A stochastic multi-objective multi-stage decision model for multiple generation mate-selection problems

H. A. Abbass Abbass@fit.qut.edu.au
G. Finn gerry@fit.qut.edu.au
M. Towsey towsey@fit.qut.edu.au

Machine Learning Research Center,
Data Mining Lab,
School of Computing Science
QUT,GPO Box 2434, Australia
Technical Report FIT-TR-99-02

Abstract

Designing a breeding program has been an issue of primary concern on much animal breeding research has focused in the last decade. In general the outcome of a breeding program is a long-term product and requires careful design. In this paper, we present a stochastic multi-objective multi-stage decision model (SMOM) that optimises simultaneously a breeding program over a number of generations. It is found that the complexity of the model is very high $O(2^{T\times(S+D)^2})$ for the search space, where S denotes the number of sires, D the number of dams, and T the number of generations. Two research directions are suggested to make the problem tractable, namely; genetic algorithms and constraint satisfaction techniques.

1 Introduction

Dairy cattle are one of the main agriculture resources in Australia. In 1995-96, there were 13,888 dairy farms[1] in Australia with 1.924 million dairy cows [2] and a total milk production of 8,716 million litres with a value of \$A3 billion. During the last decade, there has been a focus on applying successful dairy breeding programs to increase the outcome of the dairy industry.

In this paper, a description of a stochastic multi-objective multi-stage decision model (SMOM) for a multiple generation mate-selection problem is introduced. A neural network based prediction system [?] is used as a back-end to the model to supply information about the expected progeny performance. This model overcomes the representational drawbacks in previous models.

Breeding programming, is the term used throughout this paper to delineate the problem of planning a breeding program on the level of a farm reckoning with the strategic objectives of the economy.

Dekkers [?] clearly summarised the current state of the art in designing breeding programs. He suggested that the four main components of a breeding program are the formulation of a breeding goal, methods for evaluation of potential breeding stock, methods for selection and mating strategies. He claims that the crucial objectives for designing breeding programs are the rates of genetic progress and inbreeding,

[1] Source: Australian Dairy Corporation ADC.

[2] Source: Australia Bureau of Statistics

though the overall goal is an economic one and the conditions for most studies on evaluation and optimisation of breeding programs are not realistic.

Kinghorn [?] established three essential decision areas for animal breeding:

- The amount of recording effort for traits and pedigree information for each animal.
- The selection of animals allowed to act as parents.
- The optimal combination for mating selected animals.

He mainly merged the determination of a breeding objective discussed by Dekkers into the selection and the allocation processes. The first decision is a cost-benefit analysis which balances the recording effort and the accuracy required. We do not discuss this problem further here. The second decision concerns *selection* and has been covered extensively in the literature for a single generation. The third decision concerns *allocation* where dams (productive cows) are allocated to sires (productive bulls). The last two problems are combined together and called the *mate-selection* problem. This paper proposes a *Stochastic Multi-Objective Multi-stage* (SMOM) model for mate-selection. The problem is formulated so as to optimise selection and allocation simultaneously over a number of generations. Thus, it overcomes most of the problems in the models previously presented in the literature because it guarantees global optimality over both problems rather than returning partially optimal solutions as previous models do.

One of the main objectives in designing a breeding program is maximisation of genetic improvement in a herd. There are four main factors that control genetic improvement [?] namely; variation, intensity of selection and accuracy of selection and generation interval. The genetic gain per year is generally measured by the following equation :

$$Genetic\,gain/year = \frac{Genetic\,variation \times Intensity of selection \times Accuracy}{Generation\,interval}, \tag{1}$$

where the *accuracy* of predicting genetic value equals the correlation coefficient between predicted and true genetic values of an animal. *Intensity* is a relative measure of how much the average of the population after selecting and mating bulls exceeds the average of the population before selection. The *genetic variation* can be measured by the population standard deviation of genetic merit and it has an indirect relation with the amount of inbreeding in the population. The *generation interval* is the average time between the birth of an animal and that of its replacement. This is usually difficult to control since the bulls and the cows need to wait to be proven when their progeny become mature. However, with new technologies such as Embryo transfer and cloning [?], one can control the generation interval. Nevertheless, these technologies are new and still expensive. For a bull, the generation interval ranges from 5 to 9 years, whereas for a dam, it ranges from 5 to 6 years.

In general, there are three decisions to be addressed for cow selection [?]:

- Selection for culling.

- Selection for breeding to obtain sons for herd use.
- Selection for breeding to obtain replacement heifers.

If the breeding program is designed for the short term, selection should be based on the real producing ability (*i.e.* 50% of the difference between the animal production record and the herd average). If the program is designed for the long term, selection should be based on genetic values. In practice, a combination of both short and long term objectives is needed.

The problem is represented using a SMOM model. If the stochastic and the multi-objective components are ignored and selection carried out in advance, the problem is reduced to a dynamic transportation model where we suppose the bull can be allocated to multiple dams. If the bull is allowed only to mate with at most one dam in a single generation, the problem is reduced to a dynamic assignment problem. To formulate the bull multi-objective problem as a multi-stage model, we have to create variables representing animals before the start of solution process even though animals are actually born in the future generations incorporated into the model. Moreover, we do not know in advance whether a mating will result in a son or a daughter and this needs to be represented stochastically. Also eventually, some of the animals may be culled and thus cease to exist in later generations. The model must also allow children born during the period modelled to be unavailable for allocation until becoming mature and thus productive. Each animal is represented by a time-dependent binary vector component which takes the value one only when the animal it represents is alive. The difference between the conventional transportation model and the one inherited in the mate-selection model and presented in this paper, is that source and destination nodes in the conventional model are known and are available for allocation in advance. In the latter model, these nodes are non-deterministic. The way non-determinism is handled in the model by attaching to these nodes binary state variables that determine animal status (yet to be born, dead, or alive) and reproductive behaviour in the generation.

In order to substantiate the results of this study, it will be interesting to survey the literature on the mate-selection problem in animal breeding especially in dairy cattle. In Section 2, the literature review is presented while the assumptions of the model are discussed in Section 3 and the advantages of the model in Section 3. In Section 3, the model is discussed in detail and issues of complexity covered in Section 4. Proposed methods for solving the model are discussed in Section 5 then conclusions are drawn in Section 6.

2 Background

Initially, research in breeding decisions focused on the problem of selecting bulls and cows in a herd for mating. The methodology for selection is to define either a *selection index* (for long term planing) or *profit index* (for short term planing) [?]. The former expresses the likely merit of a cow's progeny thereby indicating its worth as a breeding animal. For the latter, the quality of milk produced with respect to measures such as fat and protein contents and milk volume represent the worth of a production animal. A combination of these two methodologies compromises between long-term and short-term planning. A set of *economic weights* is employed to express preferences among different animal traits. For the profit index, the economic weights in the selection index are replaced by the actual profit represented by a profit function. In this section, some key works for the selection and mate-selection problems are presented.

2.1 Selection

Beard [?] developed a linear breeding objective for the additive genetic merit for lactation yield of the three components of milk (*i.e.* fat, protein and carrier). The economic weight for a trait is calculated as the difference between the change in returns and change in costs due to a one-unit increase in the trait. With a linear objective function, this rate of change is valid over the whole domain of each variable. The environmental effects are not considered in the work. The resulting economic weights for fat, protein and carrier were $0.8695, $1.1162 and -$0.0294 per kg, respectively. It is clear that these economics weights are different from the markets' prices.

Quinton [?] presented a very useful comparison between different selection methods taking into consideration inbreeding side-effects. He concludes that neither Best Linear Unbiased Predictor (BLUP) nor unrestricted selection based on estimated breeding values may always be optimal. Moreover, he queries the current theory of selection methods and highlights the need for a revision of the methodology and reappraisal of the results of selection theory and optimisation on net genetic response rates.

The Victorian Department of Agriculture [?] developed a computer program (Selectabull) for selecting the most suitable set of bulls for a farmer's herd and ranking them according to the predictive profitability of each. The profitability is determined according to financial information submitted by the farmer. Four scenarios are implemented in Selectabull namely; Manufacturing, Liquid/Quota, Workability, and Altered prices. In each scenario, a selection index is formulated and bulls ranked accordingly. The program also allows for different payment systems in operation in Australia. As an example, the breeding objective for Victoria is calculated as $\$A - 0.02 * Milk + \$A1.2 * Fat + \$A3.9 * Protein + \$A4.9 * Survival + \$A1.1 * Milkingspeed + \$A1.8 * Temperament - \$A0.7 * Mature\_body\_weight$. The weights in the objective function are calculated according to the farmer's herd data, milk payment system and the farmers own value judgement if desired. The program is already in use, though from the author's point of view it has a number of limitations :

- It ranks bulls, according to the selected scenario for a breeding objective, without considering the significance of the ranking. It was noted by [?] that although the economic weights used for each scenario differed substantially, the correlations between the selection indices were very high. Also, the top ten bulls for the Liquid/Quota and Workability scenarios were the same though ranked differently.

- It chooses the bulls for the whole herd without considering individual characteristics of the cows in the herd (i.e. it does not recommend any allocation of sires to dams).
- It does not capture important genetic principles such as inbreeding.

In short, Selectabull is insufficient for handling the full selection-allocation problem.

A system developed by the Machine Learning Research Center at the Queensland University of Technology called Connectibull [?] showed good results. It was designed around a neural network specific for each bull in the system. It has a number of limitations.

- It does not allow for sire data to be represented explicitly in the input. In Connectibull, contributions of the sire to daughter performance are embedded in the set of weights of the artificial neural network trained on production data of sire progeny. This could be a major flaw as the traits of the daughter come from both the sire and the dam. Moreover, it increases the learning task because the network should not only learn the relation between the inputs and outputs (dam traits and its daughter) but also should learn the unseen sire information. In addition, because of the previous argument, the interpretation of the weights became more difficult as they represent different concepts (the correlation between the dam and the daughter as well as the contribution of the sire in the daughter traits).
- It allows environmental data to be incorporated only indirectly through the inclusion of herd mean production variables.
- Its employment of a single neural network for each sire does not allow for network generalisation of the effects of sire traits nor the general relation between the sires, dams and environment and the progenies.
- It ranks sires for mating with a dam only in terms of predicted production traits of potential daughters (*i.e.* it ignores other factors such as inbreeding).
- In its present form, its neural networks are trained on production figures without correcting for age at lactation or season of calving.

Howarth [?] discusses selection in terms of multiple objectives. He studies two strategies with the farmer optimising two different objectives. In the first strategy (he calls it an SPL strategy), the herd is divided into two sub-herds and each is selected for one of the objectives. In the second strategy, called AVG, the average of the two objectives is calculated and the selection is carried out a single composite objective. The study suggests using the AVG strategy if the correlations between objectives are positive and the time horizon to be considered is short-term, or if the size of the herd limits the usefulness of the SPL strategy. Moreover, SPL generates the greatest response in the short-term, but it does this only if the correlations between the objectives are negative. Actually, this result is not new in multi-objective optimisation. the negative correlation means simply that there is a conflict between the two objectives. Accordingly, taking the average of the two objectives to obtain a single objective does not guarantee the maximum output unless the set of Pareto-optimal solutions are generated. When the correlation is positive, it means that there is no conflict between the objectives (*i.e.* the maximisation of one objective will maximise the other one). In this case, maximising the average of the two is a reasonable approximation that will result in good solutions.

Hirooka [?] illustrated the use of five linear programming based strategies for sire selection. These are *unrestricted* selection, *zero-gain restricted* selection, *non-zero gain restricted* selection, *proportional restricted* selection and desired gain selection. In unrestricted selection, selection is done based on the selection index. In the zero and non-zero gain restricted selections, selection is carried out to maximise the selection index as well as to maintain the level of a desired trait. For zero-gain, the goal would be to equalise the level of a specific trait within the population. For non-zero gain, the goal would be to maintain a certain level of the difference between the value of sire trait and the mean value for this trait within the set of sires. In proportional restricted selection, the constraint is to maintain the ratio of sires traits to a certain level. For desired gain selection, selection can be done using zero or non-zero gain restricted selection or proportional selection with the relaxation of the constraint by altering from equality to inequality.

2.2 Mate-selection

In 1985, the first use of optimisation methods in the dairy industry appears to have been suggested by Jansen [?] with a conventional transportation model for selecting mating pairs. With this model, it is very simple to incorporate different aspects of a real life problem such as inbreeding and mate-selection over a number of generations. however, it cannot handle a nonlinear genetic objective.

Kinghorn [?] compares four mate-selection strategies, three of which are heuristic while the fourth employs linear programming technique. He found that the time for the linear programming calculation increases dramatically with increased population size. The heuristic of sequential allocation of the best pair from remaining candidates is found to be close to optimal in most cases; though the study is limited to a single generation. However, linear programming computerised solvers have improved dramatically during the last decade and this would change the conclusion of the study. For example, until 1987, Linear programming solvers depended on the Simplex method which is an exponential algorithm. In 1987, Karmarkar introduced interior point methods which are polynomial-time algorithms.

Klieve [?] proposed an approximate quadratic programming model for mate-selection. Portofolio analysis is used to construct a quadratic objective for the problem. Additive genetic merit is used as a measure of the gain (profit) and inbreeding is used as a measure of the risk associated with differing levels of merit. Stochastic simulation was used to evaluate different selection strategies with respect to both additive genetic response and inbreeding. The proposed quadratic model resulted in about a 50% reduction in inbreeding over the random mating strategy.

In recent work, Kinghorn [?] proposes animal breeding clustering to facilitate group mate selection together with an individual-animal mate selection step for fine-tuning. Cluster analysis is used to form groups of animals within each sex based on nominated factors such as breed, herd, age, and EBV. Mate selection is then performed on groups. He wanted to select 20 sires and 500 dams for breeding from 816 sires and 1784 dams across 3 breeds. The study reported a reduction in computation from 394.2 seconds for full mate selection to 0.28 seconds for group mate selection though the necessary cluster analysis took an additional 162 seconds. In some respects, computation time need not be a major consideration in breeding programs. Building a breeding program that would be implemented over a year or more would not be impractical even if a week of computation time is involved. The maximum number of clusters was chosen arbitrarily as four times the square root of the number of candidates available. For 100 candidates, the maximum number of clusters is 40. Hence, it is clear that the method should be practical even for herds of a few hundred animals. However, the work is certainly important for reducing the complexity of the search space.

Shepherd [?] proposed a binary nonlinear programming model for mate-selection over two generations with cross-breeding. The model accounts for two groups of predicted progeny, those to be used for commercial purposes and those for selection and mating. The breeding objective is to maximise the combined return from commercial animals for both one and two generations into the future. The model is limited since it is very difficult to generalise over more than two generations and it does not account for the stochastic nature of the gender of the progeny.

3 The model

Mathematical programming is the branch of optimisation concerned with representing problems using mathematical models taking the form:-

$$\text{Maximise or Minimise } \mathcal{F}(\S).$$
$$\text{S.T. } \mathcal{G}(\S) \leq \prime.$$

where $\mathcal{F}(\S)$ is the set of objective functions and $\mathcal{G}(\S)$ is the set of constraints. *Multi-stage decision modelling* is a modelling approach rather than a solution mechanism (eg. *linear programming*). Thus, it is an art for decomposing large scale problems in terms of stages linked with a set of transition functions. The model can be solved afterward with *dynamic programming* [?] or other mathematical programming techniques. A multi-stage decision model is defined as [?]:

A collection (N, S, D, T, S_1, g) where

1. N is a positive integer specifying the number of *decision stages* comprising the decision-making process. Define $\mathcal{N} := \infty, \in, \ni, \ldots, \mathcal{N}$. Thus, for $N = \infty$ the set $\mathcal{N}$ denotes the set of positive integers.

2. S is a non-empty set entitled the *state space*. Its elements are called *states*.

3. D is a function that to each pair $(n, s) \in \mathcal{N} \times \mathcal{S}$, it assigns a subset of some set $\mathcal{D}$. The set $D(n, s)$ is referred to as the *set of feasible decisions* pertaining to state s at stage n. If the set $D(n, s)$ is empty, the state s is considered to be non-feasible at stage n. The set $\mathcal{D}$ is named the *decision space*.

4. T is a function on $\mathcal{N} \times \mathcal{S} \times \mathcal{D}$ with values in S called the *transition function*. For any triplet (n, s, x) such that $n \in \mathcal{N}$, $s \in S$, and $x \in D(n, s)$, the element $s = T(n, s, x)$ designates the states at stage $n + 1$ resulting from the application of the decision x to the state s at stage n.

5. S_1 is a non-empty subset of S whose elements are named *initial states*.

6. g is a real-valued function on $S_1 \times \mathcal{D}^{\mathcal{N}}$ termed the *objective function*. The value yielded by $g(s, x_1, x_2, \ldots, x_N)$ is understood to specify the overall cost or benefit arising from the decisions $(x_1, \ldots, x_N)$ made at stages $1, 2, 3, \ldots, N$ respectively, given that the initial state of the process is $s \in S_1$.

Assumptions of the SMOM model

The assumptions of the proposed stochastic multi-objective multi-stage decision model are :-

1. There are no overlapping generations. By this, we assume that every event belongs only to a single generation. This is not a severe limitation as the generation interval may be chosen small enough not to be limitation.

2. Every animal has a unique index (position in state vectors) that it carries through all generations whether or not it is yet to be born, alive, or dead.

3. Each mating results in a subsequent birth. The gender of the offspring is determined stochastically.

4. Cross-breeding, between breeds, is not considered in SMOM.

The advantages of the model

The advantages of the proposed SMOM model are :

1. It accounts for different age groups among sires and dams.

2. There is no limit on the number of generations apart from the requirements of computer storage.

3. It accounts for both the short-term (real producing abilities) and long-term (real genetic gains) dairy breeding objective.

4. It models environmental and managerial effects as well as genetic effects by considering the season of calving and the average herd production figures while training the neural network [?] for predicting the progeny merit used in the objective function.

3.1 Model parameters and decision variables

Each generation is equivalent to a stage in the model. The existence of productive animals, cows and bulls, is represented using the state variables SD and SP respectively. The existence of unproductive animals, one-year old cows and bulls and two-year old cows, is represented using the state variables $D1$, $S1$, and $D2$ respectively.

The introduction of progeny into the herd is represented relative to their parents and genders. These are four state variables, DD, SD, DS, and SS for the female child from her mother and father and the male child from his mother and father respectively. The availability of an animal, cow or bull, for mating is represented using two state variables, SD and SP respectively.

The transitions are done using the logical constraints A, ..., H. The goal state is the optimal (best) herd composition from the first to the last generation. Finally, the decision in each stage is the determination of whether an animal should be culled or mated. Moreover, in the case of mating, which animal from the other sex should be involved in the mating. The variable M_{ij} represents these types of decisions between sire i and dam j. If the index of i is that of a special pseudo-sire $c(t)$, it means that the j^{th} dam is allocated for culling. The same apply for the sire when $j = c(t)$ and the bull is paired with a special pseudo-dam.

Table **??** defines the model parameters whereas the decision variables are defined in Table **??**. The generations are denoted by $t = 1, 2, \ldots, T$, where T is the total number of generations being modelled. We adopt the term "bull" generally for any male animals in the herd especially for young immature ones. We reserve the term "sire" specifically for male animals that have been mated. Thus $S1$ defines young bulls and SP defines productive bulls some of which may actually be sires already. The vector $X(t)$ identifies those bulls that have been mated in generation t and so we may more properly call them sires. Similar remarks apply to the vector $Y(t)$ which identifies for cows (corresponding to bulls) and dams (corresponding to sires).

If we assume that the model will be solved using *Simulation* techniques, at the start of simulation unit t, all bulls and cows that are productive are reflected by the vectors SP and DP. Thus if bull i is mature (that is, was conceived in or before time $t - 3$, and so was born in or before generation $t - 2$ and so is older than one unit of simulation time, may or may not have offspring but has not been culled) then

$SP_i(t) = 1$. If at time t, it is selected for mating then the i^{th} component of X, $X_i(t)$ is set to 1, otherwise $X_i(t)$ is set to 0. At simulation time $t+1$, $SP_i(t+1)$ is set to 1 if $S1_i(t) = 1$ (that is, the bull has just become productive at $t+1$ and will be now available for mating) or $X_i(t) = 1$ (that is, the bull was productive at time t and was selected for mating). Accordingly, the variable $X(t)$ is a transition variable used to hold components for the purpose of helping update the new set of productive bulls $SP(t+1)$. In a strict mathematical sense the variable X is superfluous, however, it is very convenient for purposes of defining the constructive and programmatice solution. Similar reasons apply to the corresponding transition vector Y, used for cows.

The list of parameters and constants used in the model

$S(t)$:	Number of sires required in generation t
$D(t)$:	Number of dams required in generation t
T:	Total number of generations
S:	Total number of sires
D:	Total number of dams
$N_{mate}(t)$:	The number of matings needed at generation t
$S_{max}(t)$:	The maximum number of matings per sire at generation t
$D_{max}(t)$:	The maximum number of matings per dam at generation t
$c(t)$:	Dummy variable to be allocated to any culled animal that is, an animal that is not selected for mating, and therefore is to be culled, is assigned a pseudo-mating to c(t).
$S_{cull}(t)$:	Number of sires to be culled at generation t
$D_{cull}(t)$:	Number of dams to be culled at generation t

The list of decision variables used in the model

$S1_i(t)$:	Bull number i is one year old (and unproductive) in generation t
$SP_i(t)$:	Bull number i is productive ($age \geq 2$) in generation t
$D1_j(t)$:	Cow number j is one year old (and unproductive) in generation t
$D2_j(t)$:	Cow number j is two year old (and unproductive) in generation t
$DP_j(t)$:	Cow number j is productive ($age \geq 3$) in generation t
$X_i(t)$:	A binary variable set to 1 if bull (productive male) i exists in generation t and is not culled in generation t; otherwise 0
$Y_j(t)$:	A binary variable set to 1 if cow (productive female) j exists in generation t and is not culled in generation t; otherwise 0
$M_{ij}(t)$:	A binary variable set to 1 if the i^{th} sire is mated with the j^{th} cow in generation t; otherwise 0 An extra row and column is added to represent culling.
$DD_{ij}(t+1)$:	The density function of a Bernoulli distribution with probability $DD_{ij}(t+1) = f(x) = \left\vert \begin{array}{ll} 1, & if\, the\, stochastic\, variable\, x \geq 0.5 \\ 0, & otherwise \end{array} \right. .$ If $DD_{ij}(t+1)$ is 1 then at $t+1$ dam i become the mother of cow j, resulting from a mating at t. x is a random variable.
$DS_{ij}(t+1)$:	The density function of a Bernoulli distribution with probability $DS_{ij}(t+1) = f(x) = \left\vert \begin{array}{ll} 1, & if\, the\, stochastic\, variable\, x \geq 0.5 \\ 0, & otherwise \end{array} \right. .$ If $DS_{ij}(t+1)$ is 1 then at $t+1$ dam i become the mother of bull j, resulting from a mating at t. x is a random variable.
$SD_{ij}(t+1)$:	The density function of a Bernoulli distribution with probability $SD_{ij}(t+1) = f(x) = \left\vert \begin{array}{ll} 1, & if\, the\, stochastic\, variable\, x \geq 0.5 \\ 0, & otherwise \end{array} \right. .$ If $SD_{ij}(t+1)$ is 1 then at $t+1$ sire i become the father of cow, resulting from a mating at t. x is a random variable.
$SS_{ij}(t+1)$:	The density function of a Bernoulli distribution with probability $SS_{ij}(t+1) = f(x) = \left\vert \begin{array}{ll} 1, & if\, the\, stochastic\, variable\, x \geq 0.5 \\ 0, & otherwise \end{array} \right. .$ If $SS_{ij}(t+1)$ is 1 then at $t+1$ sire i become the father of bull j, resulting from a mating at t. x is a random variable.

3.2 Objective functions

A typical objective to be maximised in a breeding program is the genetic gain. Kinghorn [?] finds that mate-selection has no value when the genetic traits are additive (that is the expected genetic value of the daughter is half the sum of the genetic value of the parents). In such a case, the selection index theory will result in the optimal solution; that is, mate the best bulls with the highest genetic values with the best cows randomly. The mate-selection problem can be useful when there is an interaction (dominance or epistasis) between the parents, a nonlinear genetic trait, an inbreeding, and/or crossbreeding.

In our design for the breeding objectives, we have considered a compromise between short and long term objectives. With short term objectives, we mean those criteria which we shall contribute during the short run (the following year or two). For example, culling cows with low phenotypic production value will increase the average production per cow within the farm. We considered three types of short term objectives; the profit of not culling an animal in the current period represented with its expected production profit next year, the profit of culling an animal represented with the average weight of an animal time the profit of 1kg weight, and the short term inbreeding depression represented with the amount of inbreeding in the second year multiplied by the cost of 1 unit inbreeding depression. These three objectives can be formulated as follows:

$$\uparrow \sum_t (sum_i b_i \times P_{ij}(t))\mathbf{Y}^t(t), \quad t = 1 \ldots T. \tag{2}$$

$$\uparrow \sum_t C_{culling}\ times(\sum_i M_{i,c(t)} + \sum_j M_{c(t),j}), \quad t = 1 \ldots T. \tag{3}$$

$$\downarrow \sum_t C_{inbreeding}[\mathbf{X}^t(t), \mathbf{Y}^t(t)]\mathbf{F}, \quad t = 1 \ldots T. \tag{4}$$

The set of objective functions of the breeding program will be

$$\uparrow \sum_t \mathbf{Z^t(t)G}, \quad t = 1 \ldots T. \tag{5}$$

$$\uparrow \sum_t \mathbf{Z^t(t)E}, \quad t = 1 \ldots T. \tag{6}$$

$$\downarrow \sum_t \mathbf{Z^t(t)B}, \quad t = 1 \ldots T. \tag{7}$$

$$\downarrow \sum_t \mathbf{M_s^t(t)C_s}, \quad t = 1 \ldots T. \tag{8}$$

$$\uparrow \sum_t N_{cull(t)}P, \quad t = 1 \ldots T. \tag{9}$$

where $\uparrow$ and $\downarrow$ indicate maximise and minimise respectively, all these are in the units of dollars, where,

- $\mathbf{Z(t)}$. A column vector of animals formed by the concatenation of X and Y that is,

$$\mathbf{Z(t)} = (X_i(t), Y_j(t))^t \tag{10}$$

- $\mathbf{M_s(t)}$. A column vector whose i^{th} element is the number of mating of sire i in generation t, that is,

$$M_{s_i}(t) = \sum_j M_{ij}(t) \tag{11}$$

- $N_{cull}(t)$. A scalar denoting the total number of animals culled in generation t, that is,

$$N_{cull}(t) = \sum_i M_{i,c(t)} + \sum_j M_{c(t),j} \tag{12}$$

- **G**. A column vector of the genetic merits of each animal (selection index). The j^{th} element G_j, takes the form $\sum_i W_i \times h_i \times T_{ij}$ where W_i is the economic weight of the i^{th} trait, h_i is the heritability of the i^{th} trait, and T_{ij} is the real producing ability of j^{th} animal for the i^{th} trait[3] (for example the i^{th} trait might be the number of litres of milk).

- **E**. A column vector of the production traits of each animal (profit index).

$$E_j = \sum_i PT_i \times T_{ij} \times (1 - h_i) \tag{13}$$

  where PT_i is the price of the production trait i, for example; i=milk, fat, protein.

- **B** the cost of inbreeding given by,

$$\mathbf{B} = C_I \times \mathbf{I_C} \tag{14}$$

  where $\mathbf{I_C}$ represents the inbreeding coefficients calculated using the algorithm provided in [?] and C_I is a scalar representing the cost of one unit inbreeding coefficient.

- $\mathbf{C_s}$. A vector of semen costs for the bulls in the system.

- P. The price of culling an animal (usually constant).

The problem can also be formulated as a single objective stochastic multi-stage decision model by introducing a set of preferential weights, $\alpha, \beta, \gamma, \delta$, and η. In this case, the objective function will be,

$$Max. \sum_t \alpha \mathbf{Z^t(t)G} + \beta \mathbf{Z^t(t)E} + \gamma \mathbf{Z^t(t)B} + \delta \mathbf{M_s^t(t)C_s} + \eta N_{cull(t)} P \tag{15}$$

3.3 Description of logical constraints

The SMOM involves a considerable number of logical constraints to be satisfied by the state variables listed in Table **??**. These constraints fall into categories, we call a, b, ..., h, with the constraints within a category being somewhat similar. If we consider state variables in their vectors or matrix form (for example, $X = (X_1, X_2, ...)$), then each constraint affects a number of state variables that tend to occur in pairs. As an aid in visualising the constraints, we have presented in Figure 1 a graph with nodes labelled with the state variables and arcs labelled with constraint categories. The arcs connect pairs of variables (nodes) are labelled with the relevant constraint category. For example, the arc labelled "a" joining nodes M (the mating matrix) and DD (the mother-daughter relationship matrix) refers to a constraint between M and DD that is similar to constraints between M and DS, SS and SD which are also labelled "a" in the

[3] Recall that the estimated genetic value = heritability $\times$ real producing ability [?].

figure. As an aid in understanding arrows arc employed to more-or-less suggest causation. For example, the matrix DD follows as a result of a mating via the matrix M producing a female offspring. We have thus directed the arc from M to DD.

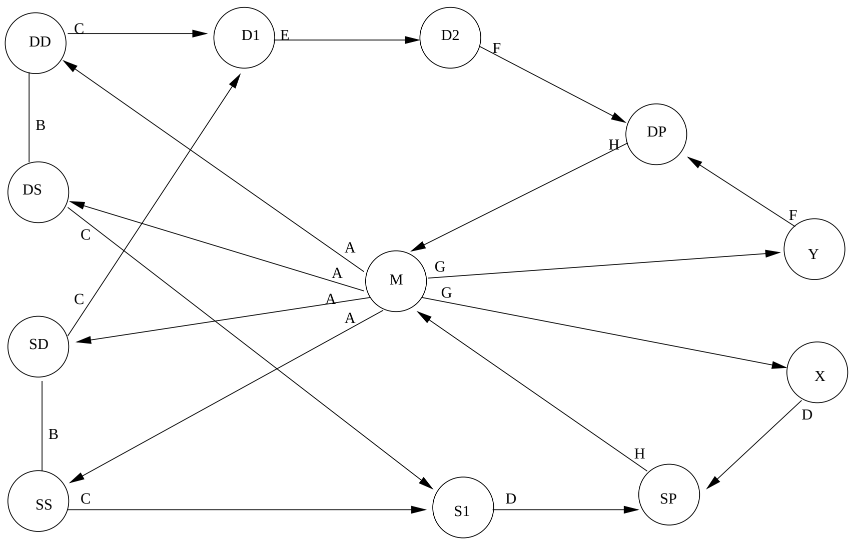

Figure 1: The logical relationships between the model's variables

3.4 Constraints

1. The number of matings for each sire (dam) is not to exceed the maximum number of allowed matings for a sire (dam) in generation t (the matings take place between vectors $SP(t)$ (productive sires) and $DP(t)$ (productive dams).

$$\sum_{i \neq c(t)} M_{ij}(t) \leq D_{max}(t), \quad j = 1 \ldots D, t = 1 \ldots T. \tag{16}$$

$$\sum_{j \neq c(t)} M_{ij}(t) \leq S_{max}(t), \quad i = 1 \ldots S, t = 1 \ldots T. \tag{17}$$

2. The total number of allocations should equal the total number of matings required.

$$\sum_{i \neq c(t)} \sum_{j \neq c(t)} M_{ij}(t) = N_{mat}(t), \quad t = 1 \ldots T. \tag{18}$$

3. The total number of bulls/cows allocated for culling should equal the total number of culling required.

$$\sum_j M_{c(t),j}(t) = D_{cull}(t), \quad t = 1 \dots T. \tag{19}$$

$$\sum_i M_{i,c(t)}(t) = S_{cull}(t), \quad t = 1 \dots T. \tag{20}$$

4. The total number of bulls/cows, that are alive in generation t and are not culled should equal the difference between the total number of bulls (cows) required in the generation minus the total number of culled bulls (cows).

$$\sum_i X_i(t) = S(t) - S_{cull}(t), \quad t = 1 \dots T. \tag{21}$$

$$\sum_j Y_j(t) = D(t) - D_{cull}(t), \quad t = 1 \dots T. \tag{22}$$

5. A sire or dam that is allocated for mating cannot be allocated for culling in the same generations.

$$M_{i,c(t)}(t) \times \sum_{j \neq c(t)} M_{ij}(t) = 0, \quad i = 1 \dots S, t = 1 \dots T. \tag{23}$$

$$M_{c(t),j}(t) \times \sum_{i \neq c(t)} M_{ij}(t) = 0, \quad j = 1 \dots D, t = 1 \dots T. \tag{24}$$

6. Logical constraints preventing the mating between the two special culling nodes in each generation.

$$M_{c(t),c(t)}(t) = 0, \quad t = 1 \dots T. \tag{25}$$

7. Each sire (dam) mated at generation t has a number of children in generation $(t+1)$ equal to the number of its matings. (Group A)

$$\sum_{i \neq c(t)} M_{ij}(t) - \sum_k DD_{jk}(t+1) - \sum_l DS_{jl}(t+1) = 0, \quad j = 1 \dots D, t = 1 \dots (T-1). \tag{26}$$

$$\sum_{j \neq c(t)} M_{ij}(t) - \sum_k SD_{ik}(t+1) - \sum_l SS_{il}(t+1) = 0, \quad i = 1 \dots S, t = 1 \dots (T-1). \tag{27}$$

8. If a mating M_{ij} occurred, the gender of the subsequent offspring in the next generation for both the sire and the dam is the same. (Group A)

$$M_{ij}(t) - \sum_k DD_{jk}(t+1) \times SD_{ik}(t+1) - \sum_l DS_{jl}(t+1) \times SS_{il}(t+1) = 0 \tag{28}$$

$$, \quad i = 1 \dots S, j = 1 \dots D, t = 1 \dots T. \tag{29}$$

9. For all sires in each generation, the number of male children equals the number of female children. Here it is required that the actual number of males and females offspring equal the expected number. However, to count for the case where the total number of children is an odd number, it is required that the difference between the number of males and females children to be less than or equal to one. (Group B)

$$\sum_i \sum_k SD_{ik}(t+1) - \sum_i \sum_l SS_{il}(t+1) \leq 1, \quad t = 1 \ldots (T-1). \tag{30}$$

10. Children conceived in the current generation (t) will be one-year-old unproductive individuals in the next generation ($t+1$) and each child has precisely one mother and one father. (Group C)

   - Each male child has one mother.

$$\sum_l DS_{li}(t) = S1_i(t+1), \quad i = 1 \ldots S, t = 1 \ldots (T-1). \tag{31}$$

   - A one-year-old male child in generation t cannot be a new child of any dam in generation t.

$$S1_i(t) + \sum_l DS_{li}(t) \leq 1, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{32}$$

   - A productive male child in generation t cannot be a new child of any dam in generation t.

$$SP_i(t) + \sum_l DS_{li}(t) \leq 1, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{33}$$

   - Each male child has one father.

$$\sum_l SS_{li}(t) = S1_i(t+1), \quad i = 1 \ldots S, t = 1 \ldots (T-1). \tag{34}$$

   - A one-year-old male child in generation t cannot be a new child of any sire in generation t.

$$S1_i(t) + \sum_l SS_{li}(t) \leq 1, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{35}$$

   - Each female child has one mother.

$$\sum_l DD_{lj}(t) = D1_j(t+1), \quad j = 1 \ldots D, t = 1 \ldots (T-1). \tag{36}$$

   - A one-year-old female child in generation t cannot be a new child of any dam in generation t.

$$D1_j(t) + \sum_l DD_{lj}(t) \leq 1, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{37}$$

- A productive female child in generation t cannot be a new child of any dam in generation t.

$$DP_j(t) + \sum_l DD_{lj}(t) \leq 1, \quad j = 1 \dots D, t = 1 \dots T. \tag{38}$$

- Each female child has one father.

$$\sum_l SD_{lj}(t) = D1_j(t+1), \quad j = 1 \dots D, t = 1 \dots (T-1). \tag{39}$$

- A one-year-old female child in generation t cannot be a new child of any sire in generation t.

$$D1_j(t) + \sum_l SD_{lj}(t) \leq 1, \quad j = 1 \dots D, t = 1 \dots T. \tag{40}$$

11. Any productive bull in the next generation is either a one year old bull in the current generation or it is a productive bull in the current generation and *has not been culled*. This also forces a bull to be either productive or unproductive. (Group D)

$$SP_i(t+1) = S1_i(t) + X_i(t), \quad i = 1 \dots S, t = 1 \dots (T-1). \tag{41}$$

$$SP_i(t) + S1_i(t) \leq 1, \quad i = 1 \dots S, t = 1 \dots T. \tag{42}$$

12. One-year-old (unproductive) cows in the current generation will be two-year-old (unproductive) cows in the next generation. (Group E)

$$D2_j(t+1) = D1_j(t), \quad j = 1 \dots D, t = 1 \dots (T-1). \tag{43}$$

$$D2_j(t) + D1_j(t) \leq 1, \quad j = 1 \dots D, t = 1 \dots T. \tag{44}$$

13. A productive cow in the next generation is either a two year old dam in the current generation or it is a productive cow in the current generation and has not been culled. Also logically, a cow can either be productive or unproductive but not both. (Group F)

$$DP_j(t+1) = D2_j(t) + Y_j(t), \quad j = 1 \dots D, t = 1 \dots (T-1). \tag{45}$$

$$DP_j(t) + D2_j(t) \leq 1, \quad j = 1 \dots D, t = 1 \dots T. \tag{46}$$

$$DP_j(t) + D1_j(t) \leq 1, \quad j = 1 \dots D, t = 1 \dots T. \tag{47}$$

14. A flag constrains $X_i(t)$ to refer to sires that are not culled in generation t. If sire i had a real mating in generation t then we must force $X_i(t)$ to be 1. Similar results apply to $Y_j(t)$ for dams. (Group G)

$$(1 - X_i(t)) \times \sum_{j \neq c(t)} M_{ij}(t) = 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{48}$$

$$-X_i(t) + \sum_{j \neq c(t)} M_{ij}(t) \geq 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{49}$$

$$(1 - Y_j(t)) \times \sum_{i \neq c(t)} M_{ij}(t) = 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{50}$$

$$-Y_j(t) + \sum_{i \neq c(t)} M_{ij}(t) \geq 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{51}$$

15. The bulls/cows are available for allocation in generation t if and only if they are productive in generation t. In the first constraint, if bull i is mated or culled, it must be productive and $SP_i(t)$ forced to 1. In the second constraint, if bull i is not mated or culled then it is unproductive and $SP_i(t)$ is forced to 0. The same interpretation applies for the subsequent two constraints for a cow. (Group H)

- If bull i is allocated then it must be productive.

$$(1 - SP_i(t)) \times \sum_{j} M_{ij}(t) = 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{52}$$

- If it is productive, then it must be allocated (even for culling).

$$-SP_i(t) + \sum_{j} M_{ij}(t) \geq 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{53}$$

- If cow i is allocated then it must be productive.

$$(1 - DP_j(t)) \times \sum_{i} M_{ij}(t) = 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{54}$$

- If it is productive, then it must be allocated (even for culling).

$$-DP_j(t) + \sum_{i} M_{ij}(t) \geq 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{55}$$

- The children in generation t must be guaranteed not to exist as individuals in previous generations (*i.e.* 0 ... $(t-1)$).

$$\sum_{k} DS_{ki}(t) \times \sum_{l=1}^{(t-1)} X_i(l) = 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{56}$$

$$\sum_k SS_{ki}(t) \times \sum_{l=1}^{(t-1)} X_i(l) = 0, \quad i = 1 \ldots S, t = 1 \ldots T. \tag{57}$$

$$\sum_k DD_{kj}(t) \times \sum_{l=1}^{(t-1)} Y_j(l) = 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{58}$$

$$\sum_k SD_{kj}(t) \times \sum_{l=1}^{(t-1)} Y_j(l) = 0, \quad j = 1 \ldots D, t = 1 \ldots T. \tag{59}$$

4 Complexity analysis

This model can be either solved as a stochastic multi-objective binary multi-stage decision model (SMOM) or expanded and solved as a stochastic multi-objective binary nonlinear model (SMOBNM).

In the former, the number of variables in each generation is $(D+S)^2 + 3 \times (D+S) + D \times S + D$ with up to $19 \times D + 16 \times S + D \times S + 7$ constraints. Also, there will be approximately $2 \times D \times S + D^2 + S^2 - 15 \times D - 13 \times S - 7$ degrees of freedom in each generation. The search space in each generation will be approximately $2^{(D+S)^2+3\times(D+S)+D\times S+D}$. For example, if we have a 100 dams and a 100 sires and we would like to optimise for 5 generations, there will be 50,700 variables in each generation with at most 13,507 constraints, 37,193 degrees of freedom and approximately search space size of 10^{15270}.

In the latter, the number of variables is $T \times ((D+S)^2 + 3 \times (D+S) + D \times S + D)$ with $19 \times D \times T + 16 \times S \times T + D \times S \times T + 7 \times T - 4 \times S - 5 \times D - 1$ constraints. Also, there will be $2 \times D \times S \times T + D^2 \times T + S^2 \times T + 5 \times D + 4 \times S + 1 - 15 \times D \times T - 13 \times S \times T - 7 \times T$ degrees of freedom. The search space will be $2^{T\times((D+S)^2+3\times(D+S)+D\times S+D)}$. In the previous example, there will be 253,500 variables in all generations with 66,634 constraints, 186,866 degrees of freedom and search space size of 10^{76350}.

5 Proposed methods for solutions

It is clear that the complexity of the model is extremely high to be solved by conventional optimisation techniques. We propose a hybrid method that combines *genetic algorithms* as a powerful paradigm for function optimisation and constraint satisfaction which provides strong algorithms for solving efficiently problems over finite domains. Moreover, the high complexity of the mathematical model will be reduced by using the mapping from mathematical relations into logical relations and restricting the production of solutions in genetic algorithms to feasible solutions only by the development of algorithms to repair the chromosomes. The representational issues play a significant rule in complexity reduction and this will cut the search space extensively. The proposed technique for solving the problem sketched here indicates that the problem can still be solved efficiently despite the explosion in the search space through efficient representation as well as efficient algorithms. Moreover, the use of constraint logic programming for decomposing the problem [**?**; **?**] should also prove fruitful.

The model can be highly simplified within the genetic algorithm representation by generating a strong bias in the search. The initial population is guaranteed in the preliminary design to be feasible and the feasibility of the logical constraints is preserved in a projection. Generating a solution should be quite fast though the efficiency of convergence to optimality remains to be demonstrated by experiments.

6 Conclusion

In this paper, a stochastic multi-objective multi-stage decision model has been formulated for mate-allocation problem. The model has very high complexity and conventional optimisation algorithms do not seem to be feasible for solving the problem. There are two approaches suggested to tackle this problem. First, a hybrid algorithm that combines genetic algorithms and constraint satisfaction. Second, the use of problem-specific knowledge to to decompose the problem while maintaining optimality. A constraint logic programming paradigm should be useful for the latter, since it is a powerful paradigm combining symbolic search mechanisms and non-symbolic algorithmic mechanisms. The problem can be classified as a stochastic constraint satisfaction problem which is an interesting class of problem. To our knowledge, this type of problem has not been discussed in the field of artificial intelligence.