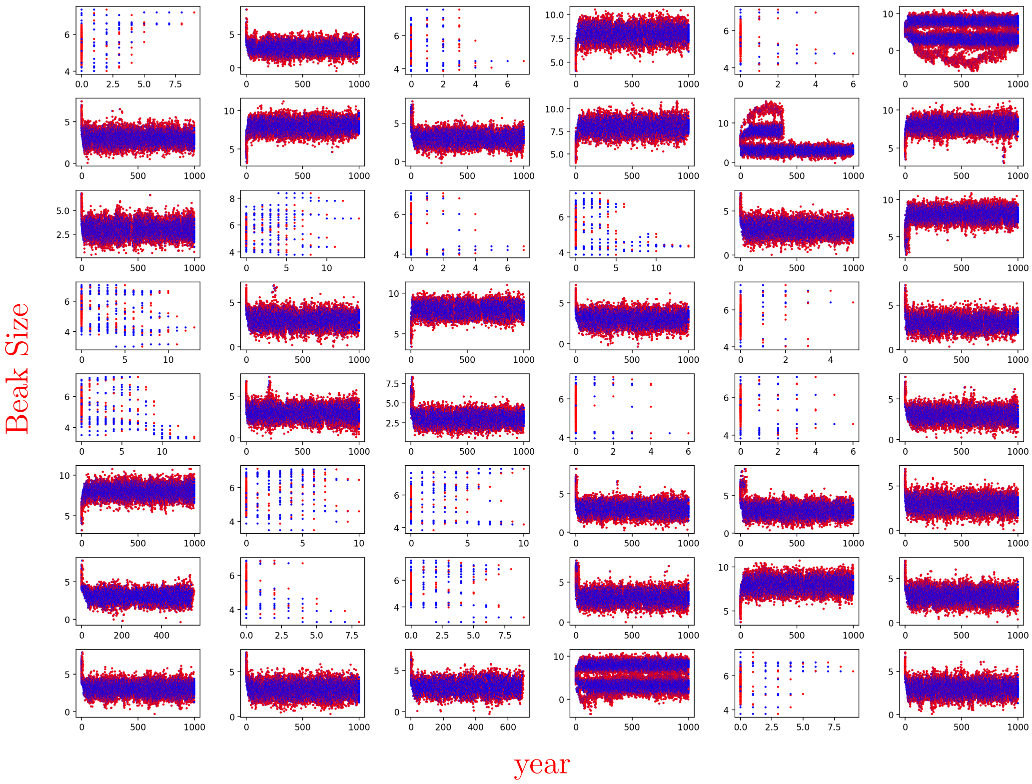

Figure A.5: Performance of the poly-weighted dominance model with high variance under the bimodal seed distribution. Part 1.

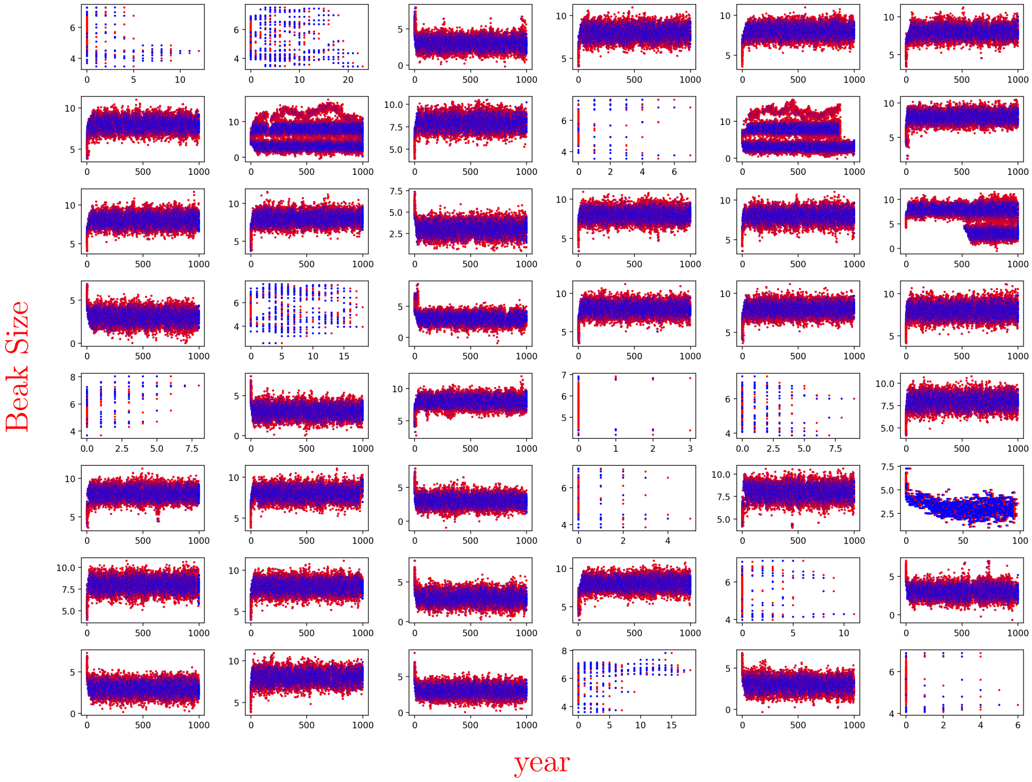

Figure A.6: Performance of the poly-weighted dominance model with high variance under the bimodal seed distribution. Part 2.

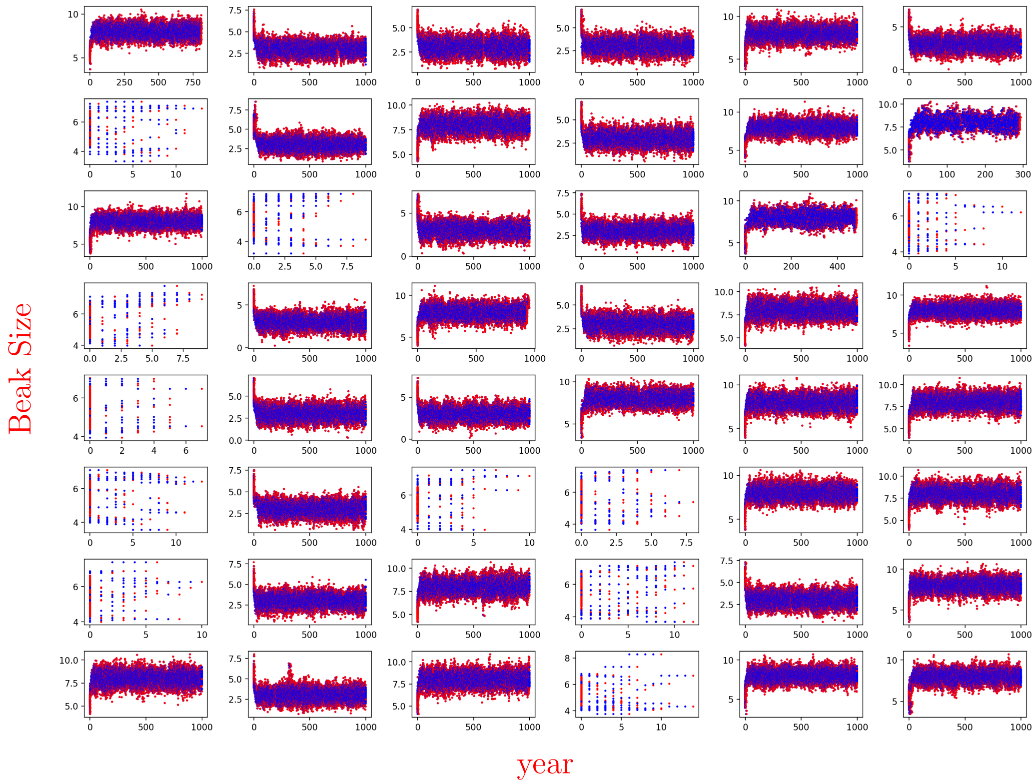

Figure A.7: Performance of the poly-weighted dominance model with low variance under the bimodal seed distribution. Part 1.

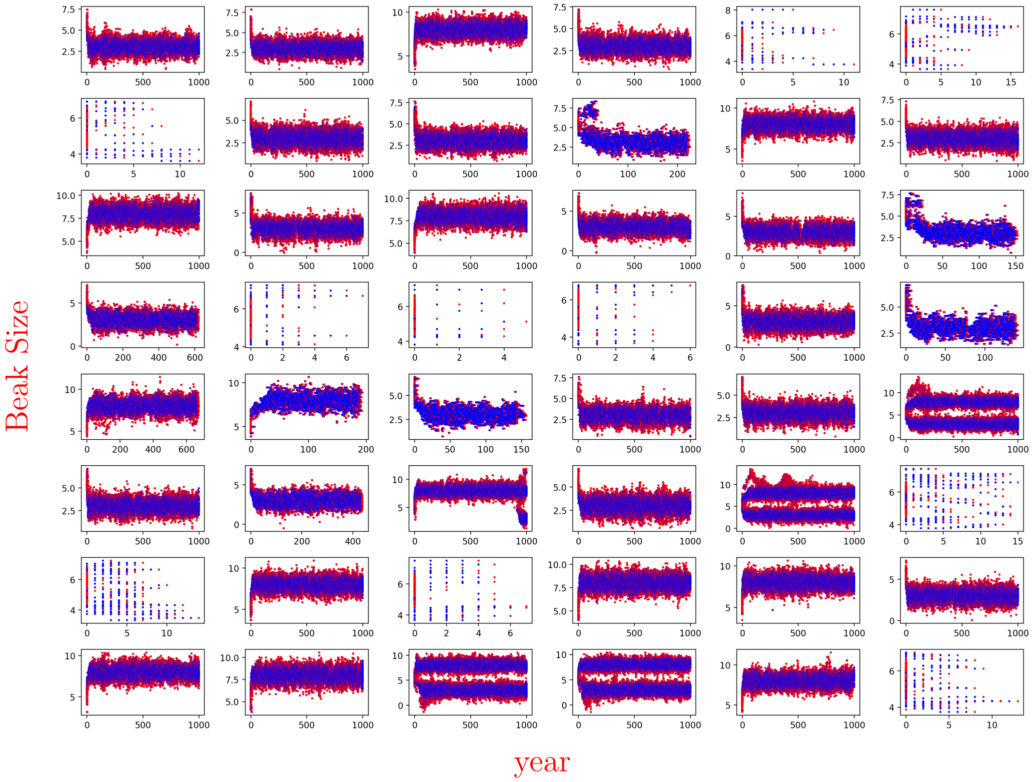

Figure A.8: Performance of the poly-weighted dominance model with low variance under the bimodal seed distribution. Part 2.

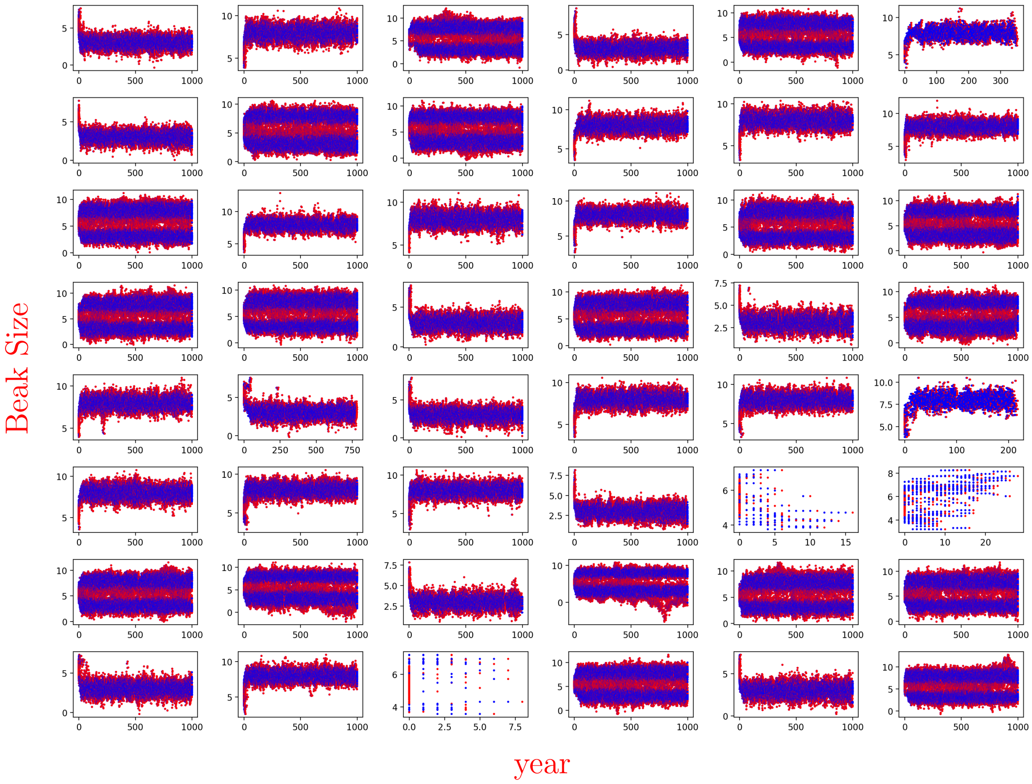

Figure A.9: Performance of the mono-weighted dominance model with high variance under the bimodal seed distribution. Part 1.

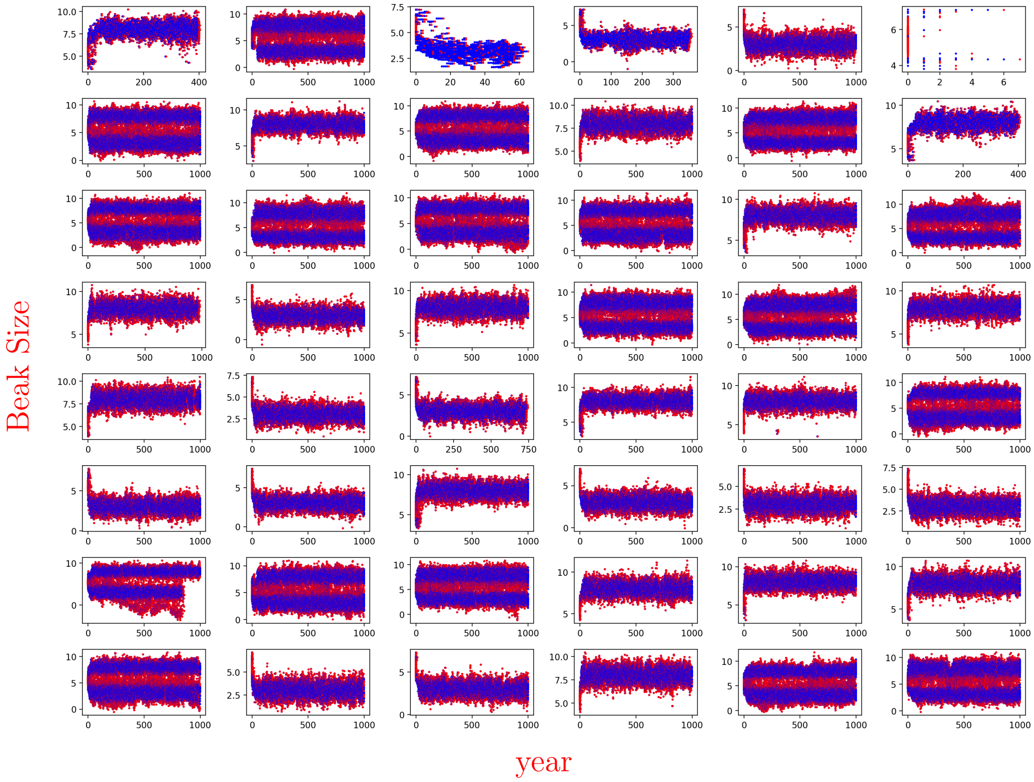

Figure A.10: Performance of the mono-weighted dominance model with high variance under the bimodal seed distribution. Part 2.

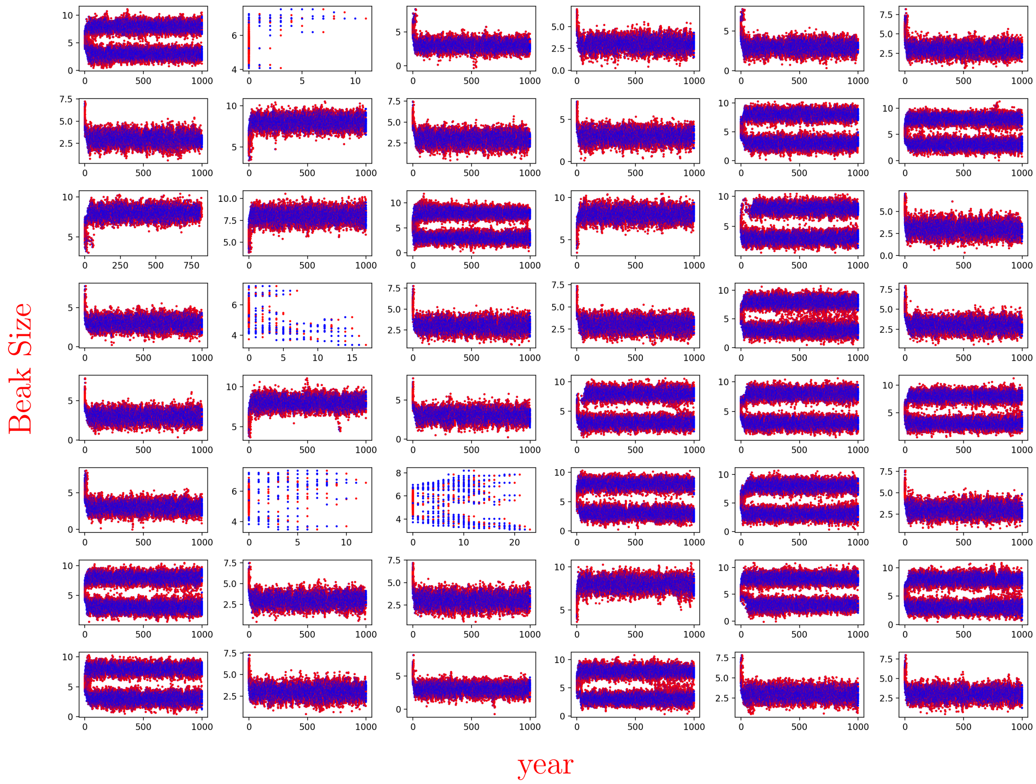

Figure A.11: Performance of the mono-weighted dominance model with low variance under the bimodal seed distribution. Part 1.

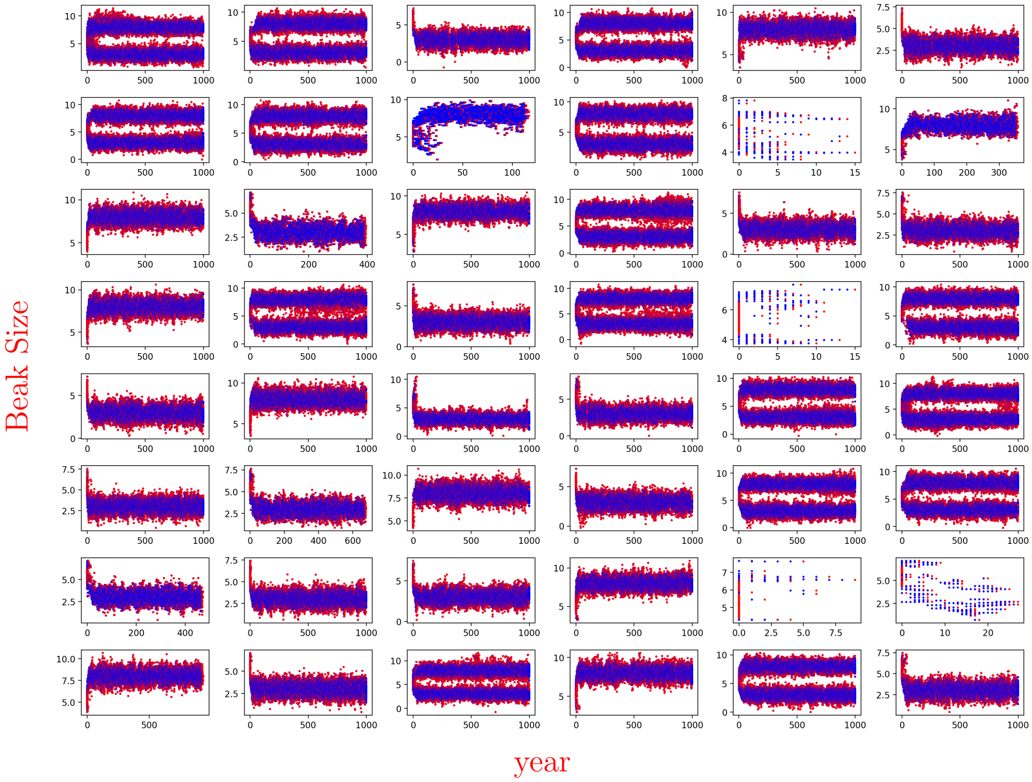

Figure A.12: Performance of the mono-weighted dominance model with low variance under the bimodal seed distribution. Part 2.

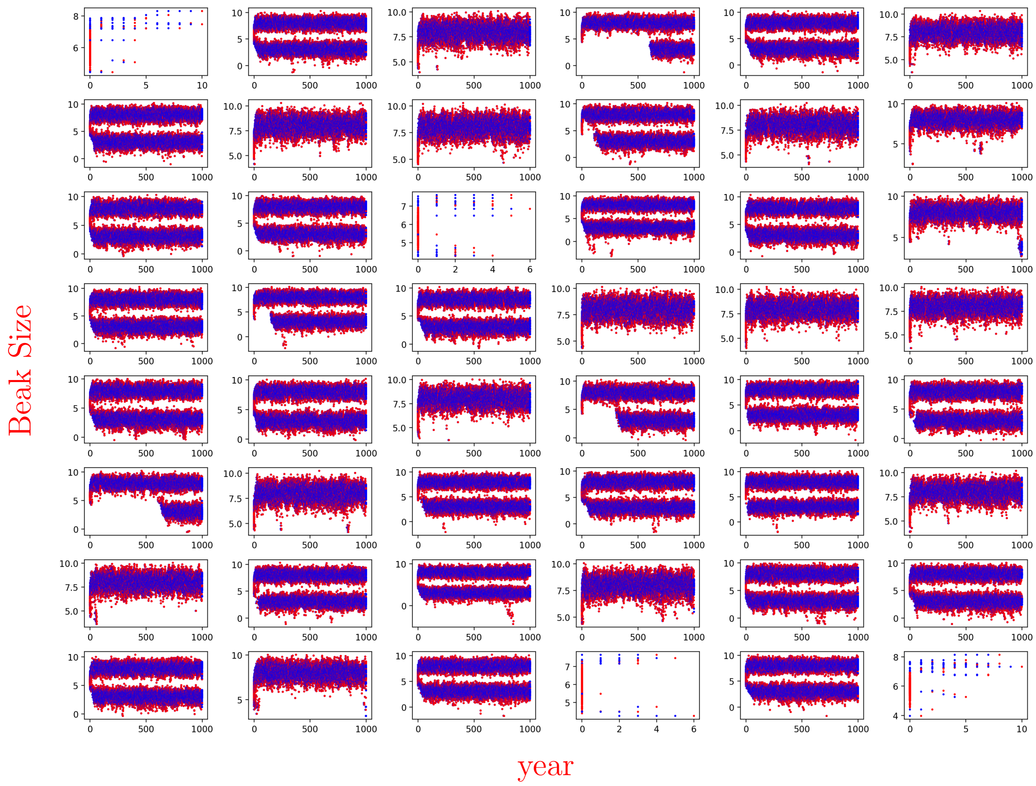

Figure A.13: Performance of the complete dominance model under the trimodal large seed distribution. Part 1.

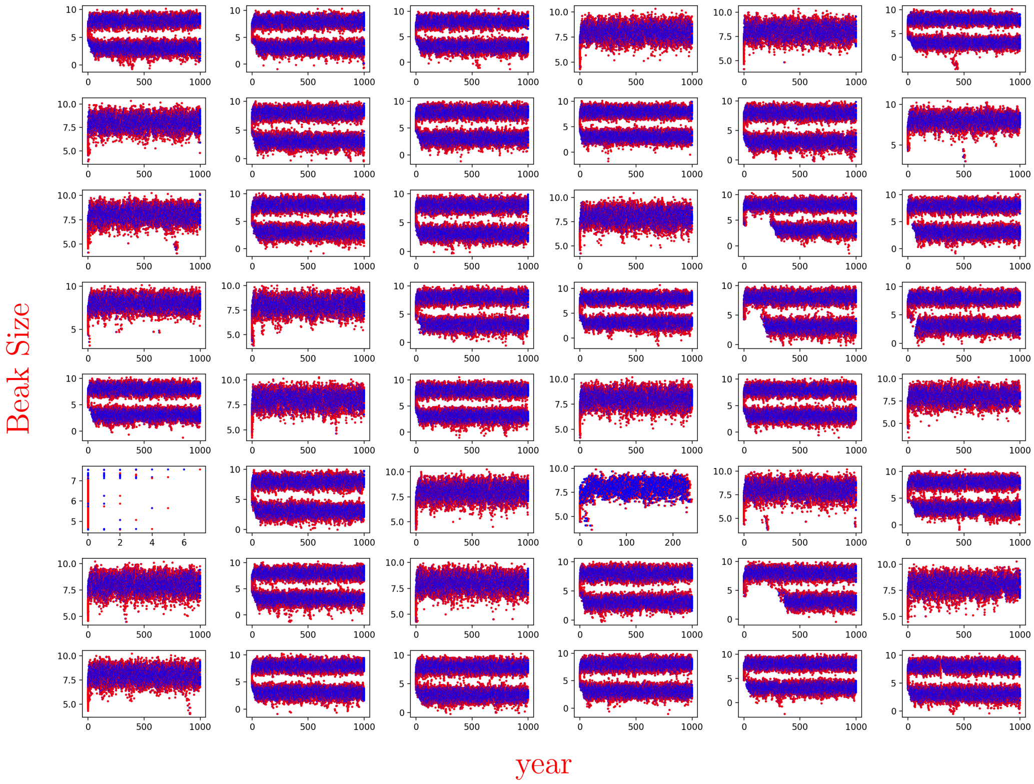

Figure A.14: Performance of the complete dominance model under the trimodal large seed distribution. Part 2.

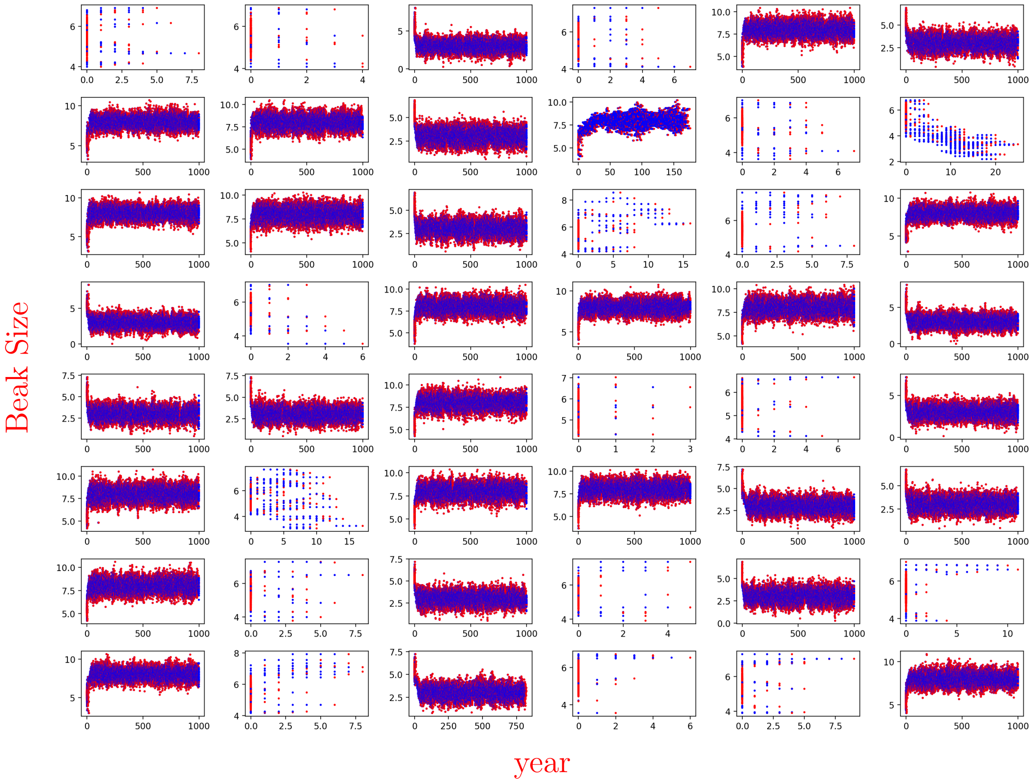

Figure A.15: Performance of the incomplete dominance model under the trimodal large seed distribution. Part 1.

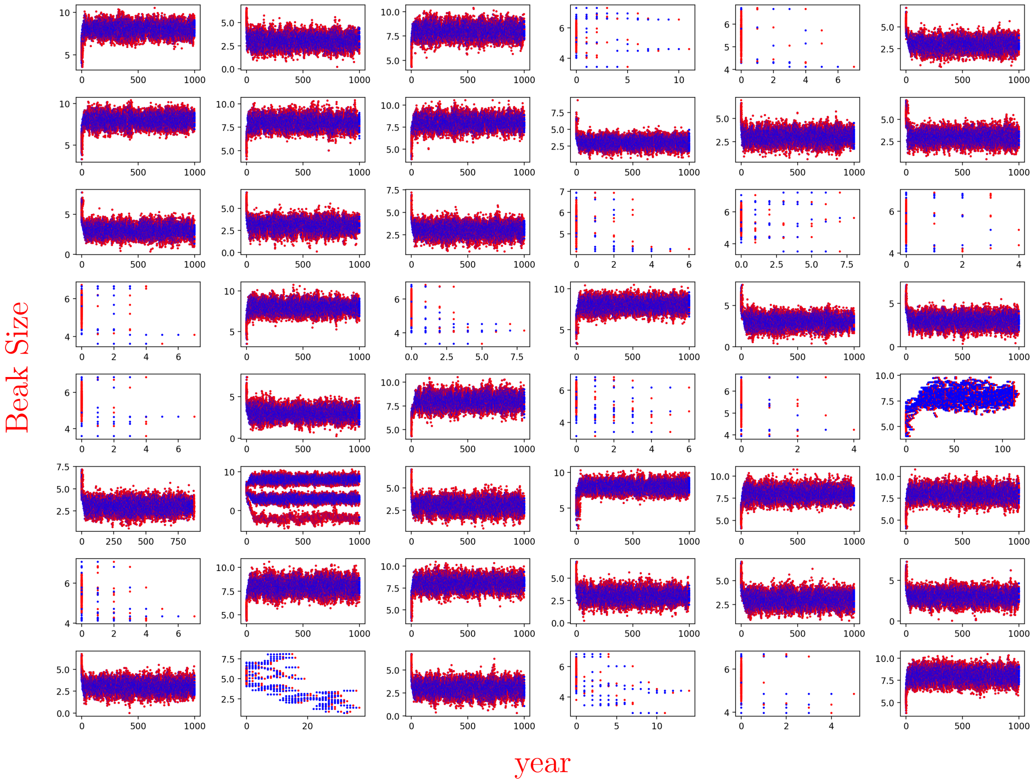

Figure A.16: Performance of the incomplete dominance model under the trimodal large seed distribution. Part 2.

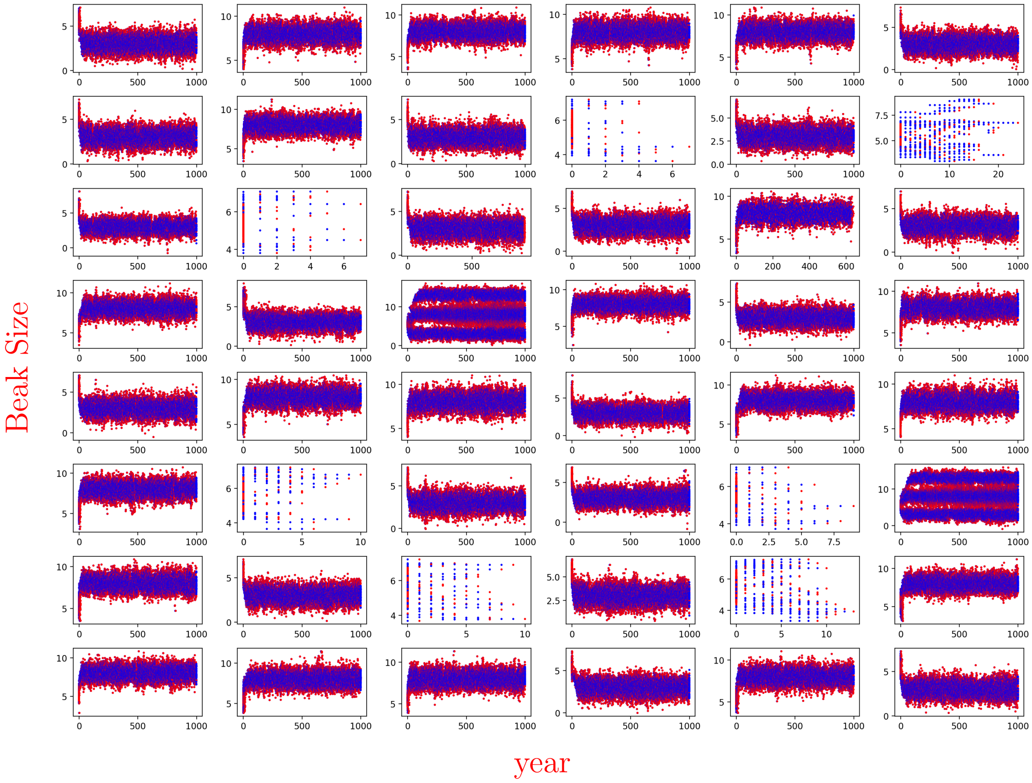

Figure A.17: Performance of the poly-weighted dominance model with high variance under the trimodal large seed distribution. Part 1.

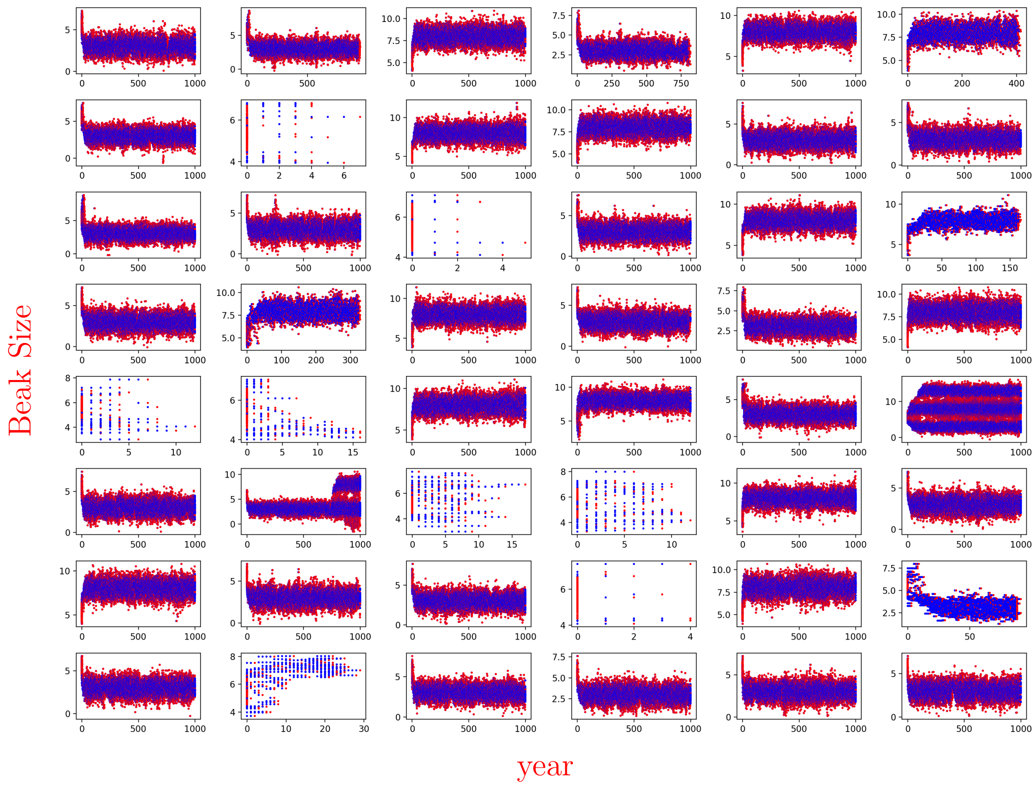

Figure A.18: Performance of the poly-weighted dominance model with high variance under the trimodal large seed distribution. Part 2.

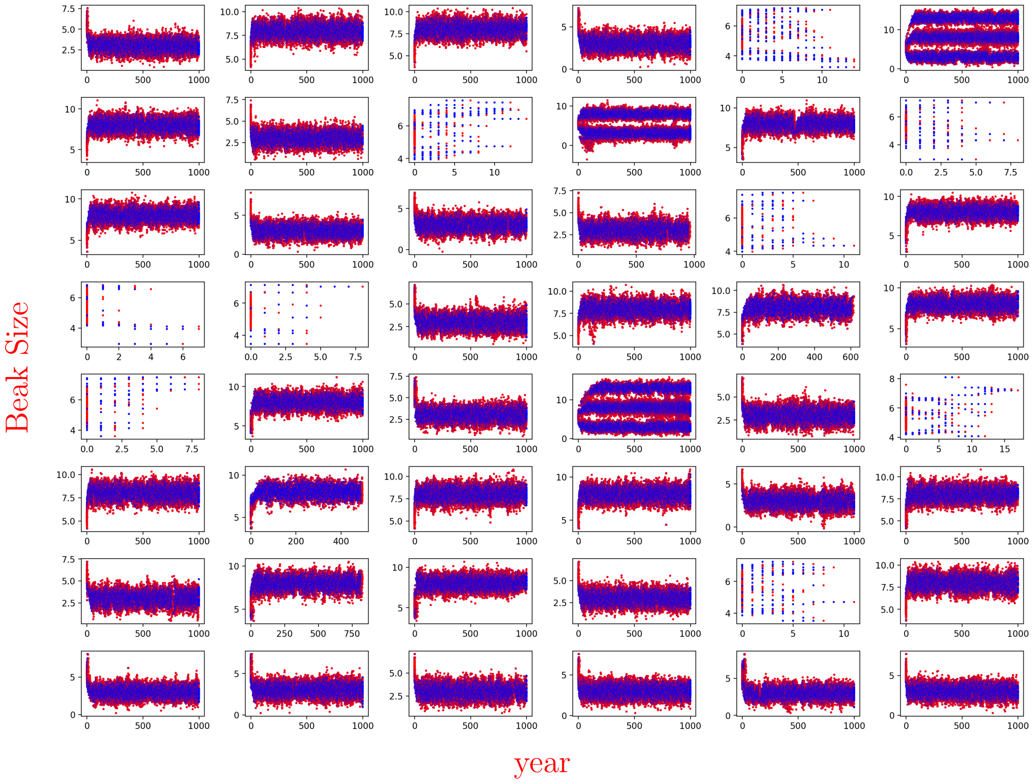

Figure A.19: Performance of the poly-weighted dominance model with low variance under the trimodal large seed distribution. Part 1.

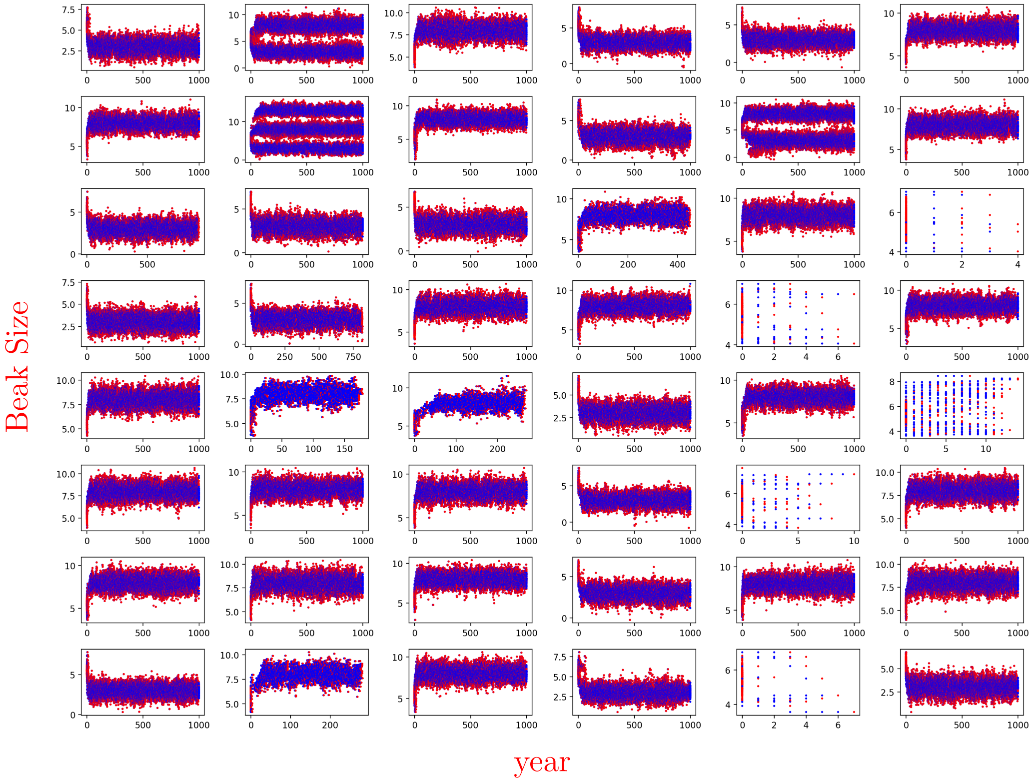

Figure A.20: Performance of the poly-weighted dominance model with low variance under the trimodal large seed distribution. Part 2.

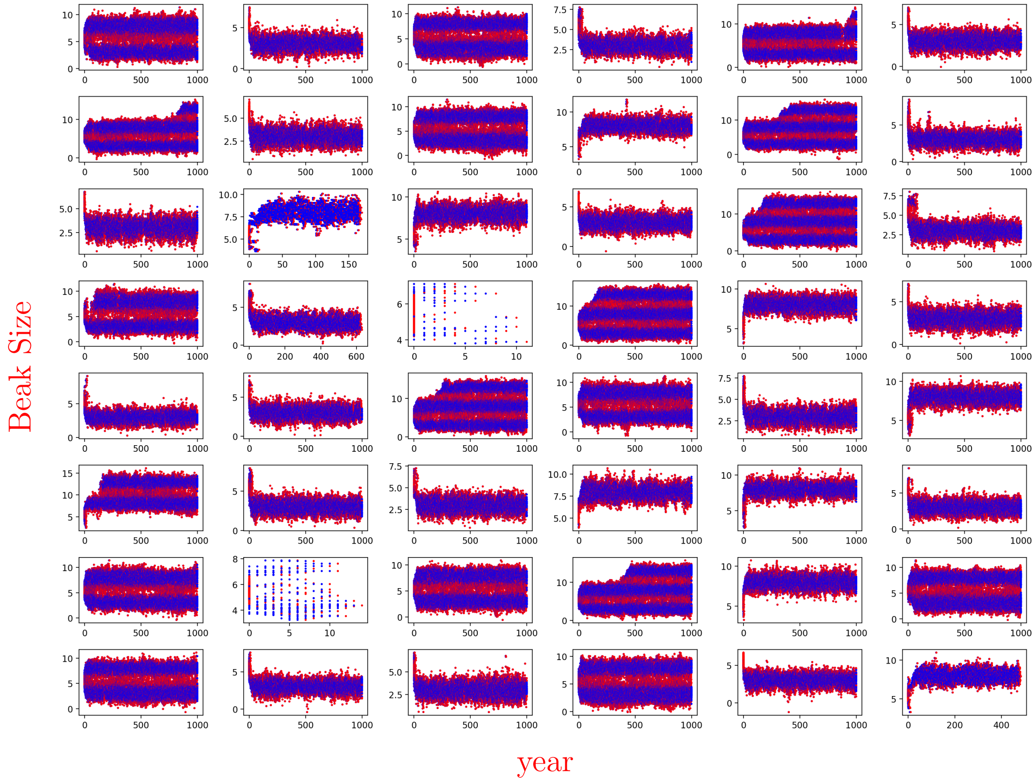

Figure A.21: Performance of the mono-weighted dominance model with high variance under the trimodal large seed distribution. Part 1.

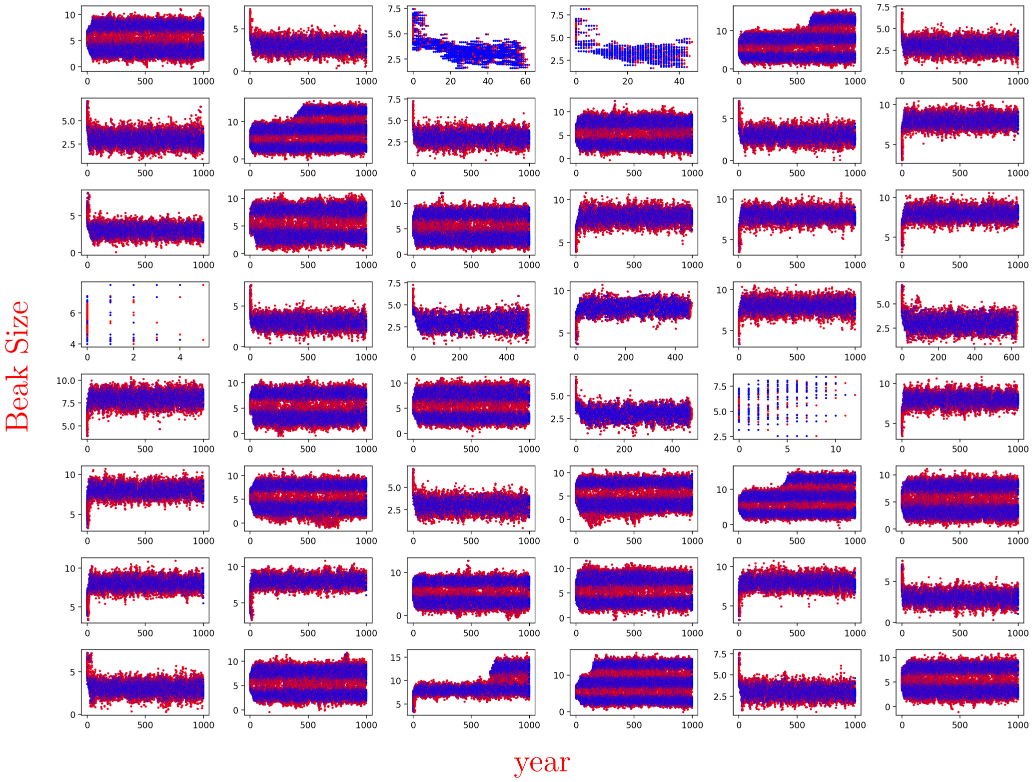

Figure A.22: Performance of the mono-weighted dominance model with high variance under the trimodal large seed distribution. Part 2.

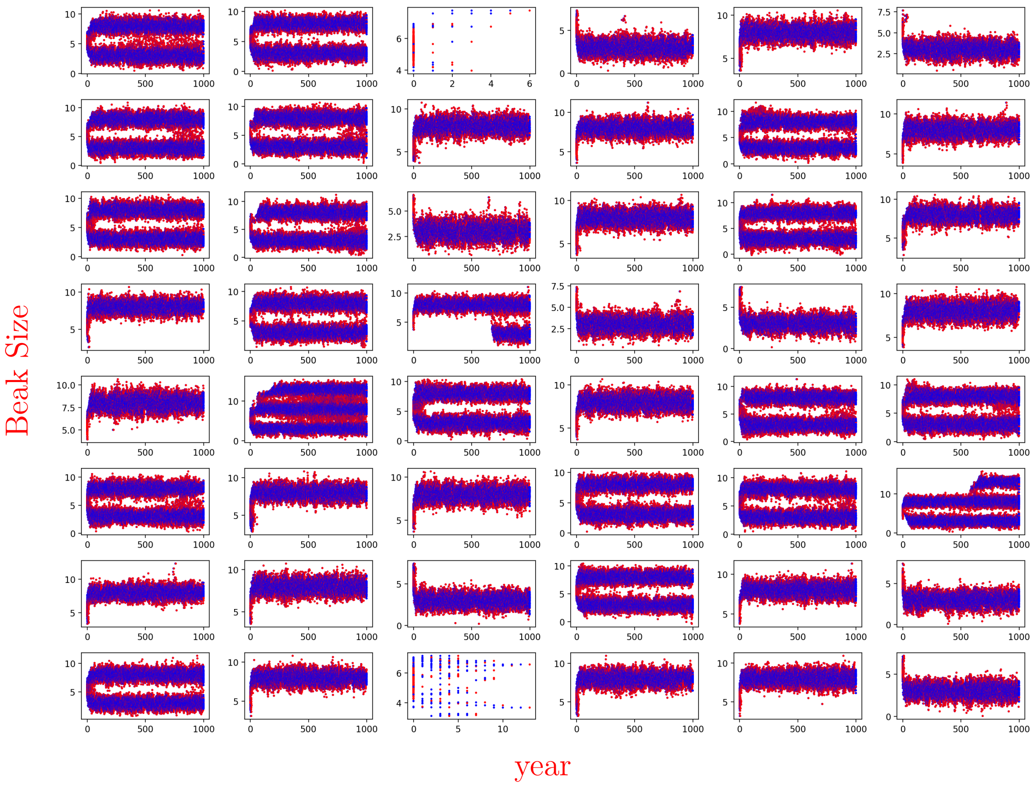

Figure A.23: Performance of the mono-weighted dominance model with low variance under the trimodal large seed distribution. Part 1.

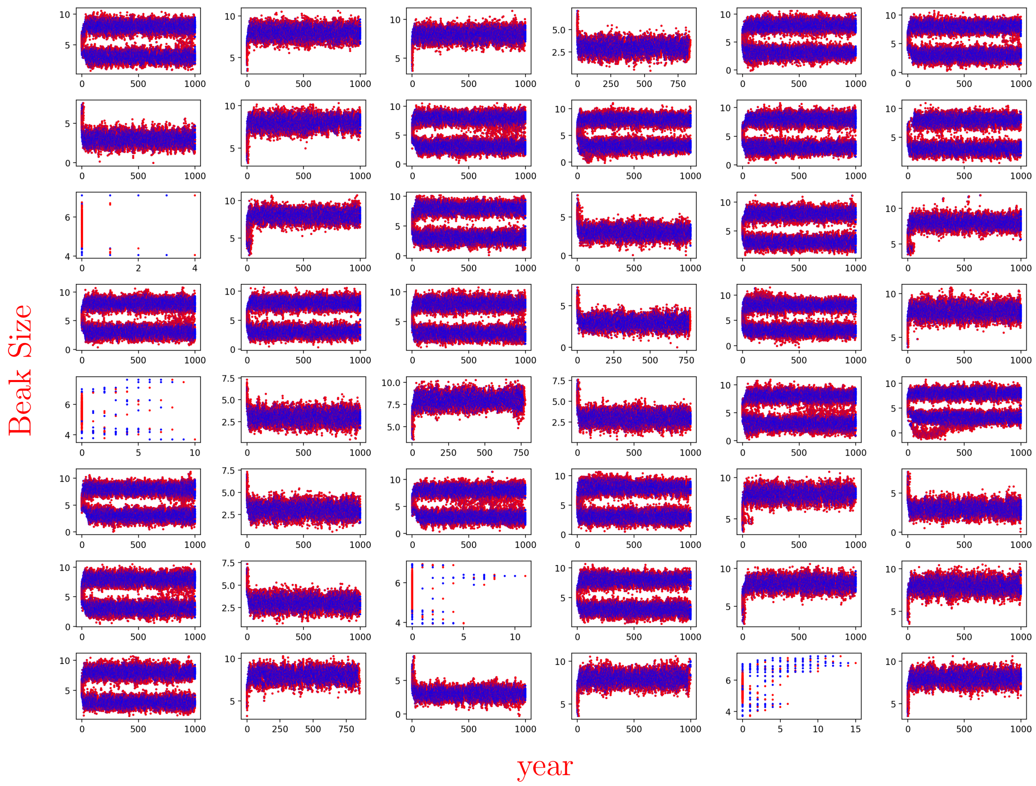

Figure A.24: Performance of the mono-weighted dominance model with low variance under the trimodal large seed distribution. Part 2.

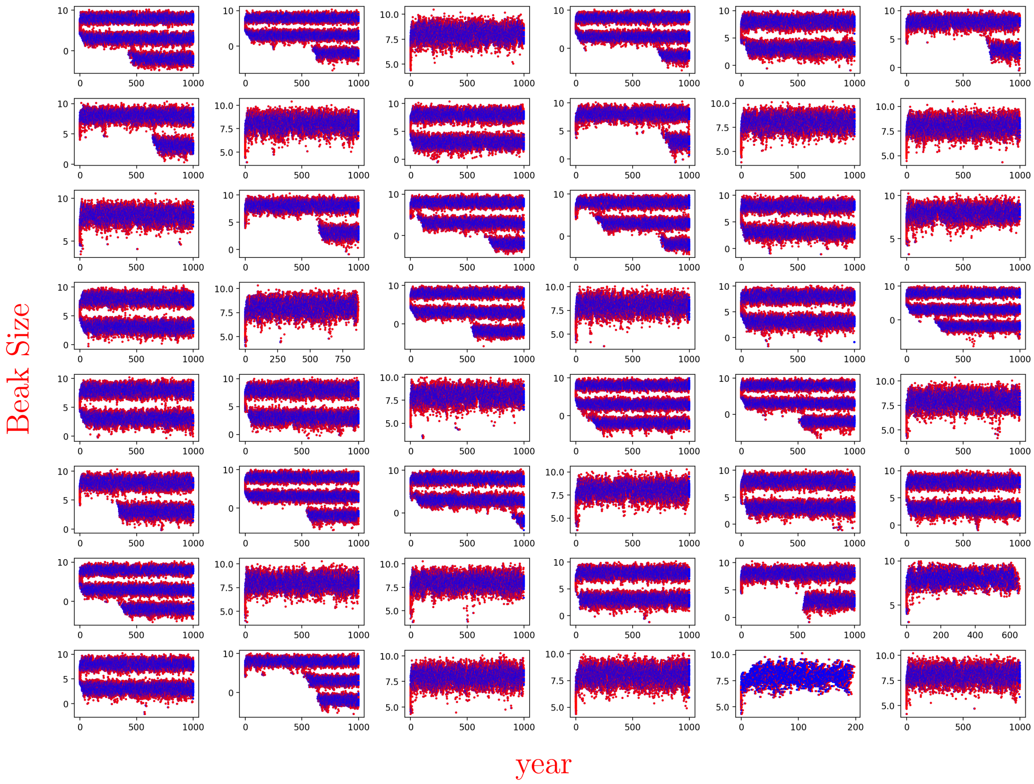

Figure A.25: Performance of the complete dominance model under the trimodal small seed distribution. Part 1.

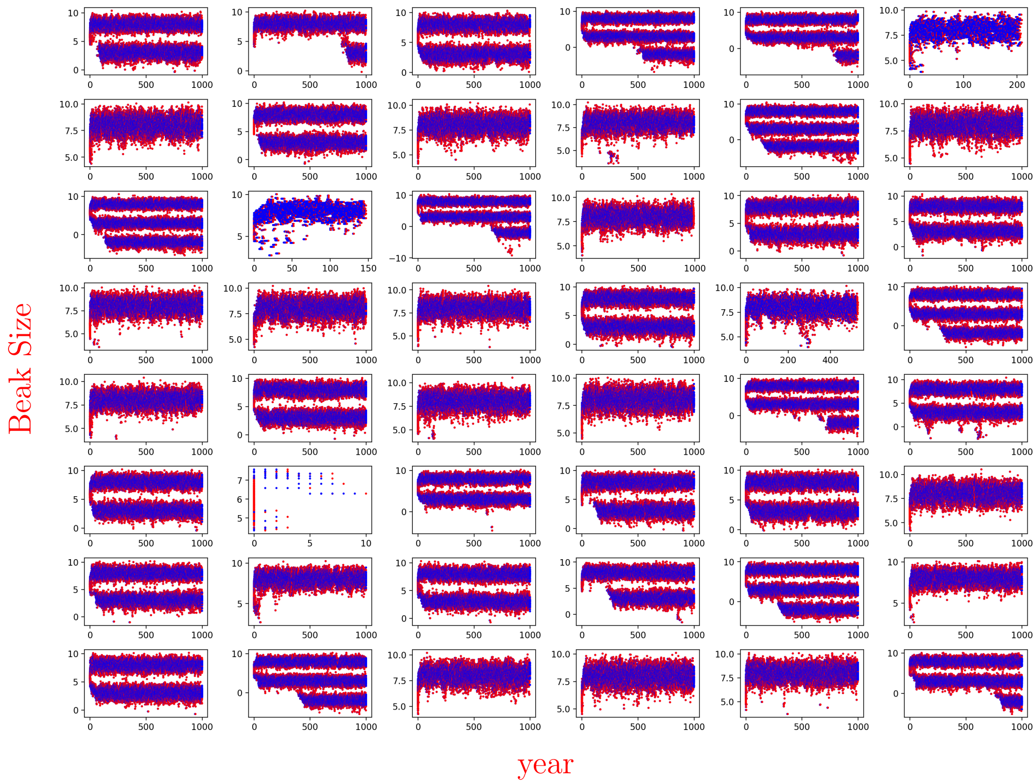

Figure A.26: Performance of the complete dominance model under the trimodal small seed distribution. Part 2.

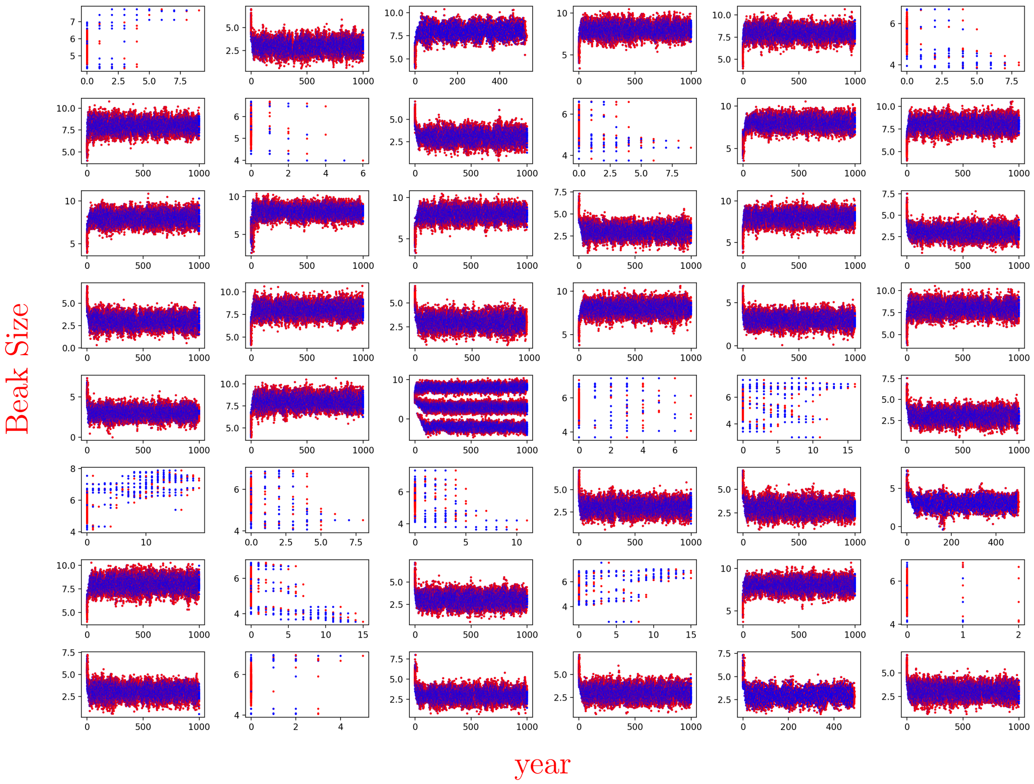

Figure A.27: Performance of the incomplete dominance model under the trimodal small seed distribution. Part 1.

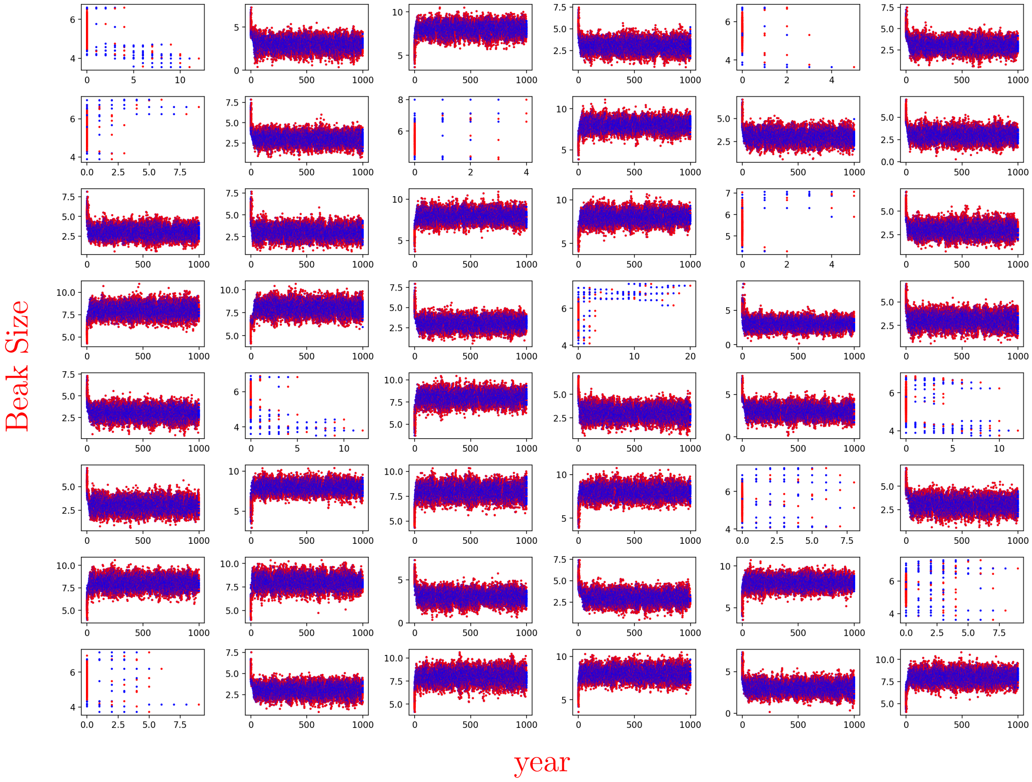

Figure A.28: Performance of the incomplete dominance model under the trimodal small seed distribution. Part 2.

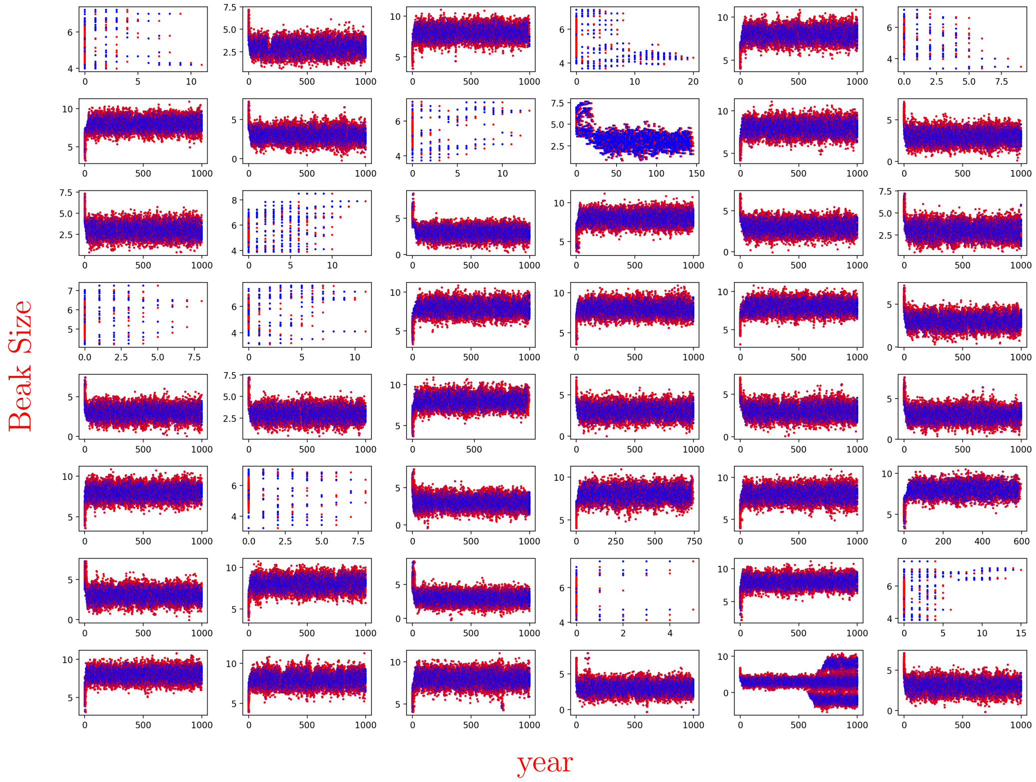

Figure A.29: Performance of the poly-weighted dominance model with high variance under the trimodal small seed distribution. Part 1.

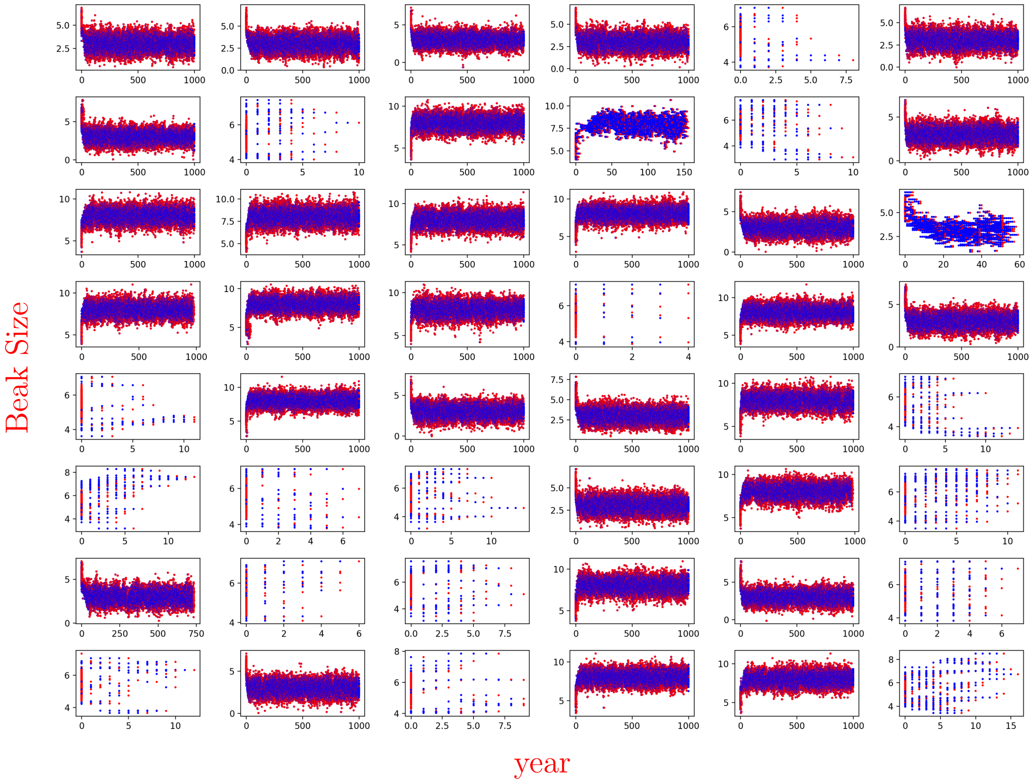

Figure A.30: Performance of the poly-weighted dominance model with high variance under the trimodal small seed distribution. Part 2.

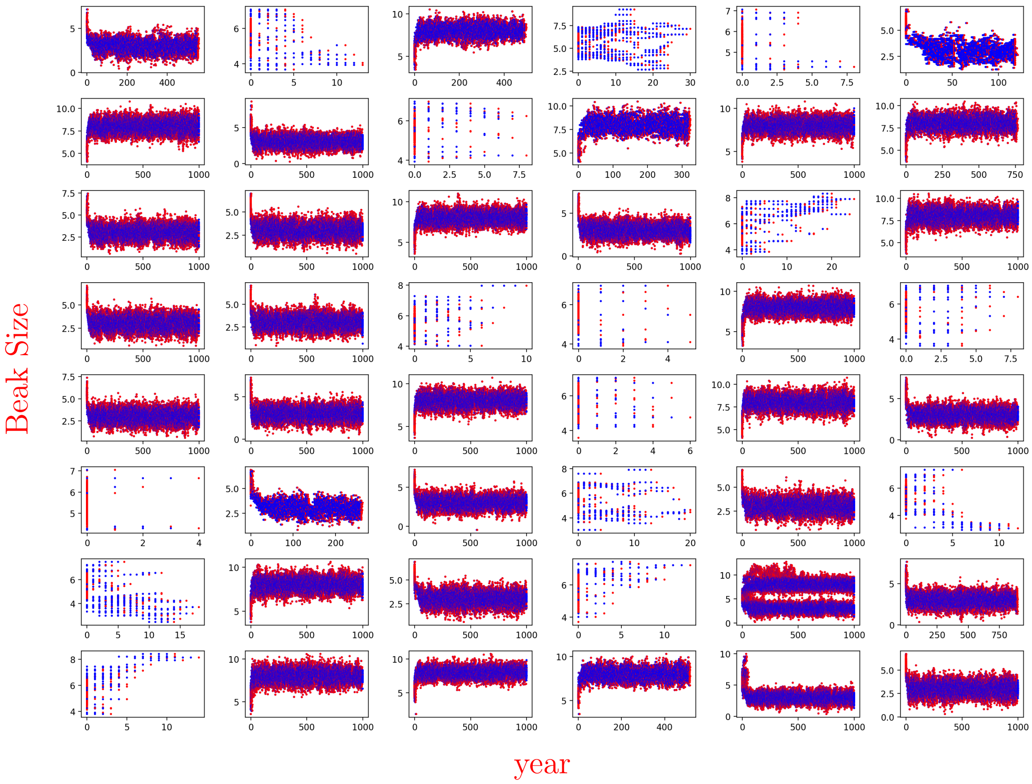

Figure A.31: Performance of the poly-weighted dominance model with low variance under the trimodal small seed distribution. Part 1.

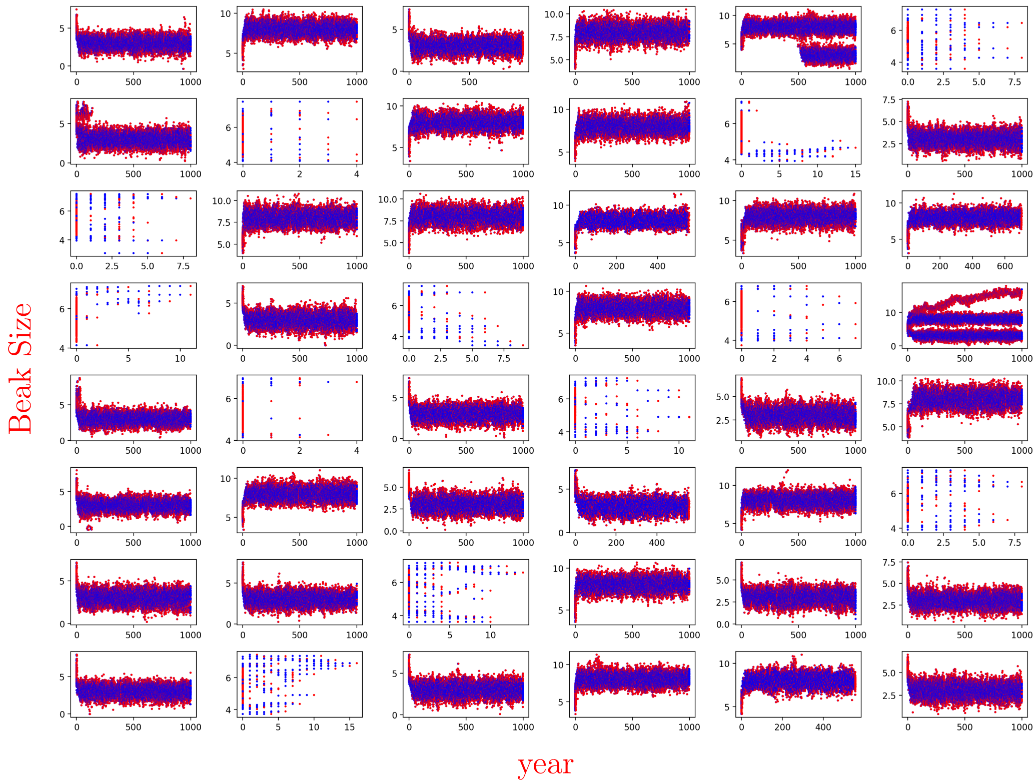

Figure A.32: Performance of the poly-weighted dominance model with low variance under the trimodal small seed distribution. Part 2.

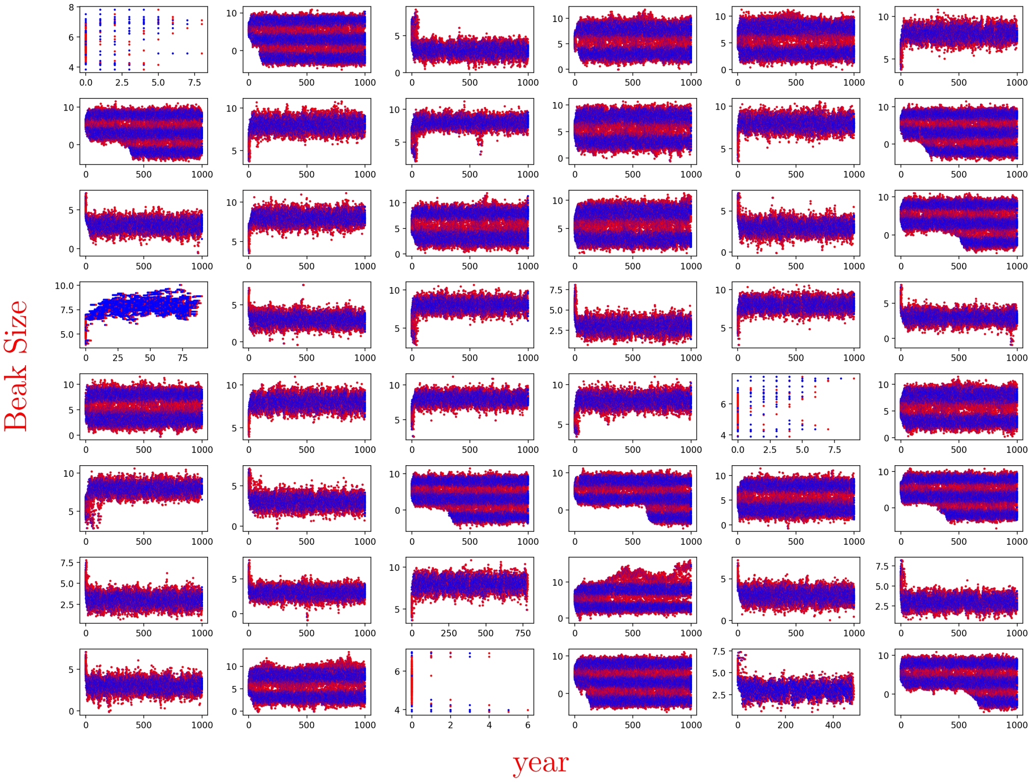

Figure A.33: Performance of the mono-weighted dominance model with high variance under the trimodal small seed distribution. Part 1.

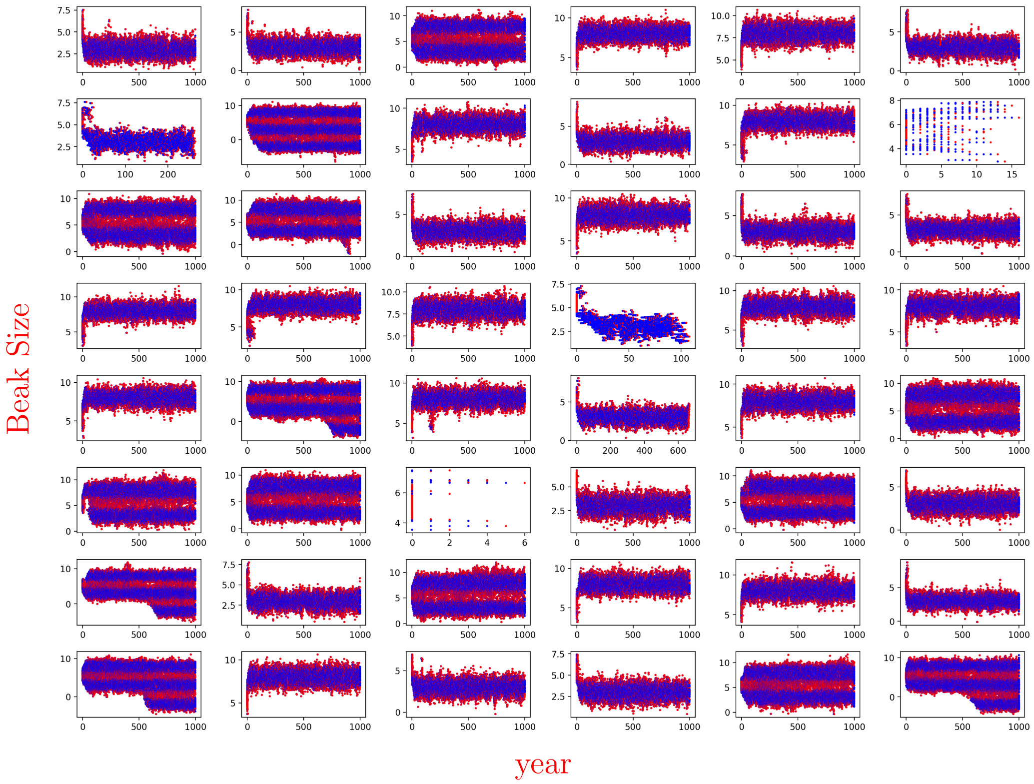

Figure A.34: Performance of the mono-weighted dominance model with high variance under the trimodal small seed distribution. Part 2.

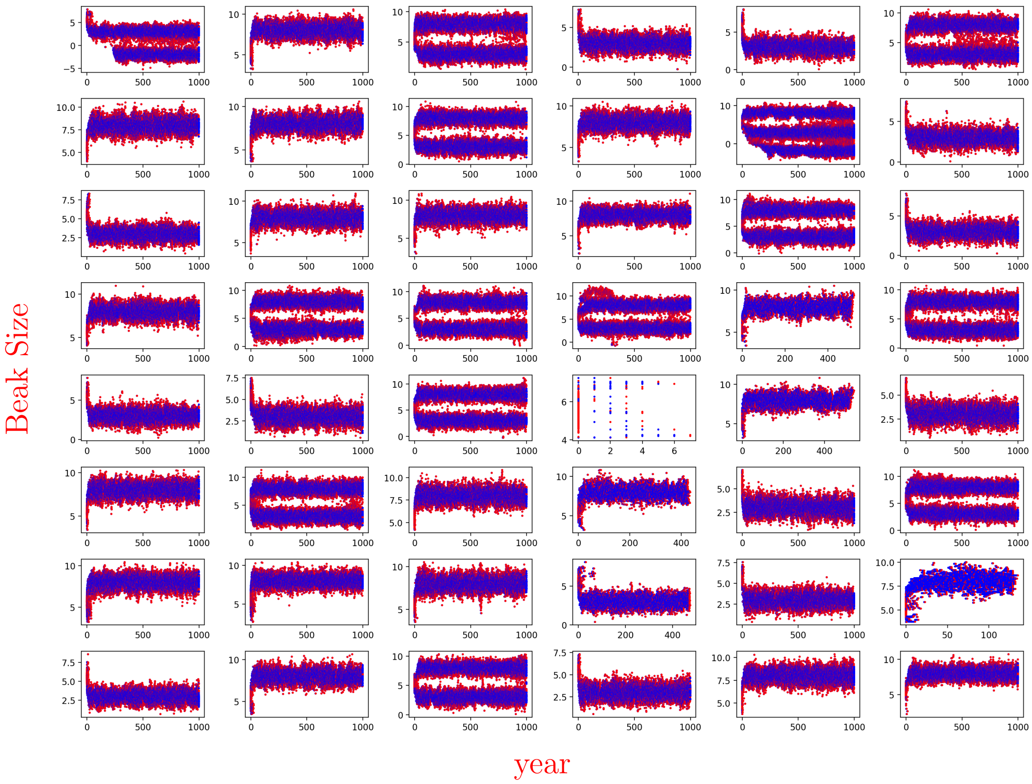

Figure A.35: Performance of the mono-weighted dominance model with low variance under the trimodal small seed distribution. Part 1.

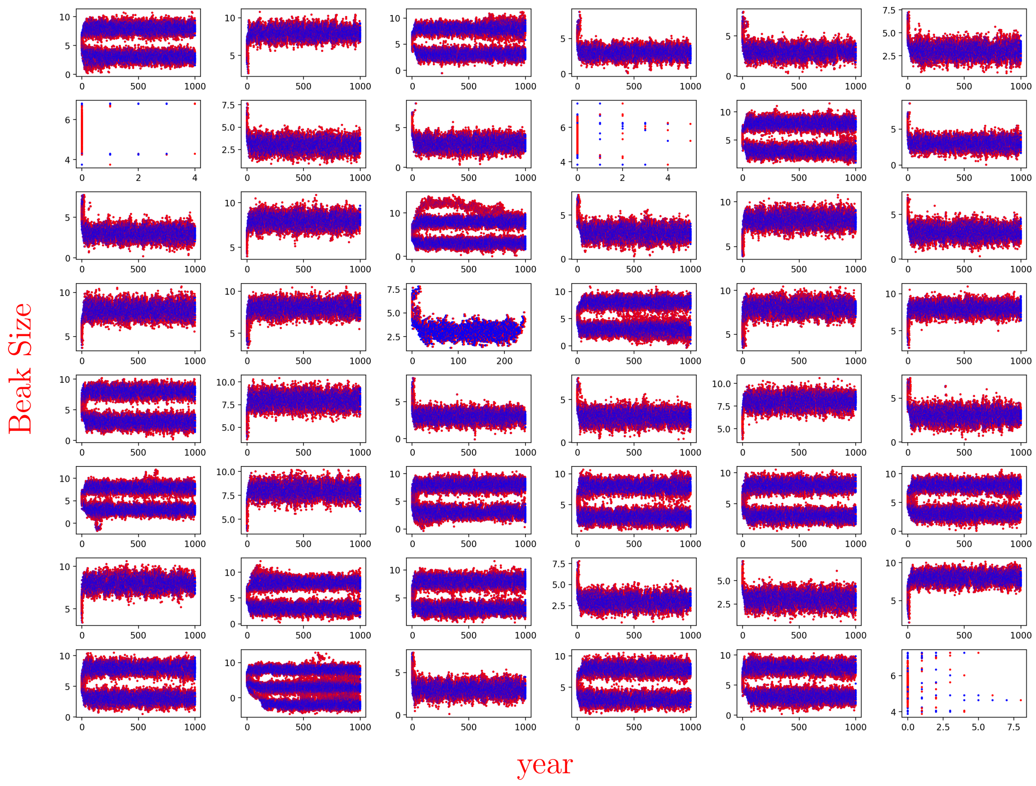

Figure A.36: Performance of the mono-weighted dominance model with low variance under the trimodal small seed distribution. Part 2.

Appendix B

Histograms

The following are the histograms constructed of each of the 96 trials for all combinations of genetic models and seed distributions. Green bars represent bins that featured a large enough population to be counted. Red bars represent bins that did not make the 15% cutoff. Due to the number of trials only half of the trials in a run can fit on one page. As such, each run is split into two parts, labeled “Part 1” and “Part 2” respectively.

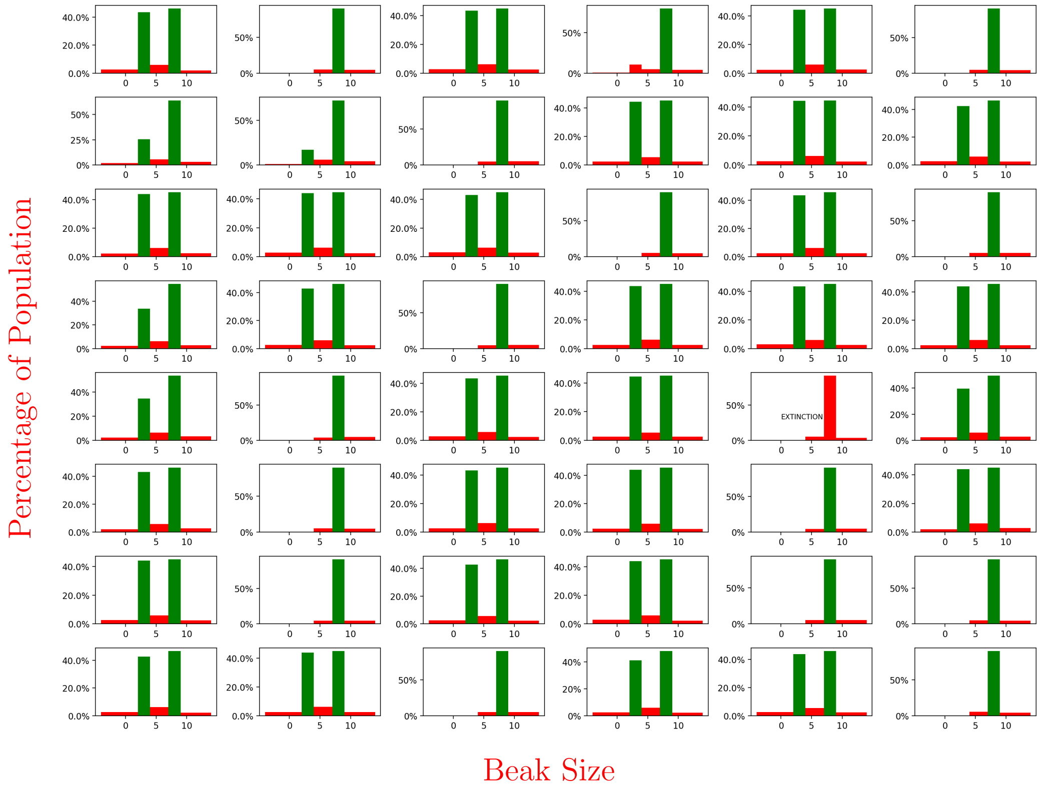

Figure B.1: Histogram of the complete dominance model under the bimodal seed distribution. Part 1.

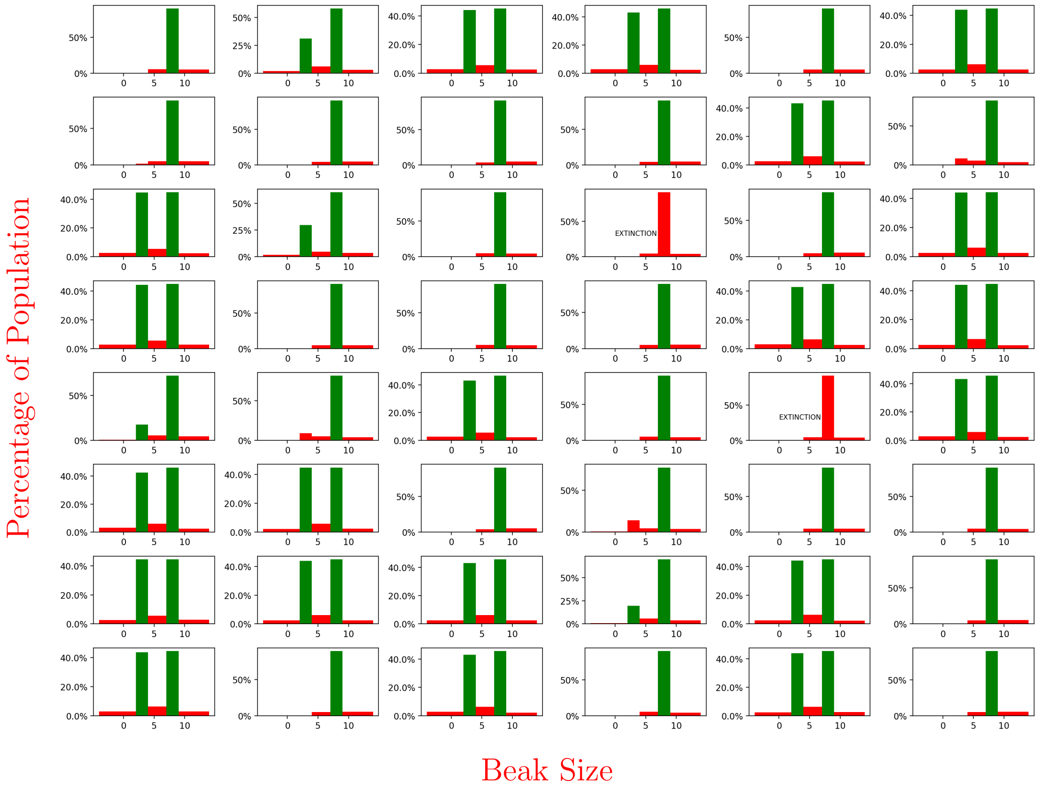

Figure B.2: Histogram of the complete dominance model under the bimodal seed distribution. Part 2.

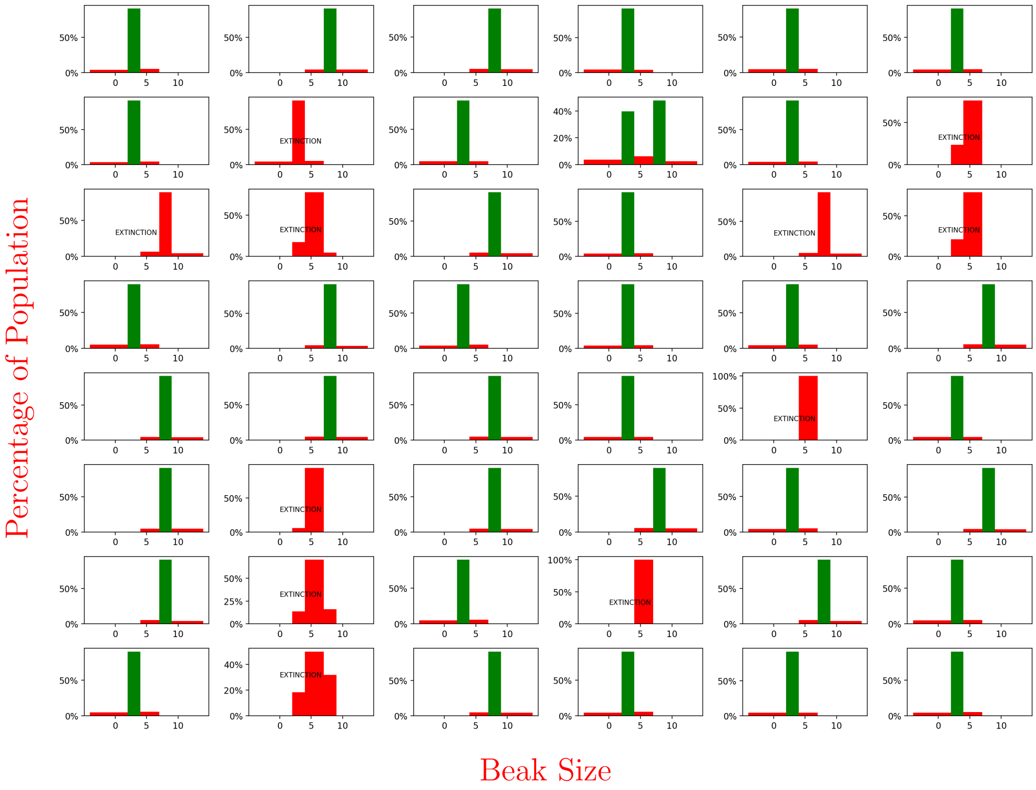

Figure B.3: Histogram of the incomplete dominance model under the bimodal seed distribution. Part 1.

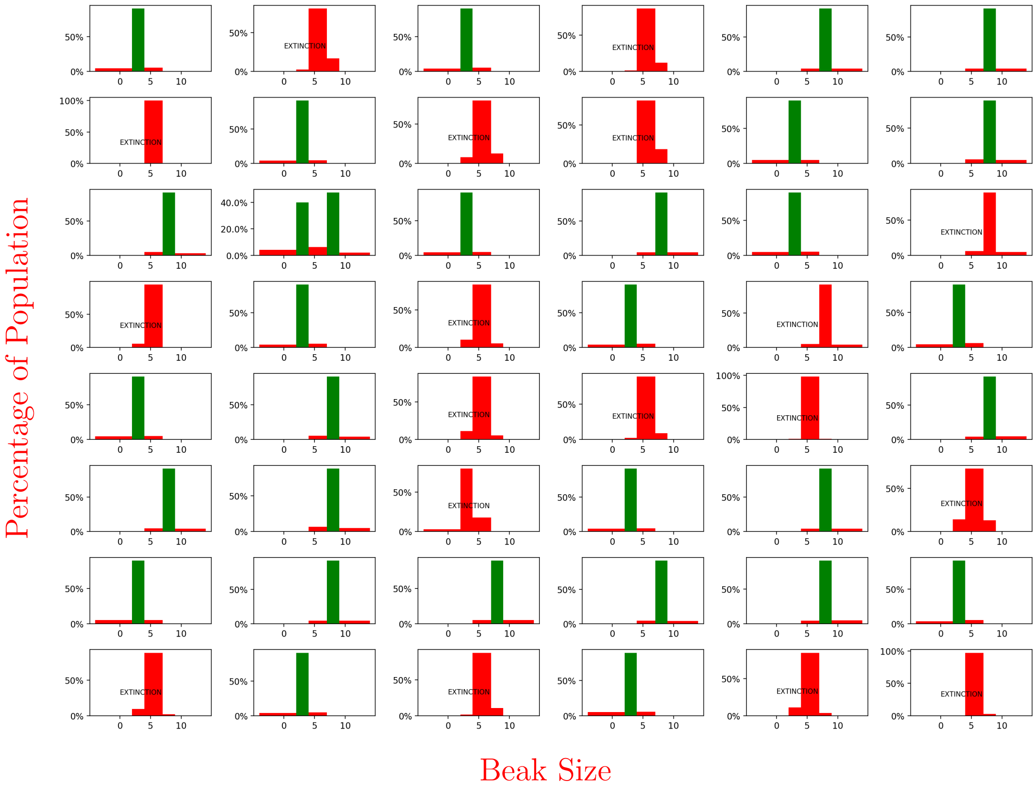

Figure B.4: Histogram of the incomplete dominance model under the bimodal seed distribution. Part 2.

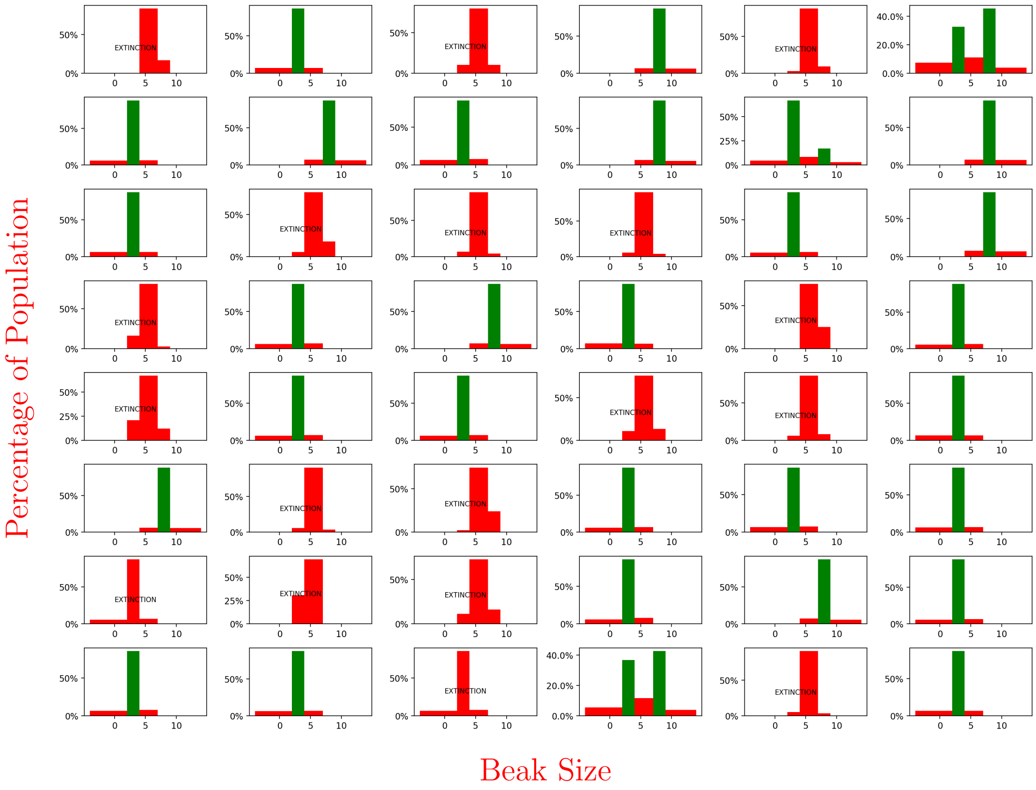

Figure B.5: Histogram of the poly-weighted dominance model with high variance under the bimodal seed distribution. Part 1.

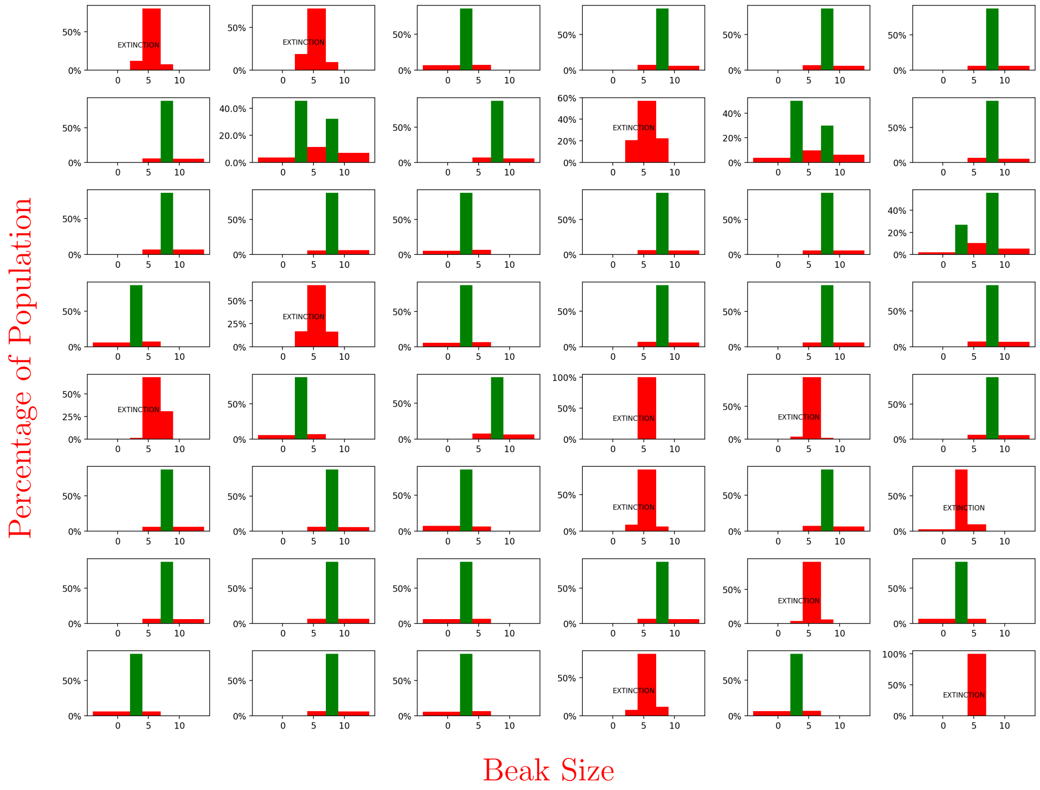

Figure B.6: Histogram of the poly-weighted dominance model with high variance under the bimodal seed distribution. Part 2.

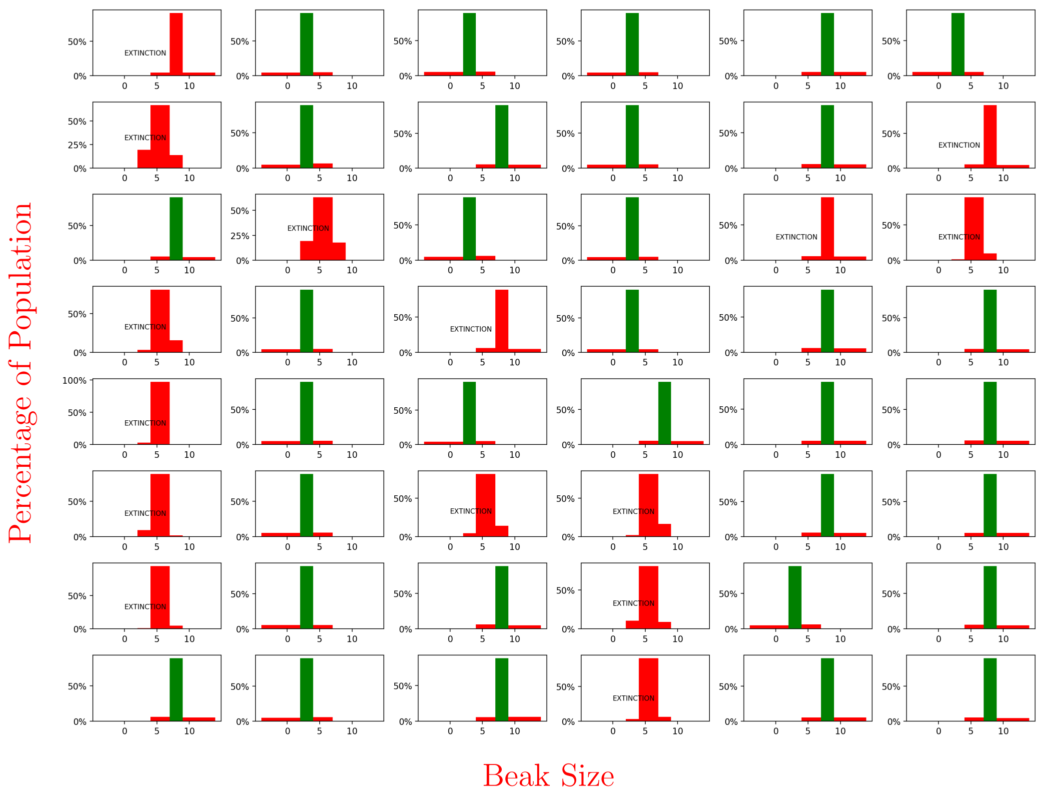

Figure B.7: Histogram of the poly-weighted dominance model with low variance under the bimodal seed distribution. Part 1.

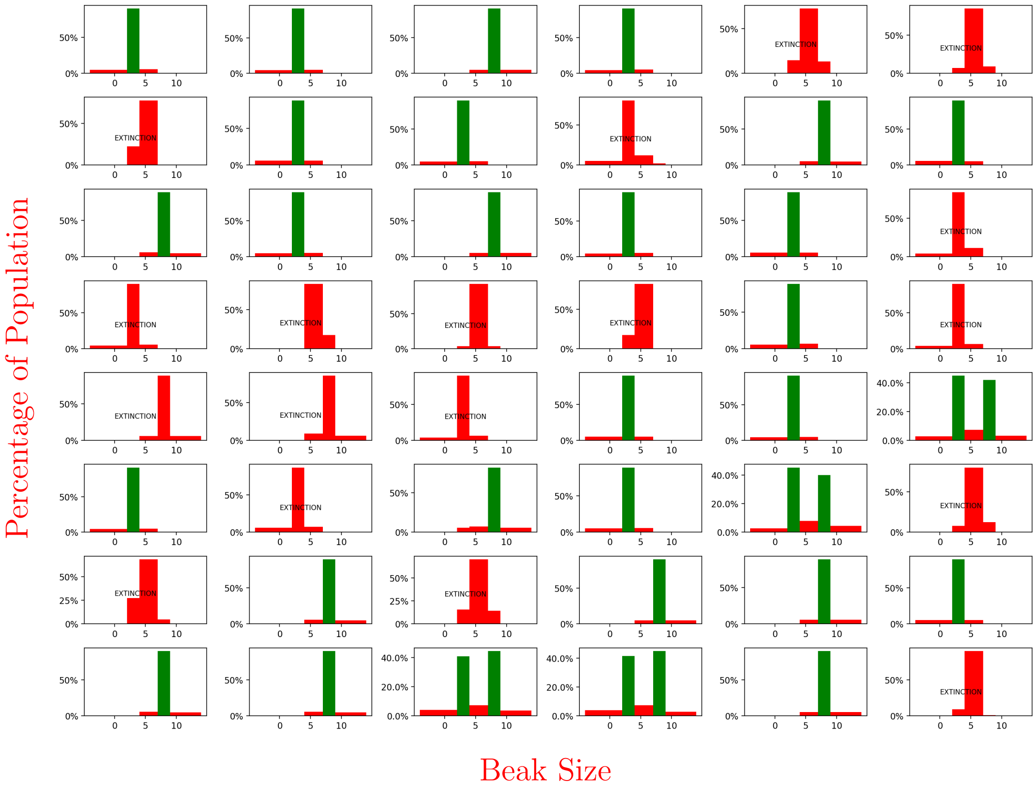

Figure B.8: Histogram of the poly-weighted dominance model with low variance under the bimodal seed distribution. Part 2.

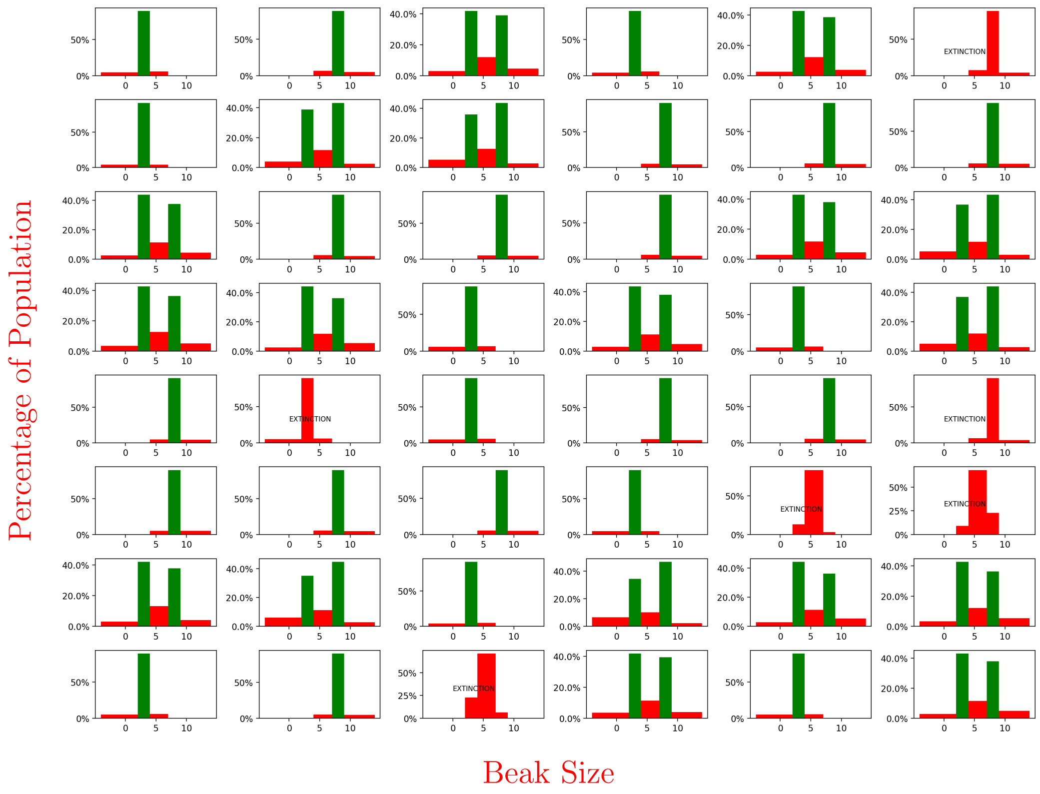

Figure B.9: Histogram of the mono-weighted dominance model with high variance under the bimodal seed distribution. Part 1.

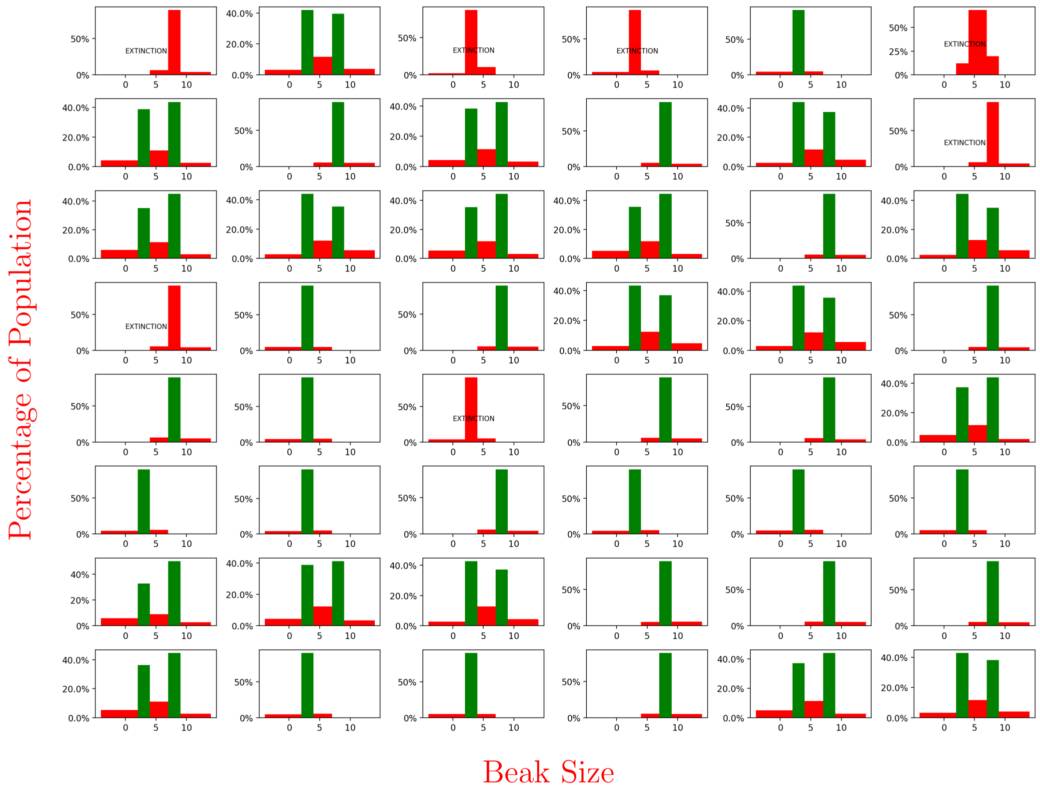

Figure B.10: Histogram of the mono-weighted dominance model with high variance under the bimodal seed distribution. Part 2.

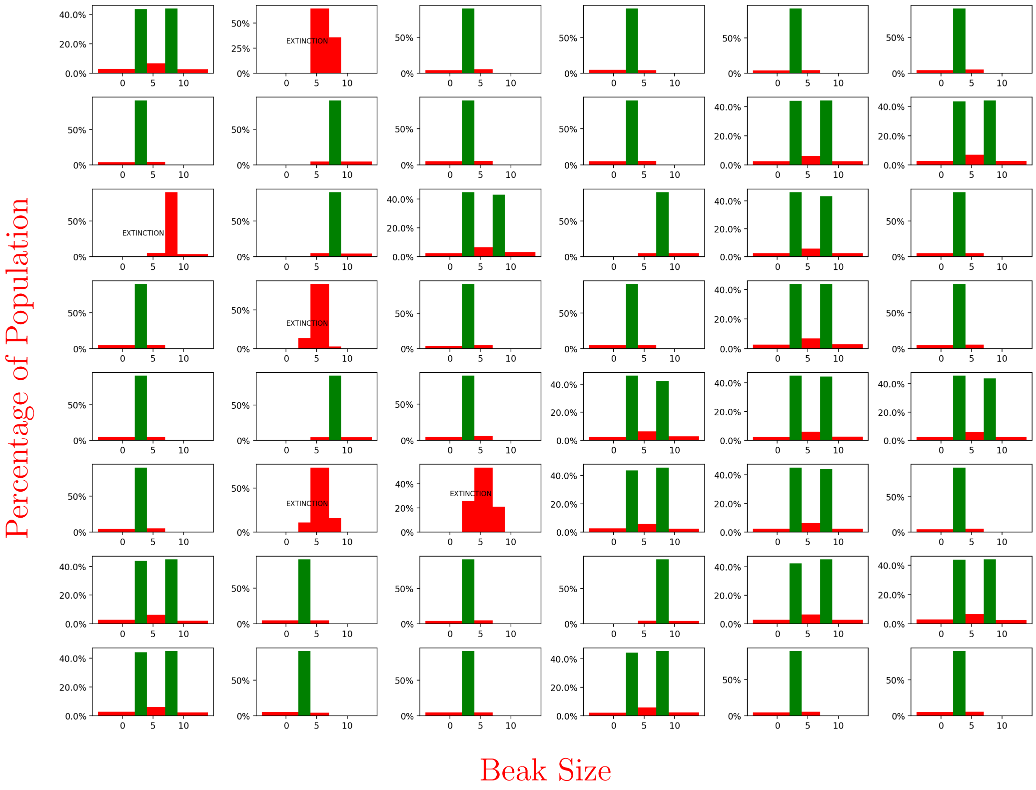

Figure B.11: Histogram of the mono-weighted dominance model with low variance under the bimodal seed distribution. Part 1.

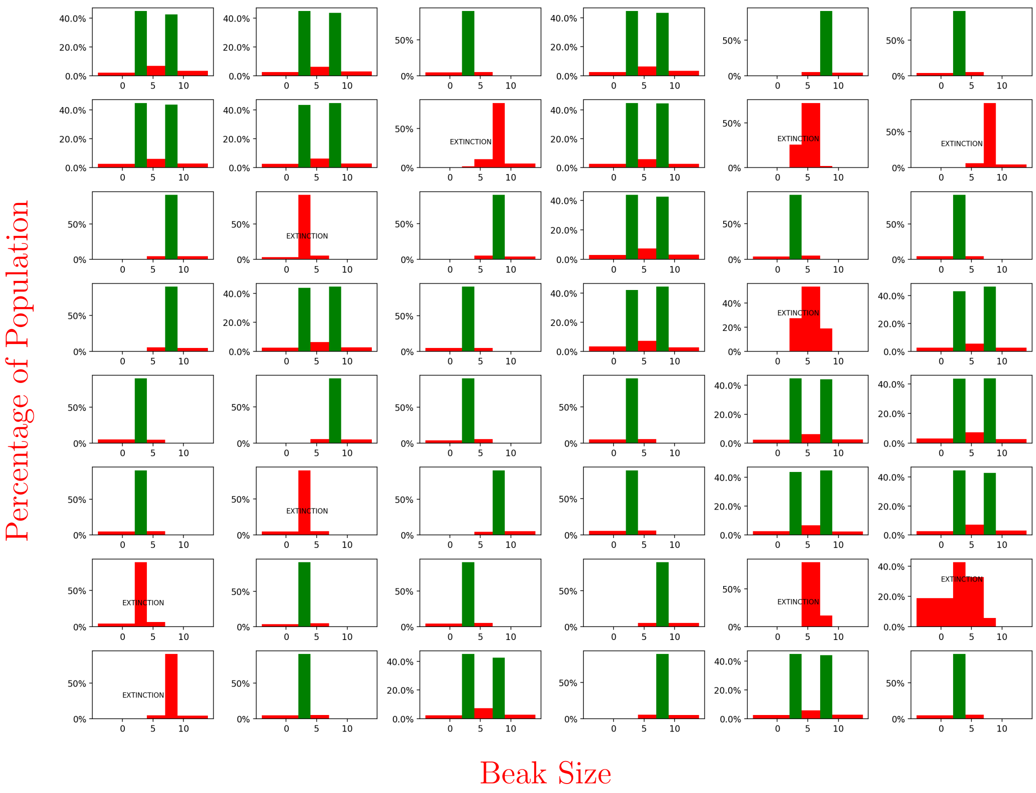

Figure B.12: Histogram of the mono-weighted dominance model with low variance under the bimodal seed distribution. Part 2.

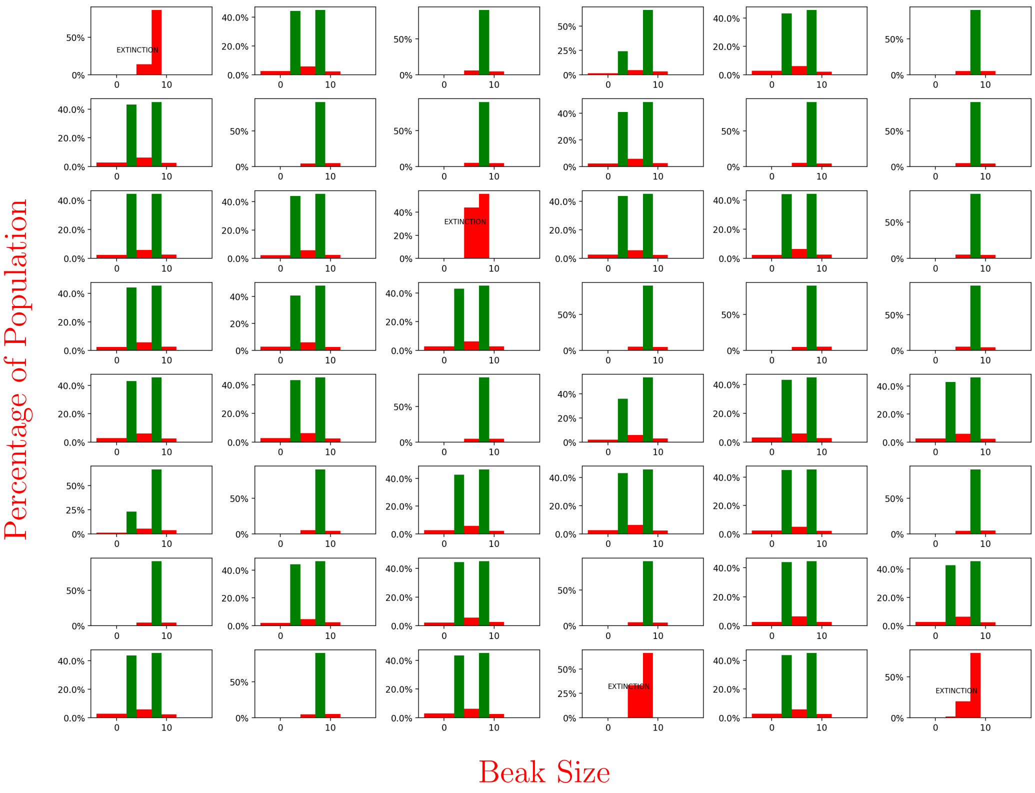

Figure B.13: Histogram of the complete dominance model under the trimodal large seed distribution. Part 1.

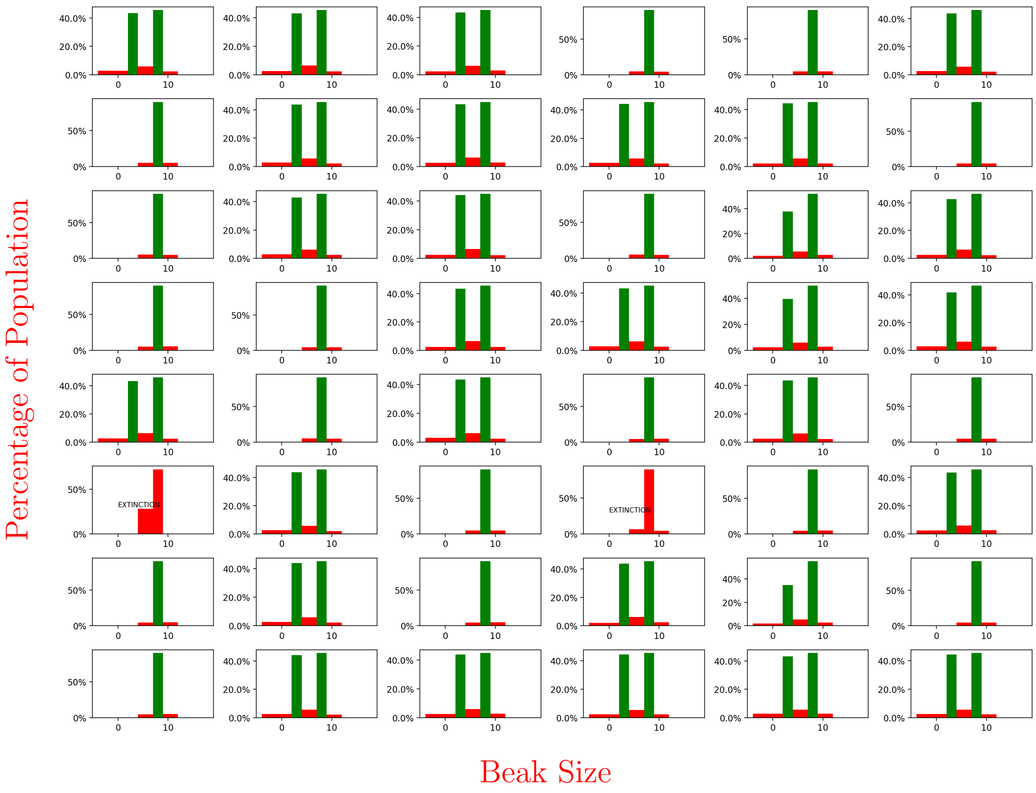

Figure B.14: Histogram of the complete dominance model under the trimodal large seed distribution. Part 2.

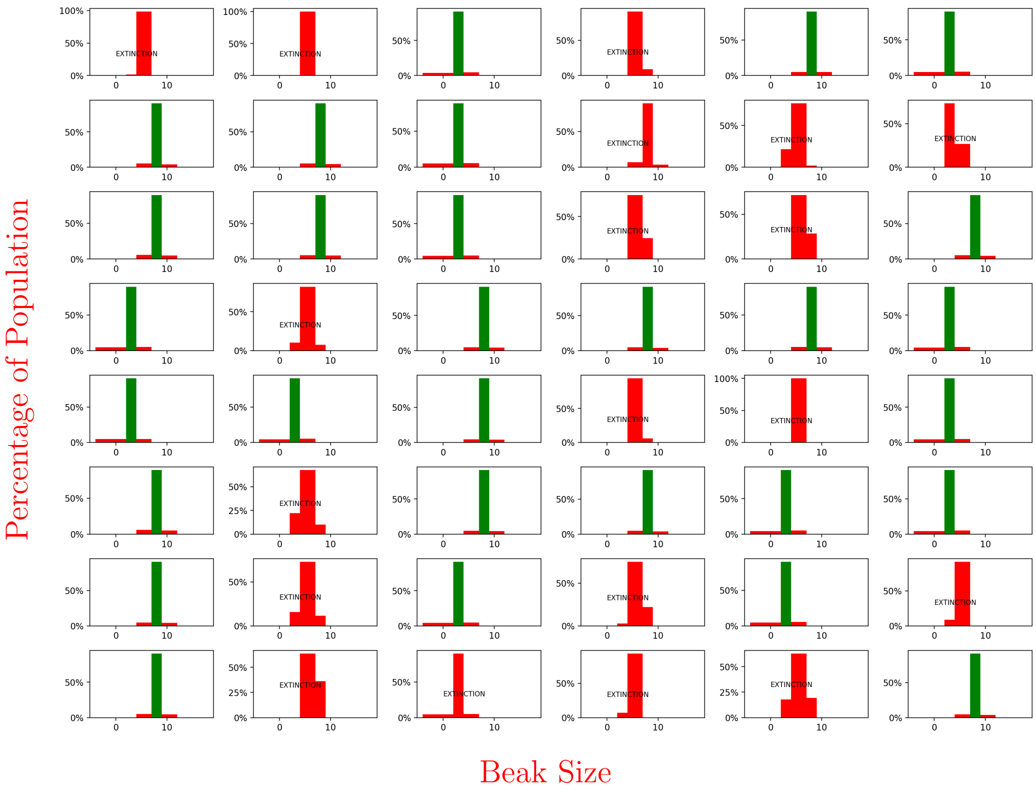

Figure B.15: Histogram of the incomplete dominance model under the trimodal large seed distribution. Part 1.

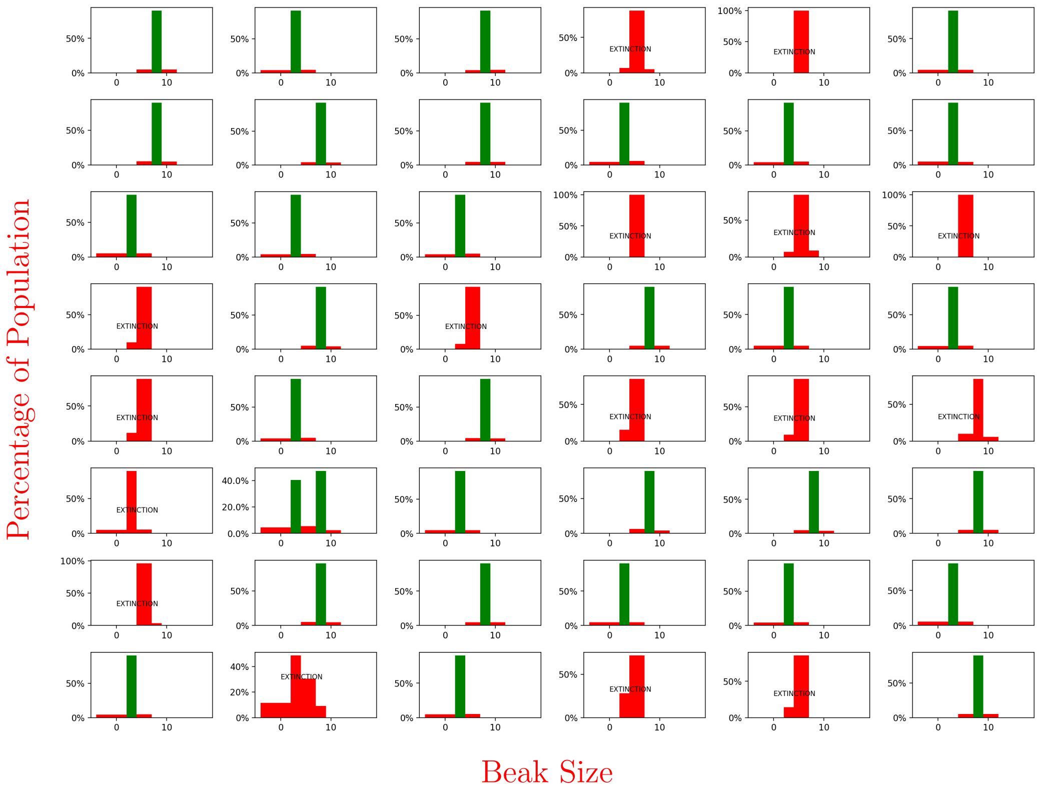

Figure B.16: Histogram of the incomplete dominance model under the trimodal large seed distribution. Part 2.

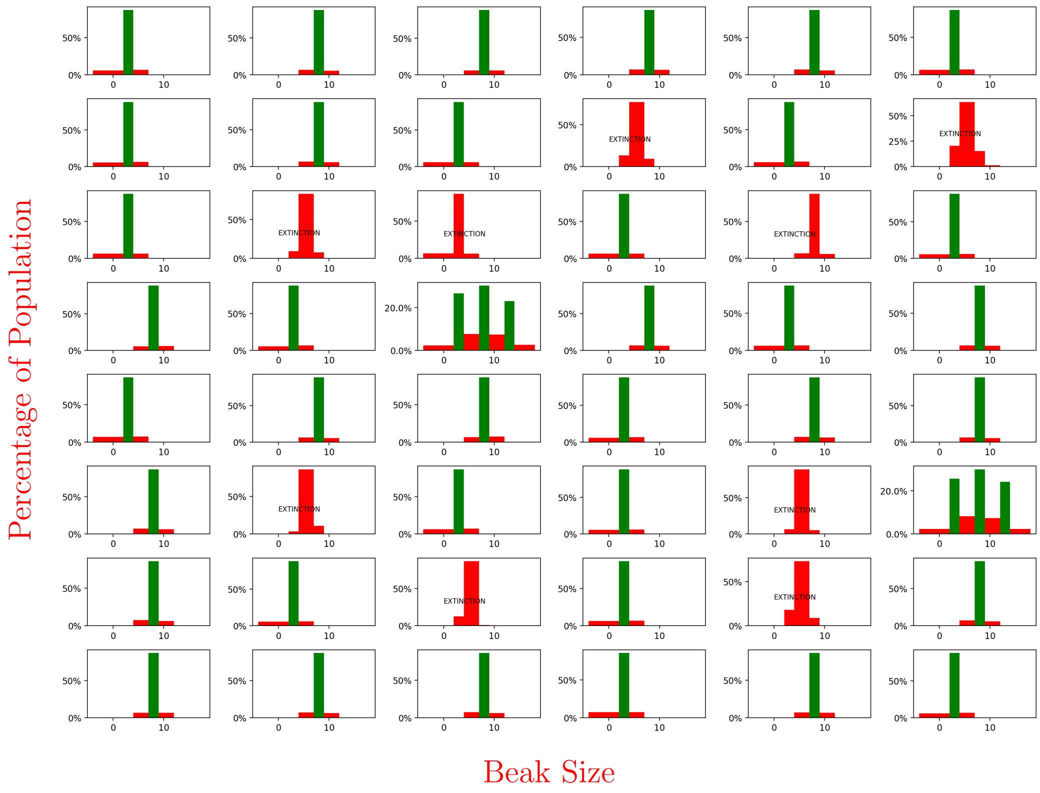

Figure B.17: Histogram of the poly-weighted dominance model with high variance under the trimodal large seed distribution. Part 1.

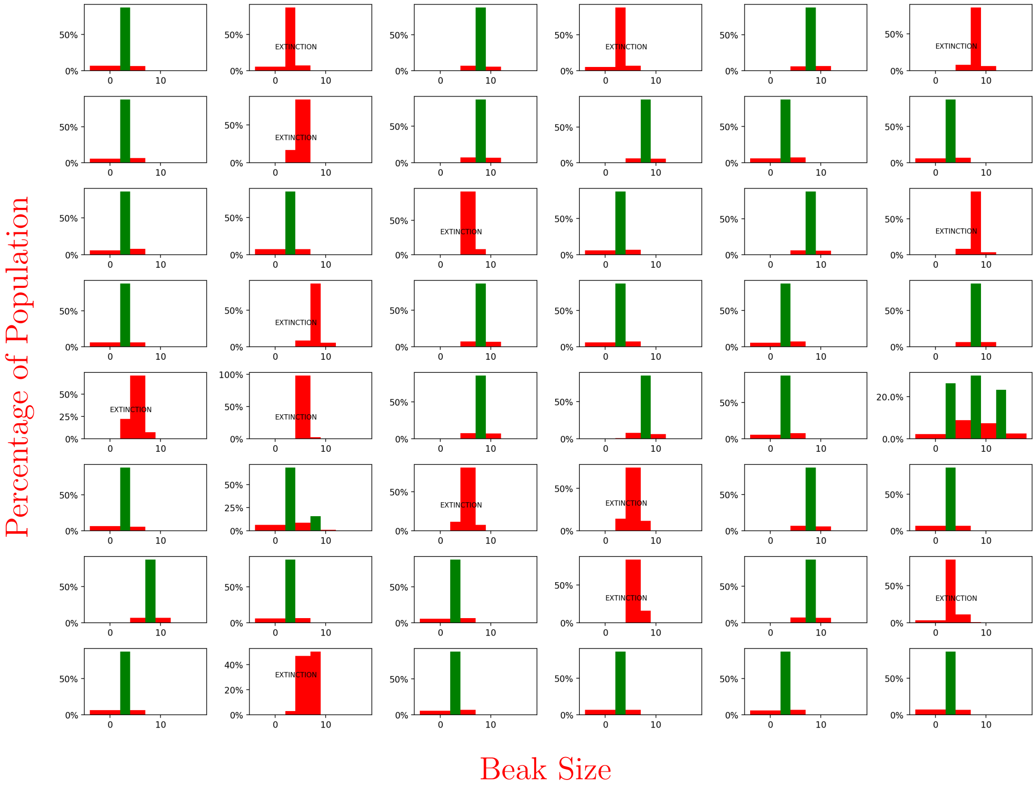

Figure B.18: Histogram of the poly-weighted dominance model with high variance under the trimodal large seed distribution. Part 2.

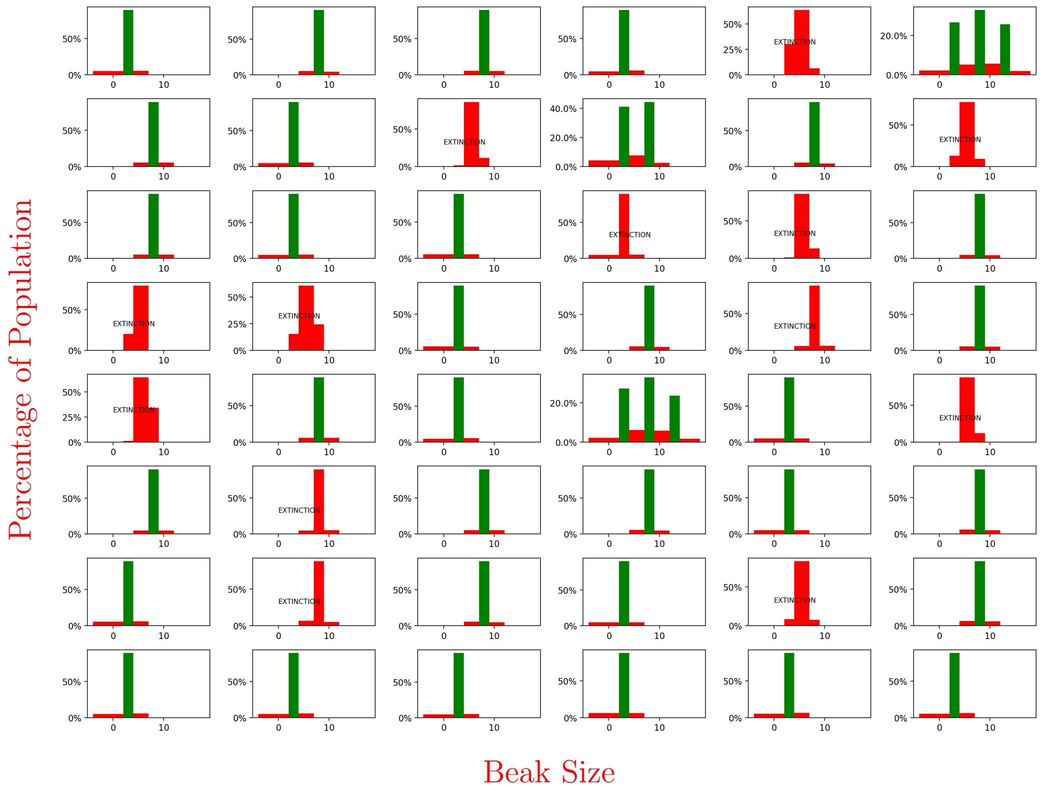

Figure B.19: Histogram of the poly-weighted dominance model with low variance under the trimodal large seed distribution. Part 1.

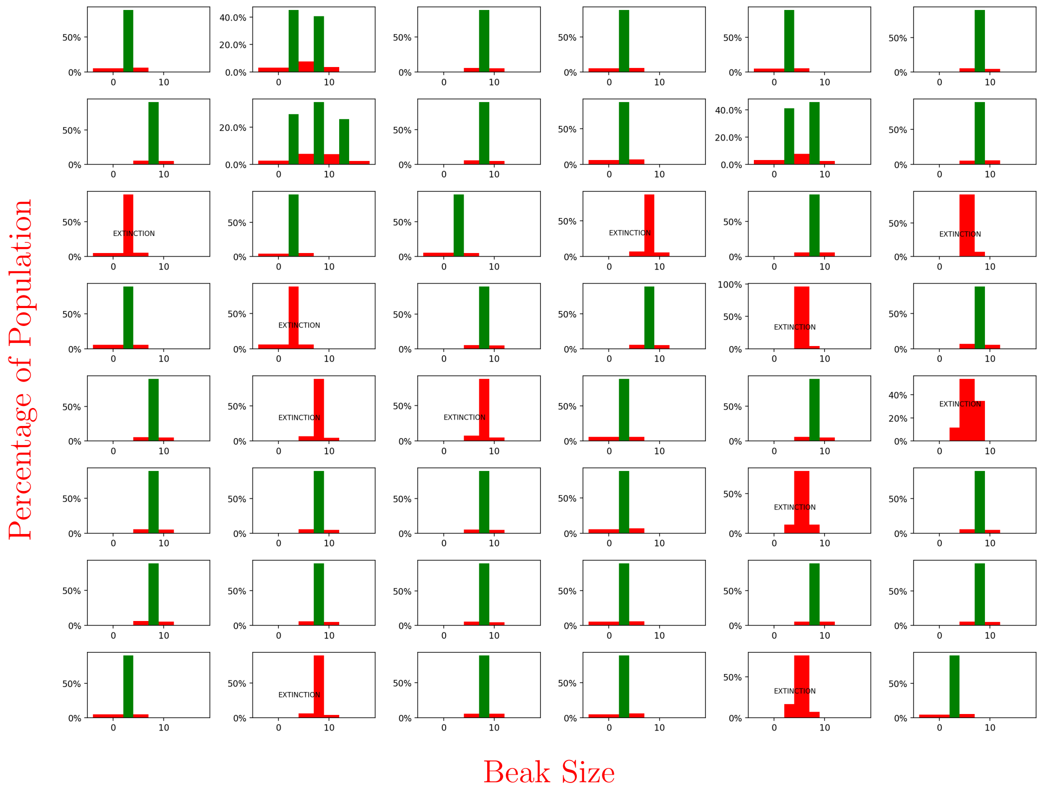

Figure B.20: Histogram of the poly-weighted dominance model with low variance under the trimodal large seed distribution. Part 2.

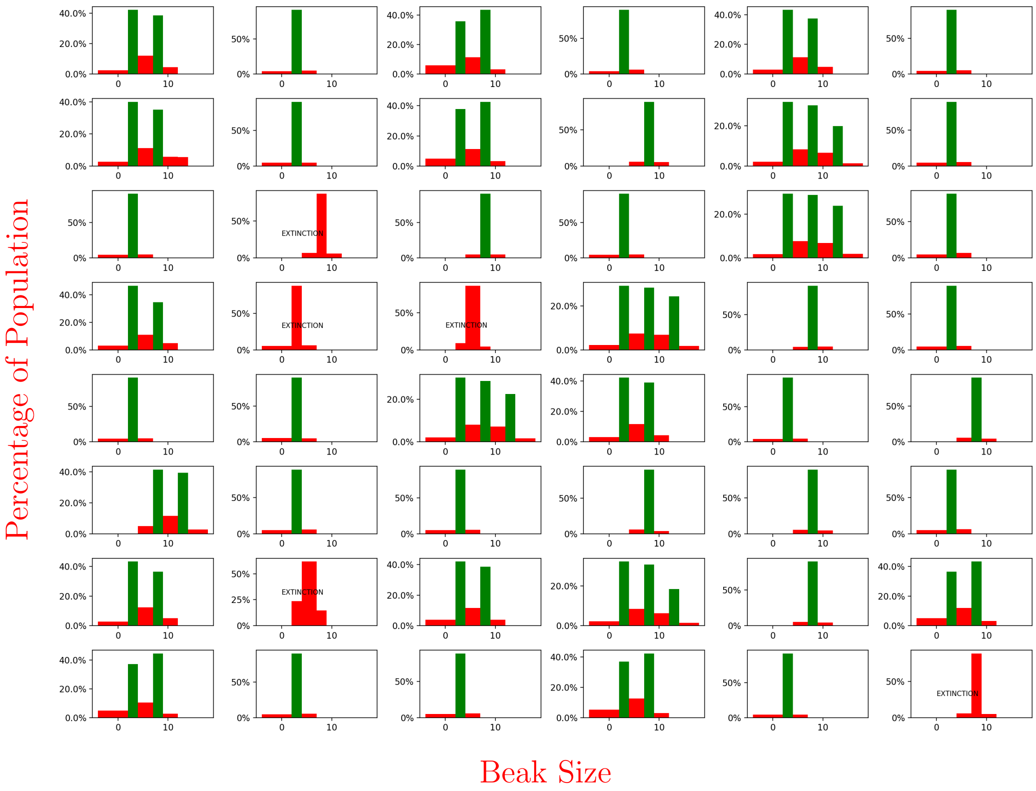

Figure B.21: Histogram of the mono-weighted dominance model with high variance under the trimodal large seed distribution. Part 1.

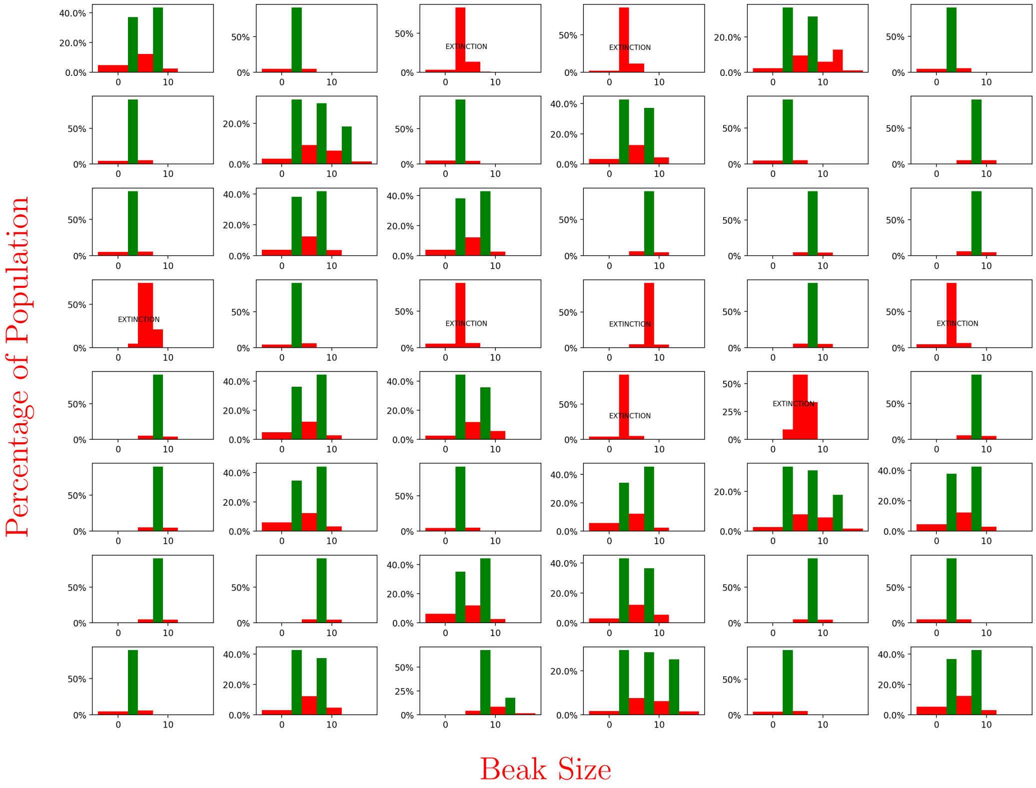

Figure B.22: Histogram of the mono-weighted dominance model with high variance under the trimodal large seed distribution. Part 2.

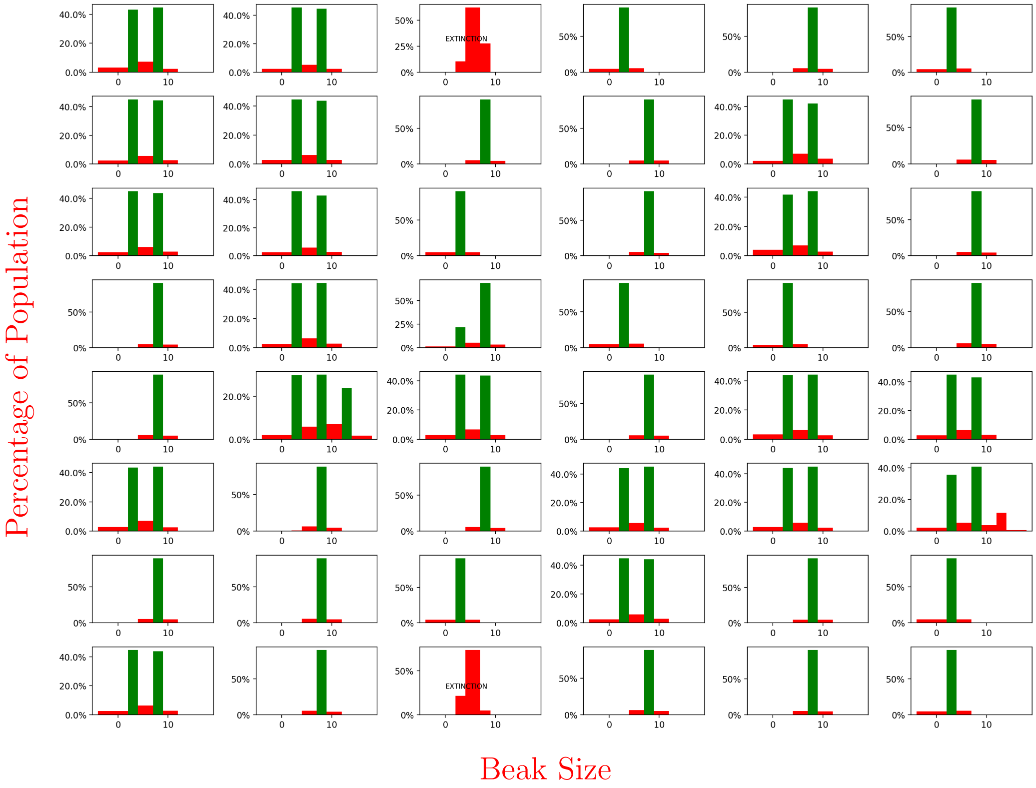

Figure B.23: Histogram of the mono-weighted dominance model with low variance under the trimodal large seed distribution. Part 1.

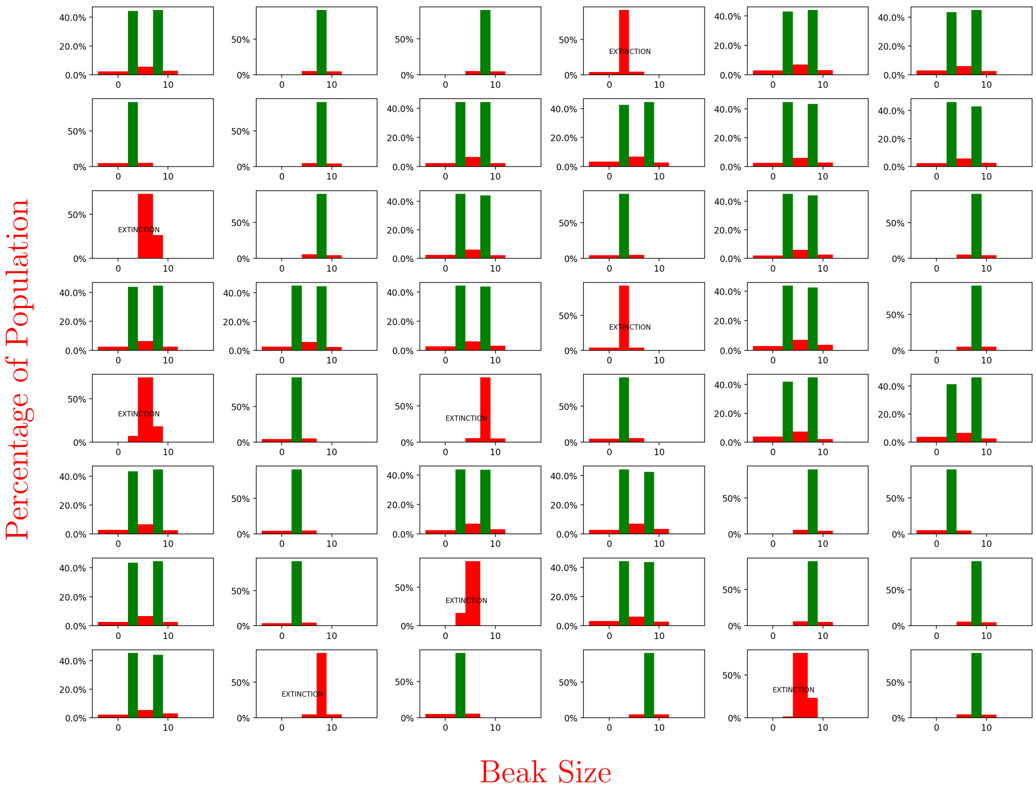

Figure B.24: Histogram of the mono-weighted dominance model with low variance under the trimodal large seed distribution. Part 2.

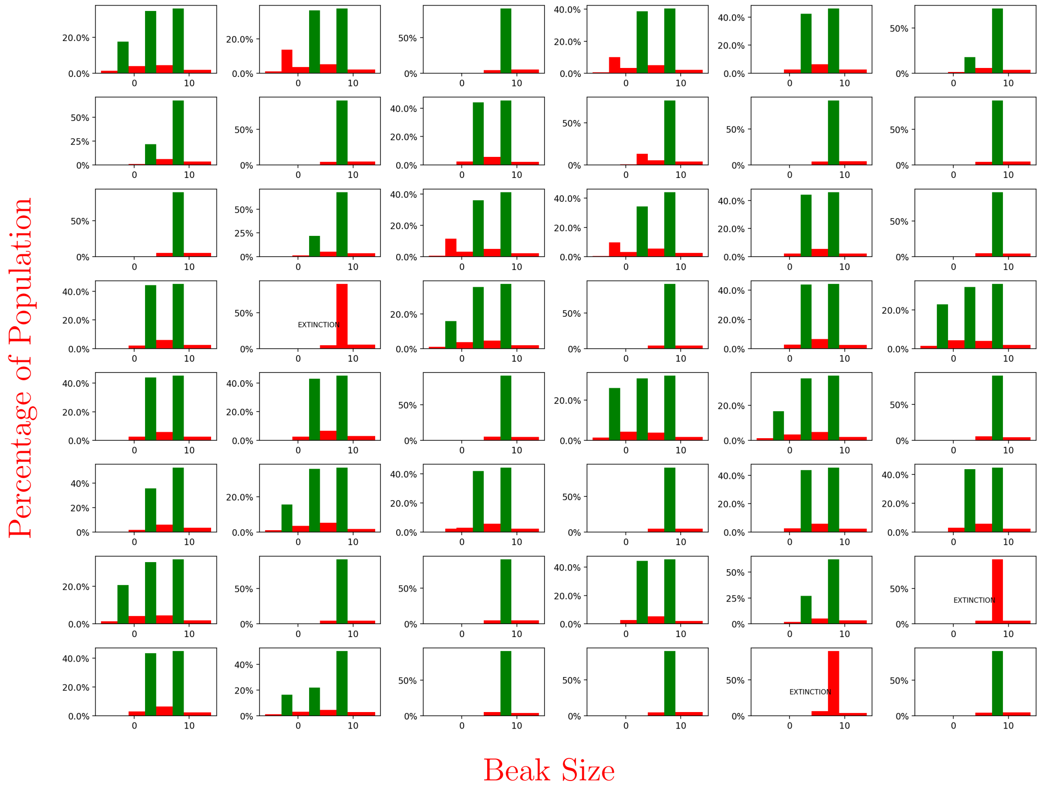

Figure B.25: Histogram of the complete dominance model under the trimodal small seed distribution. Part 1.

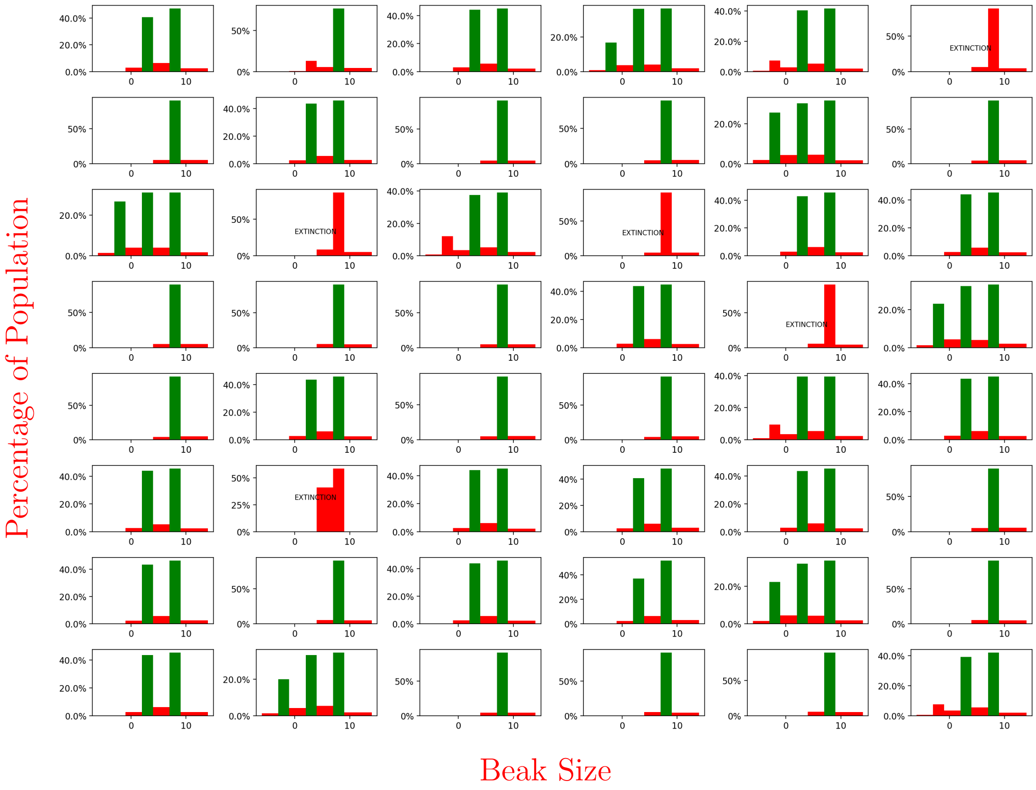

Figure B.26: Histogram of the complete dominance model under the trimodal small seed distribution. Part 2.

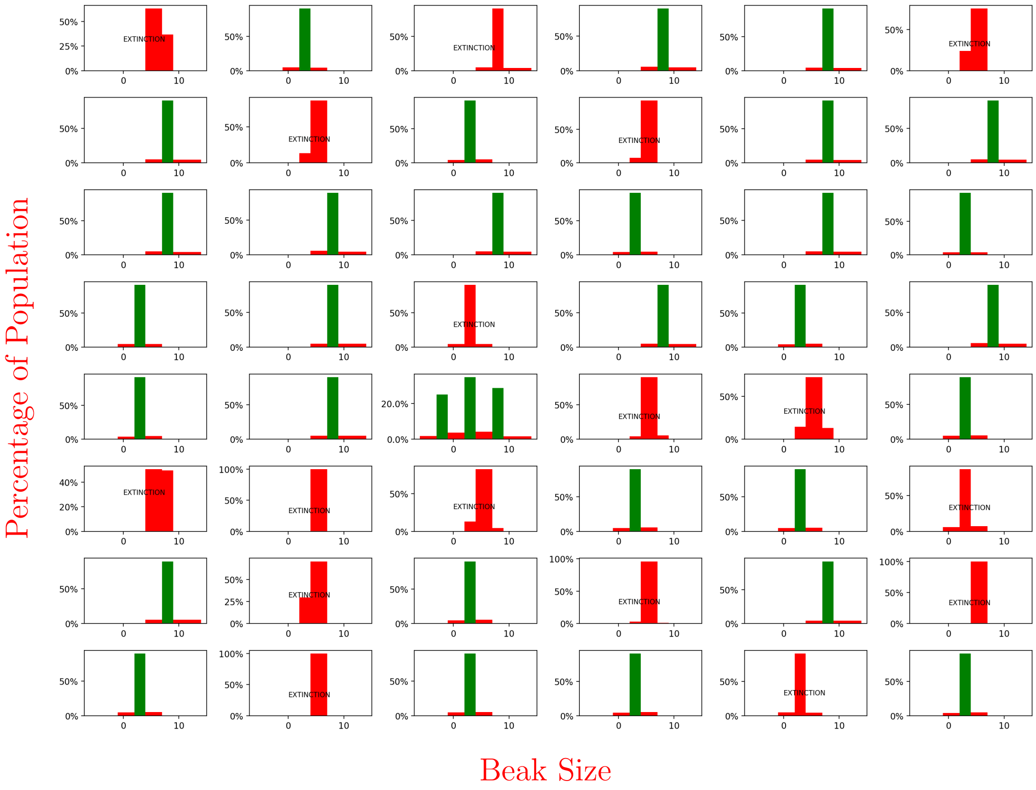

Figure B.27: Histogram of the incomplete dominance model under the trimodal small seed distribution. Part 1.

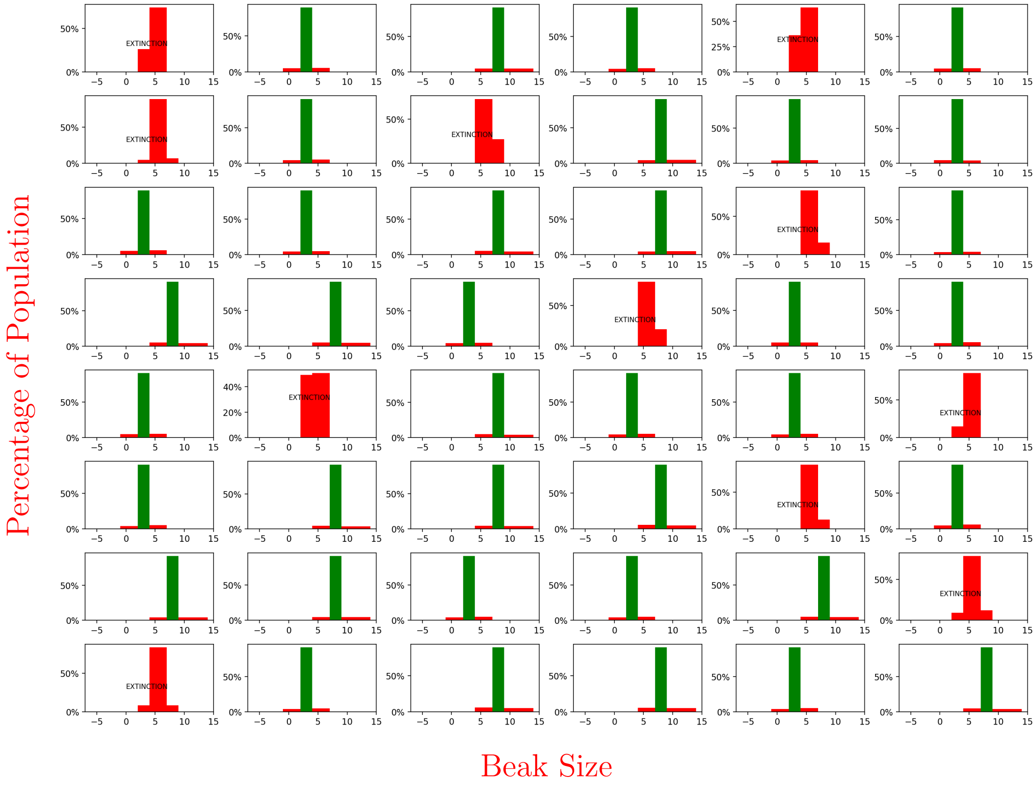

Figure B.28: Histogram of the incomplete dominance model under the trimodal small seed distribution. Part 2.

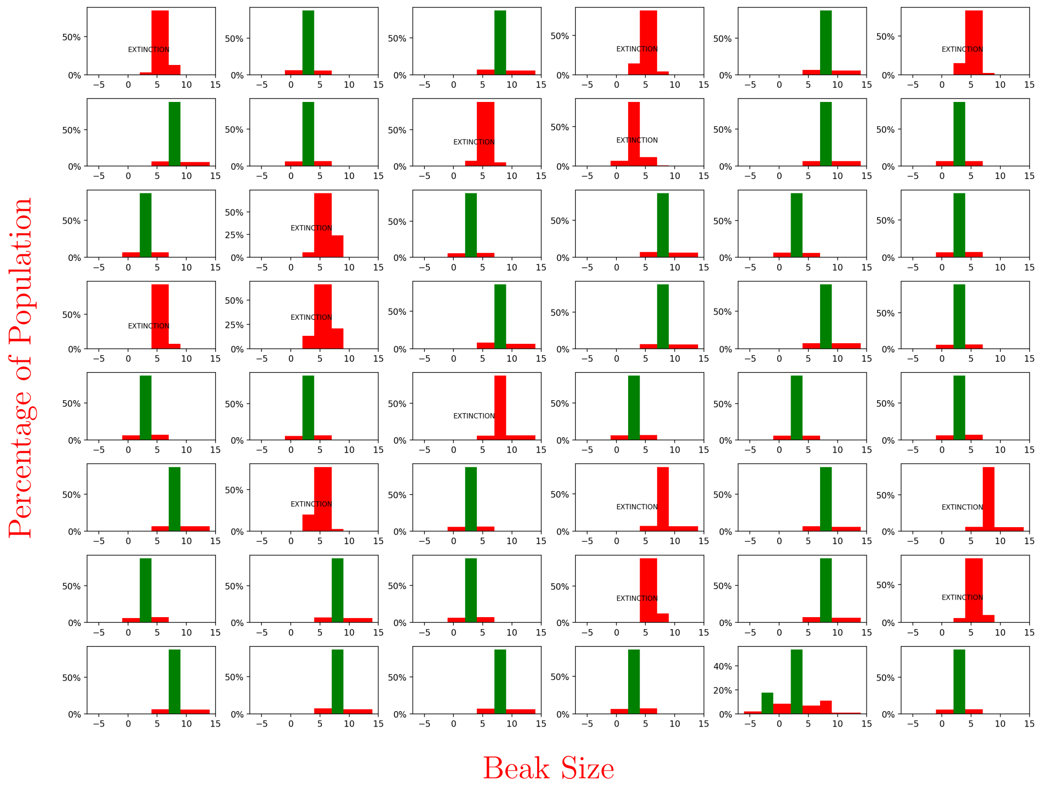

Figure B.29: Histogram of the poly-weighted dominance model with high variance under the trimodal small seed distribution. Part 1.

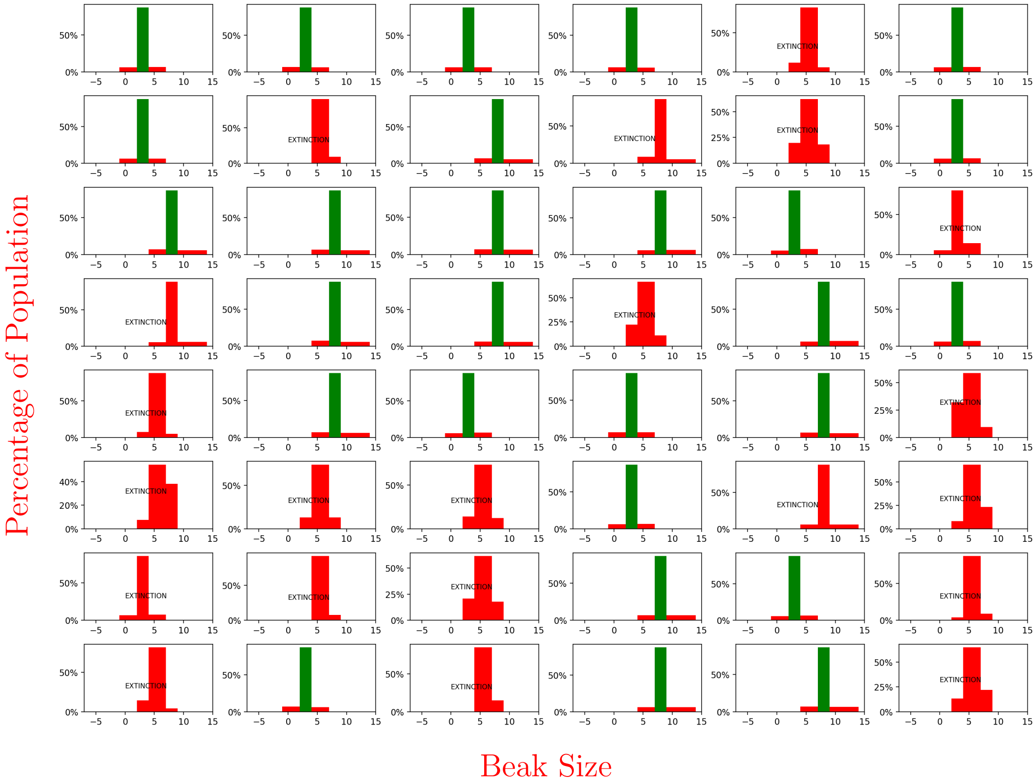

Figure B.30: Histogram of the poly-weighted dominance model with high variance under the trimodal small seed distribution. Part 2.

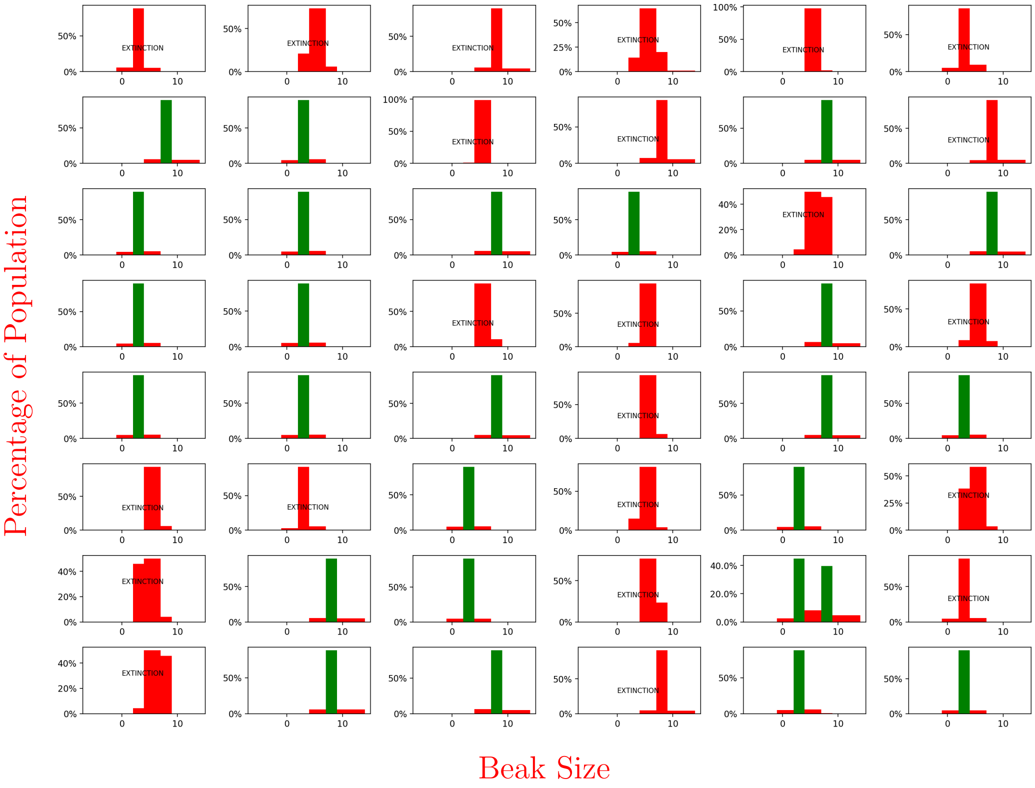

Figure B.31: Histogram of the poly-weighted dominance model with low variance under the trimodal small seed distribution. Part 1.

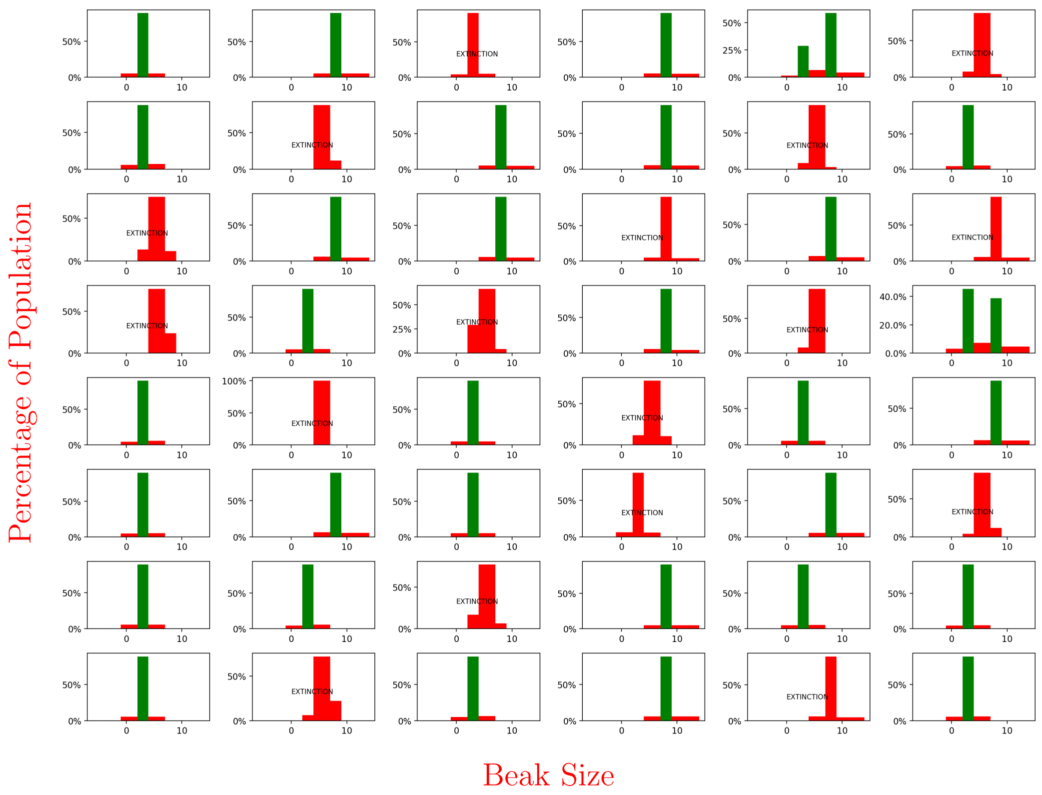

Figure B.32: Histogram of the poly-weighted dominance model with low variance under the trimodal small seed distribution. Part 2.

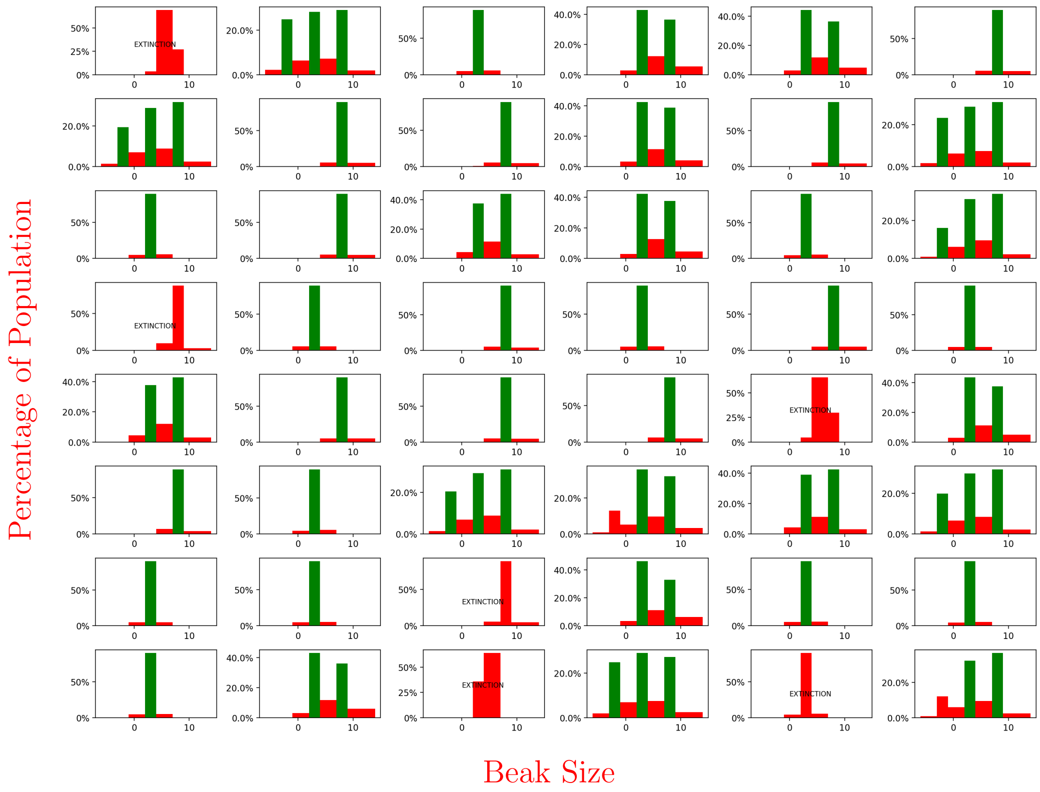

Figure B.33: Histogram of the mono-weighted dominance model with high variance under the trimodal small seed distribution. Part 1.

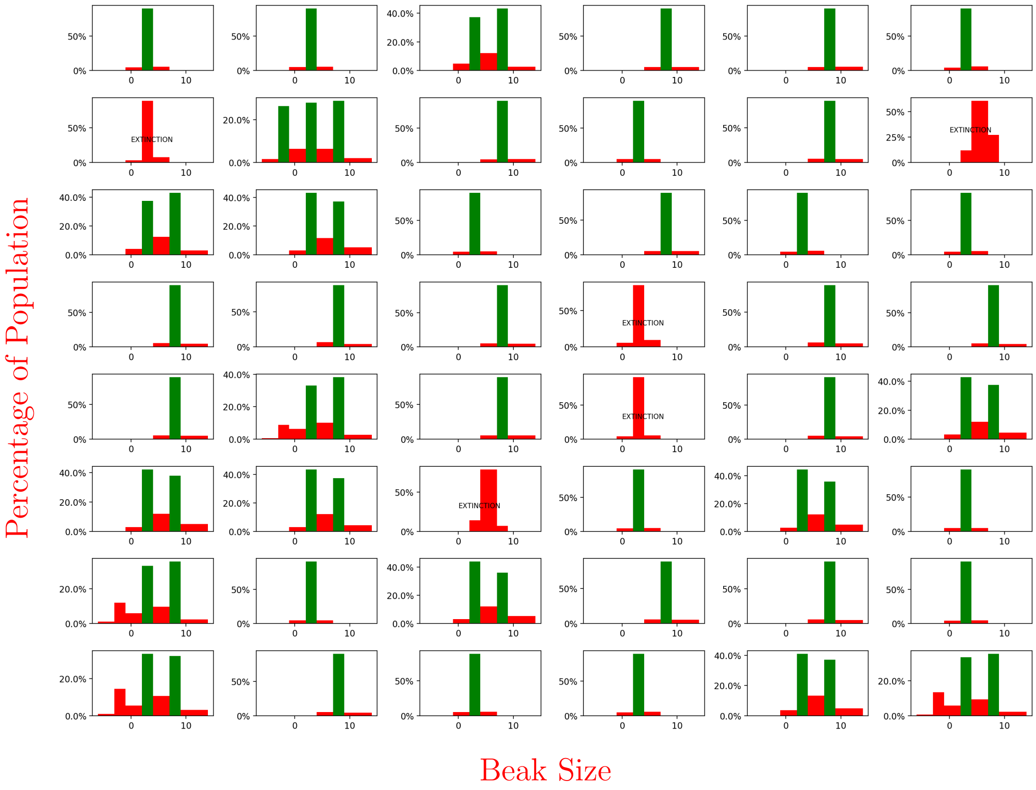

Figure B.34: Histogram of the mono-weighted dominance model with high variance under the trimodal small seed distribution. Part 2.

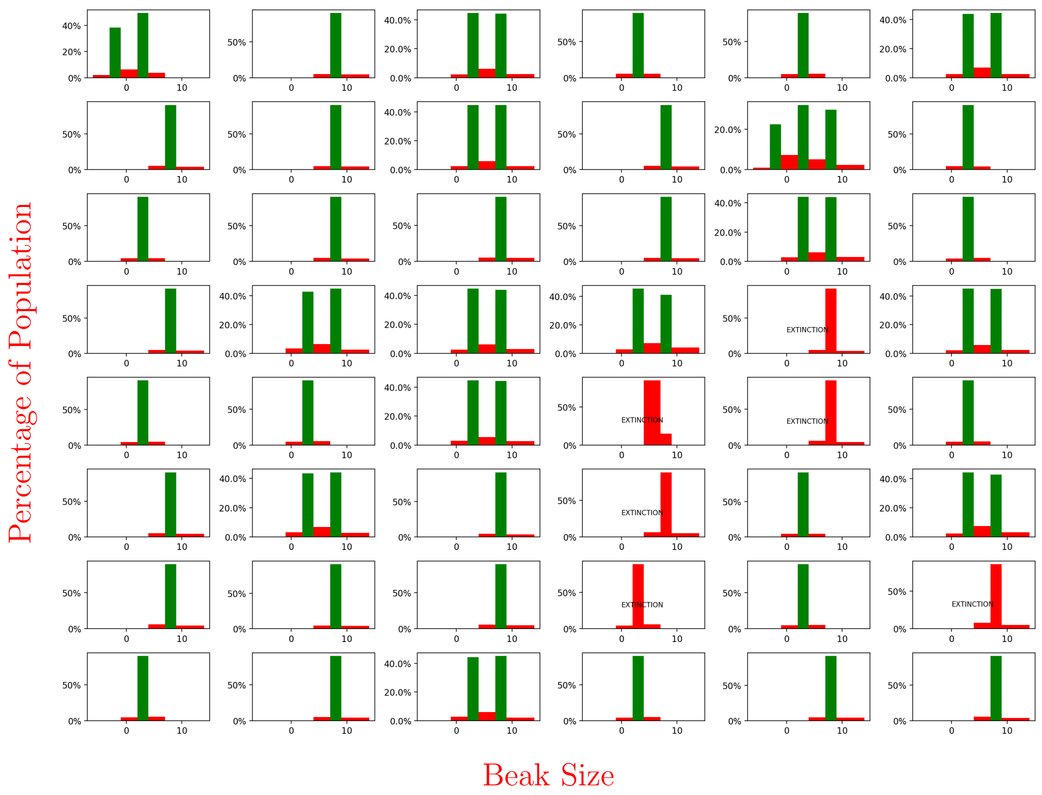

Figure B.35: Histogram of the mono-weighted dominance model with low variance under the trimodal small seed distribution. Part 1.

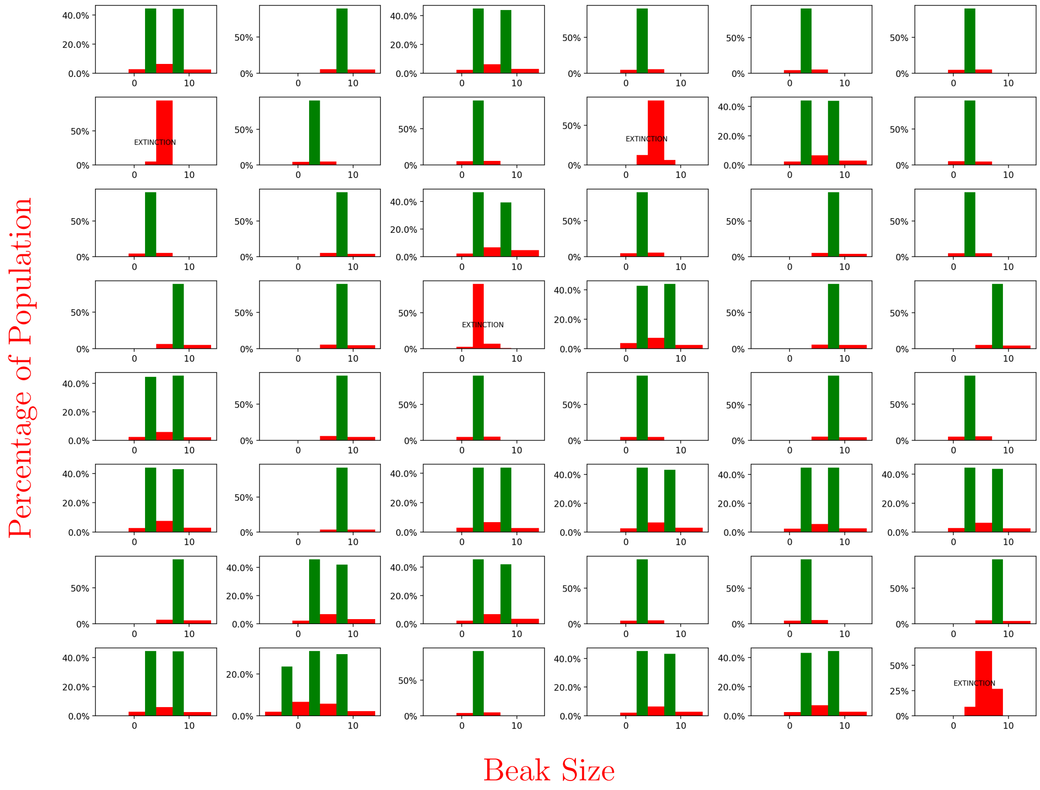

Figure B.36: Histogram of the mono-weighted dominance model with low variance under the trimodal small seed distribution. Part 2.