Applied Soft Computing 12 (2012) 1483–1492

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A population dynamics model to describe gene frequencies in evolutionary algorithms

Maury Meirelles Gouvêa Jr.[a,*], Aluizio F.R. Araújo[b]

[a] Polytechnic Institute, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Brazil
[b] Center of Informatics, Federal University of Pernambuco, Recife, Brazil

ARTICLE INFO

Article history:
Received 9 July 2010
Received in revised form 18 June 2011
Accepted 15 January 2012
Available online 31 January 2012

Keywords:
Population dynamics
Evolutionary algorithms
Stochastic processes
Diversity

ABSTRACT

The performance of evolutionary algorithms (EAs) may heavily depend severely on a suitable choice of parameters such as mutation and crossover rates. Several methods to adjust those parameters have been developed in order to enhance EA performance. For this purpose, it is important to understand the EA dynamics, i.e., to appreciate the behavior of the population. Hence, this paper presents a new model of population dynamics to describe and predict the diversity in any particular generation. The formulation is based on selecting the probability density function of each individual. The population dynamics proposed is modeled for a generational population. The model was tested in several case studies of different population sizes. The results suggest that the prediction error decreases as the population size increases.

1. Introduction

Since their introduction in the 1950s, evolutionary algorithms (EAs) have been employed to solve problems in areas, such as, telecommunications [1,2], analysis, modeling, and optimization [3–5]. The performance of EAs may be enhanced if the choice of their parameters, such as mutation rate and crossover method, is suitable. The on-line adjustment of these parameters, called parameters control, tends to improve EA performance [6,7].

The adjustment of mutation [8,9] and crossover [10,11] parameters may be used to produce greater exploration at the beginning and an appropriate exploitation at the end of the evolutionary process. A scheme to determine the population size dynamically [12,13] is useful to change either the size of subpopulations or the size of the whole population. The control of the population diversity [14,3,15–18] is a crucial issue when attempting to solve complex multimodal problems, specially in dynamic environments, because a suitable diversity prevents early convergence to a specific region of the solution space.

In order to enhance EA performance, it is important to understand its dynamics. For example, the diversity control methods do not have precise information about the evolutionary process; normally, such methods know the error between the current and reference diversity [16,3,18]. There are some EA designs, such as that of Radcliffe [19]; however, these models do not consider EA dynamics. Alternatively, there are mathematical formulations of EAs dealing with their dynamics. Some examples of these approaches were proposed by Vose and Liepins [20], Stark and Spall [21], and Bennett and Shapiro [22]. Most of these models may not be used to predict EA dynamics because they need knowledge of the transition probability between generations.

This paper presents a new model for population dynamics that allows to predict the diversity at one generation to be predicted as a function of the current gene frequency and two evolutionary factor parameters: selection pressure and mutation rate. The formulation is based on the selection probability of each individual. The proposed population dynamics is modeled for a generational population. The model was tested in several case studies with four population sizes, from 10 to 400 individuals using three different environment types. The results suggest that the prediction error decreases as the population size increases. The proposed model can be applied to different evolutionary computation problems, such as parameter and diversity control, convergence rate and sensibility analysis.

The rest of this paper is organized as follows. Section 2 presents several ways to calculate the population diversity and the procedure used by the proposed model. Section 3 describes the proposed model of population dynamics. Section 5 presents the results of several experiments on diversity measurement to support our

* Corresponding author at: Pontifical Catholic University of Minas Gerais, Polytechnic Institute, Av. Dom José Gaspar 500, 30.535-901, Belo Horizonte, MG, Brazil.
E-mail address: maury@pucminas.br (M.M. Gouvêa Jr.).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2012.01.004

measurement method. Section 6 presents the experiments with three different fitness functions and it reports their results and analysis. Finally, Section 8 draws conclusions from the paper.

2. Measuring the population diversity

In order to describe the dynamics of a population it is necessary to measure its diversity. There are several ways to describe the diversity of a population [23,3], which measure the degree of similarity between objects or individuals of a group or population of any nature. For example, Rao [24] created a diversity function based on the probability distribution of a finite set of species. The diversity function of Rao uses the distance $d(s_1, s_2)$ between two species s_1 and s_2 defined over a finite set of species, as follows

$$\Gamma = \sum_{i=1}^{n_s}\sum_{j=1}^{n_s} p_i\, p_j\, d(s_i, s_j), \tag{1}$$

where n_s is the number of species and $p_i = P(X = s_i)$.

Solow et al. [25] proposed a function, named the preservation measure, to calculate the loss of diversity when a species s_i becomes extinct, as follows

$$\Delta\Gamma = -\sum_{s_i \notin S} d(s_i, S). \tag{2}$$

Based on Rao [24], Champely and Chessel [26] introduced a function for diversity using the Euclidean distance between species, defined as

$$\Gamma = \frac{1}{2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} p_i\, p_j\, [d(s_i, s_j)]^2. \tag{3}$$

Shannon [27] derived a diversity function from the theory of communication, described as

$$\Gamma = -\sum_{i=1}^{n_a} p_i \log_2 p_i, \tag{4}$$

where p_i is the occurrence rate of the ith allele, individual, or species from the set S, and n_a is the number of alleles, individuals, or species.

In our model, the population diversity is calculated based on Simpson method [23], with respect to the gene heterozygosity, H_e, such that

$$\Gamma = 1 - \sum_{i=1}^{n_a} (p_i)^2, \tag{5}$$

where n_a is the number of alleles and p_i is the occurrence rate of the ith allele. The number of alleles depends on the representation of the problem: for binary-coded EA, $n_a = 2$, for integer-coded EA, n_a is measurable, and for real-coded EA, n_a is unmeasurable. We used this equation to measure the population diversity because of its simplicity and low computational effort. This measure also reflects satisfactorily the population diversity, from its maximum level, $1 - 1/n_a$, when the number of each type of allele is the same, to zero, when there is only one type of allele. Its quadratic feature permits that the diversity, Γ, decreases faster than the distribution of the allele rates, $p_1, \ldots, p_{n_a}$.

In our population dynamics model, the number of alleles is calculated dividing the gene length into a given number of subintervals, i.e., the number of alleles, n_a, is the number of subintervals. Thus, the allele that belongs to a given subinterval j is defined as an g_{ij} allele, i.e., a type j allele from a i gene. The subinterval length, which defines each allele type, is calculated as

$$\Delta g = \frac{g_{\max} - g_{\min}}{n_a}, \tag{6}$$

where $g_{\min}$ and $g_{\max}$ are the minimum and maximum gene limits, respectively. Thus, the first and n_ath allele type are within $[g_{\min}, g_{\min} + \Delta g)$ and $[g_{\max} - \Delta g, g_{\max}]$ subintervals, respectively. For example, if a given gene is bounded by $[-10, 10]$ and $n_a = 10$, the subinterval length is equal to 2.0 and allele types are defined by the following subintervals

$$\begin{aligned} \text{Allele type \#1}: &\quad [-10, -8),\\ \text{Allele type \#2}: &\quad [-8, -6),\\ &\vdots\\ \text{Allele type \#}(n_a - 1): &\quad [6, 8),\\ \text{Allele type \#}n_a: &\quad [8, 10]. \end{aligned}$$

The Simpson measurement method uses only one gene of the genotype to describe the gene frequency or population diversity. In this paper, we call this gene the reference gene and show experimentally, in Section 5, that it may be possible to represent the population diversity with any gene of the genotype.

3. Population dynamics model

This section presents a new population dynamics model that describes the gene frequency behavior. As shown in Section 2, the model describes the population diversity with respect to only one gene, deemed the reference gene, with n_a alleles, and considers the selection pressure and mutation rate as evolutionary factor parameters. In this population dynamics model, crossover is not considered because it does not influence the calculation of the expected gene frequency in generational populations. This occurs because when an individual is selected for crossover, the allele of the reference gene is guaranteed in the next generation when the population model is generational. The model for population dynamics proposed has the following general features:

- for real-coded genotype, reference gene defined in $[g_{\min}, g_{\max}]$ interval, with n_a alleles, i.e., with n_a sub-intervals;
- all individuals are parent candidates;
- generational population model;
- no migration, and;
- roulette wheel or tournament selection method.

The population dynamics model is formulated in two stages, the parent selection for crossover and the mutation. In the first stage, the individual selection probability is a function of the selection pressure, β, in the roulette wheel, or of the group size, G, in the tournament. In the second stage, the gene frequencies are modified as a function of the mutation probability, p_m.

The gene frequency of a given allele type in the next generation depends on the number of selected individuals in which their alleles are from the same type. This occurs because when an individual is selected for crossover, the allele of its reference gene is guaranteed in the next generation. Of course, after crossover this reference gene may be modified by mutation.

The parent selection for crossover consists of two steps: in the first one, all individuals are parent candidates; in the second one, an individual chosen in the first selection cannot be chosen again. Since these events are mutually exclusive, the probability of the allele of the reference gene of an individual being selected in the next generation is the sum of probability of this individual being selected in the first and second parent selection for crossover.

The selection probability of an individual depends on the selection method. In the fitness proportional selection, roulette wheel

selection, the probability of the ith individual being selected for crossover in the first selection can be defined as

$$p_{s_1(i)} = \frac{w_i}{\sum_{j=1}^{N} w_j}, \tag{7}$$

where $w_i = w(f_i)$ is a function of the individual fitness, f_i, typically expressed as

$$w_i = f_i, \tag{8}$$

or

$$w_i = \exp(\beta f_i), \tag{9}$$

where β is the selection pressure, which determines the fitness impact over the selection probability. Thus, it is possible to control the selection pressure when Eq. (9) is used. For example, if $\beta = 0$, the selection will be at random ($w_i = 1, \forall i = 1, \ldots, N$); since β increases, the selection becomes elitist.

In the tournament method, the selection probability depends on the group size, G, and the ranking position of the individuals. Therefore, the selection probability of the ith individual is defined as follows

$$p_{s_1(i)} = \frac{\binom{N - N_s}{G} - \binom{N - N_s - 1}{G}}{\binom{N}{G}}, \tag{10}$$

where N_s is the number of individuals in which their fitnesses are better than that of the ith individual. For the best individual, $N_s = 0$; for the worst one, $N_s = N - 1$. The combinations from $C(N - N_s, G)$ are those in which the individuals better than the ith individual are not included, the combinations from $C(N - N_s - 1, G)$ are those in which the individuals better than the ith individual and itself are not included, and $C(N, G)$ is total number of combinations. The resultant on the numerator of Eq. (10) means the number of combinations in which the ith individual wins the tournament.

Mathematically, in a combination $C(n, k)$, the number of elements, n, must be larger than or equal to the number of groups, k. Nevertheless, in Eq. (10), $N - N_s$ or $N - N_s - 1$ can be less than G, e.g., for the worst individual in which $N - 1$ individuals are better than it. This means that there is no combination in which the worst individual wins the tournament. Thus, for the calculation of the selection probability in the tournament, the combination $C(n, k)$ becomes 0 when n is less than k.

In the tournament method, the selection pressure can be expressed as

$$\beta = \frac{G}{N}, \tag{11}$$

for $\beta \in [1/N, 1]$. For example, for $G = N$, i.e., $\beta = 1$, the selection pressure is maximum, elitist, which enables the best individual to be selected; for $G = 1$, i.e., $\beta = 1/N$, there is no pressure, the selection occurs at random.

In the second selection for crossover, the selection probabilities change proportionally to the selection probability of the first selected individual. Thus, the higher the $p_{s_1(\cdot)}$ of the first selected individual is, the larger are the changes of the selection probabilities of the remaining individuals. On the other hand, the higher the population is, the smaller are the changes of selection probabilities of the remaining individuals. In the second parent selection for crossover, the selection probability of the ith individual being selected when the jth individual has already been selected in the first parent selection is defined as follows

$$p_{s(i,j)} = p_{s1(j)}\, p_{s(i)}^{(-j)}, \tag{12}$$

where $p_{s(i)}^{(-j)}$ is the selection probability of the ith individual of a population without the jth individual. In the fitness proportional selection, $p_{s(i)}^{(-j)}$ can be calculated only by removing the jth individual fitness from Eq. (7), resulting

$$p_{s(i)}^{(-j)} = \frac{w_i}{\sum_{k=1}^{N} w_k}, \quad k \neq j. \tag{13}$$

In the tournament selection, $p_{s(i)}^{(-j)}$ becomes

$$p_{s(i)}^{(-j)} = \frac{\binom{N' - N'_s}{G} - \binom{N' - N'_s - 1}{G}}{\binom{N'}{G}}, \tag{14}$$

where $N' = N - 1$, because the first selected individual does not participate in the second selection, and $N'_s = N_s + 1$ if the ranking position of the jth individual is better than that of the ith one; or $N'_s = N_s$ otherwise.

The probabilities $p_{s(i,j)}$, $\forall j = 1, \ldots, N$, come from mutually exclusive events; thus, the selection probability of the ith individual in the second parent selection results

$$p_{s_2(i)} = \sum_{j=1}^{N} p_{s(i,j)}, \quad i \neq j. \tag{15}$$

For example, a population with three individuals, labeled as A, B, and C, have fitnesses equal to 2, 3, and 5, respectively. In the fitness proportional selection, the probabilities for the first selection, $p_{s_1(i)}$, $i = 1, 2, 3$, are 2/10 (20%), 3/10 (30%), and 5/10 (50%), respectively. If the individual B is selected in the first selection, the probabilities $p_{s_2(A)}$ and $p_{s_2(C)}$ will be 2/7 (29%) and 5/7 (71%); else if the individual C is selected, the probabilities $p_{s_2(A)}$ and $p_{s_2(B)}$ will be 2/5 (40%) and 3/5 (60%); then, the selection probability of the individual A for the second selection, $p_{s_2(A)}$, from Eq. (15), will be

$$\begin{aligned} p_{s_2(A)} &= p_{s1(B)}\, p_{s(A)}^{(-B)} + p_{s1(C)}\, p_{s(A)}^{(-C)} \\ &= \frac{3}{10} \times \frac{2}{7} + \frac{1}{2} \times \frac{2}{5} \\ &= 2/7 = 0.2857. \end{aligned} \tag{16}$$

Table 1 shows all probabilities for the first and second crossover selection by roulette wheel.

The same population, as in the paragraph above, with the tournament selection, and group size equal to 2, can provide 3 possible combinations for the first selection: AB, AC, and BC. Thus, the selection probabilities for the first selection, $p_{s_1(i)}$, $i = 1, 2, 3$, are 0/3 (0%), 1/3 (33%), and 2/3 (67%), respectively. Table 2 shows all probabilities for the first and second crossover selection by tournament.

Table 1
Proportional selection probabilities for a 3 individuals population.

Probabilities	$A(f_A = 2)$	$B(f_B = 3)$	$C(f_C = 5)$	Total
$p_{s_1(\cdot)}$	0.2000	0.3000	0.5000	1.0000
$p_{s_2(\cdot)}$	0.2857	0.3750	0.3393	1.0000
$p_{xo(\cdot)}$	0.4857	0.6750	0.8393	–

Table 2
Tournament selection probabilities for a 3 individuals population.

Probabilities	$A(f_A = 2)$	$B(f_B = 3)$	$C(f_C = 5)$	Total
$p_{s_1(\cdot)}$	0	0.3333	0.6667	1.0000
$p_{s_2(\cdot)}$	0	0.6667	0.3333	1.0000
$p_{xo(\cdot)}$	0	1.0000	1.0000	–

Finally, since the selection in the first and second parent selections are mutually exclusive events, the probability of the ith individual being selected for crossover, $p_{xo(i)}$, can be expressed as

$$p_{xo(i)} = p_{s_1(i)} + p_{s_2(i)}. \tag{17}$$

The last line of Tables 1 and 2 shows the selection probability for crossover of the 3-individual example population for the roulette wheel and tournament methods, respectively.

The number of offspring of an individual, $N_{d(\cdot)}$, among λ offspring in the next generation yielded by $\lambda/2$ crossovers, or Bernoulli samples, is a random variable with a binomial distribution $B(\lambda/2, p_{xo(\cdot)})$ defined by the following probability distribution

$$f(x) = P[N_{d(i)} = x] = \binom{\lambda/2}{x} (p_{xo(i)})^x (1 - p_{xo(i)})^{\lambda/2-x}. \tag{18}$$

Therefore, the expected number of offspring of the ith individual, $N_{d(i)}$, is defined by the mean

$$\begin{aligned} E[N_{d(i)}] &= \sum_{j=0}^{\lambda/2} j \times P[N_{d(i)} = j] \\ &= \sum_{j=0}^{\lambda/2} j \binom{\lambda/2}{j} (p_{xo(i)})^j (1 - p_{xo(i)})^{\lambda/2-j} \\ &= \sum_{j=1}^{\lambda/2} \frac{\lambda}{2} \binom{\lambda/2-1}{j-1} (p_{xo(i)})^{j-1} p_{xo(i)} (1 - p_{xo(i)})^{(\lambda/2-1)-(j-1)} \\ &= \frac{\lambda}{2} p_{xo(i)} \sum_{j=1}^{\lambda/2-1} \binom{\lambda/2-1}{v} (p_{xo(i)})^v (1 - p_{xo(i)})^{\lambda/2-1-v} \\ &= \frac{\lambda}{2} p_{xo(i)}, \end{aligned} \tag{19}$$

where $j\binom{\lambda/2}{j} = \lambda/2\binom{\lambda/2-1}{j-1}$ is an identity, $0 \le j \le \lambda/2$, and $v = j - 1$.

The number of type q alleles, $N_{a(q)}$, in the next generation is equal to the sum of all offspring in which the reference gene is a type q allele. Thus, the expected number of type q alleles is defined as follows

$$E[N_{a(q)}] = \sum_{i \in S_q} E[N_{d(i)}], \tag{20}$$

where S_q is set of individuals in which the reference gene is a type q allele.

After crossover, the population gene frequencies may be changed by the mutation. The reference genes with a type q allele may be changed, in addition to which the genes with non-q type alleles may become type q. In the mutation, the probability of changing a gene from type q to type non-q is

$$p_{m(\overline{q})} = p_m \left(1 - \frac{1}{n_a}\right), \tag{21}$$

where $(1 - 1/n_a)$ is the probability of an individual mutating its reference gene to a non-q type allele, given that the mutation is a function of the uniform distribution in the reference gene domain. The probability $p_{m(\overline{q})}$ considers that, when mutated, a gene may become the same allele type as the one before the mutation (otherwise, $p_{m(\overline{q})} = p_m$). This occurs because in the uniform mutation used, the whole gene domain is considered, including its value before the mutation. Thus, despite it having been mutated, the gene may keep on having a type q allele.

Among $E[N_{a(q)}]$, the number of mutated genes that stop being type q, $M_{a(\overline{q})}$, is a random variable with a binomial distribution $B(E[N_{a(q)}], p_{m(\overline{q})})$, defined by the distribution probability

$$\begin{aligned} f(x) &= P[M_{a(\overline{q})} = x] \\ &= \binom{E[N_{a(q)}]}{x} (p_{m(\overline{q})})^x (1 - p_{m(\overline{q})})^{E[N_{a(q)}]-x}. \end{aligned} \tag{22}$$

Thus, from $E[N_{a(q)}]$ offspring in which the reference gene is a type q allele, the expected number of mutated individuals whose reference genes will become non-q type alleles is

$$E[M_{a(\overline{q})}] = E[N_{a(q)}]\, p_{m(\overline{q})}. \tag{23}$$

The mutation may contribute to creating individuals in which their reference genes will be type q alleles when $N - E[N_{a(q)}]$ offspring with non-q type alleles are mutated and become type q alleles. The probability of an offspring mutating and its reference gene becoming a type q allele is

$$p_{m(q)} = p_m \frac{1}{n_a}. \tag{24}$$

The number of mutated offspring whose reference genes become type q alleles, $M_{a(q)}$, is a random variable with a binomial distribution $B(N - E[N_{a(q)}], p_{m(q)})$. Thus, the expected number of offspring with a non-q type allele, $N - E[N_{a(q)}]$, that they are mutated and become type q is

$$E[M_{a(q)}] = (N - E[N_{a(q)}])\, p_{m(q)}. \tag{25}$$

Finally, the number of offspring whose reference genes are type q alleles, $R_{a(q)}$, is a random variable and its expected value, $E[R_{a(q)}]$, is defined by the crossover selection and mutation contributions, as follows

$$\begin{aligned} E[R_{a(q)}] &= E[N_{a(q)}] - E[M_{a(\overline{q})}] + E[M_{a(q)}] \\ &= E[N_{a(q)}] - E[N_{a(q)}]\, p_{m(\overline{q})} + (N - E[N_{a(q)}])\, p_{m(q)} \end{aligned} \tag{26}$$

Replacing Eqs. (21) and (24) into Eq. (26), the latter results in

$$\begin{aligned} E[R_{a(q)}] &= E[N_{a(q)}] - E[N_{a(q)}]\, p_m \left(1 - \frac{1}{n_a}\right) + \left(N - E[N_{a(q)}]\right) p_m \frac{1}{n_a} \\ &= E[N_{a(q)}] \left[1 - p_m \left(1 - \frac{1}{n_a}\right) - p_m \frac{1}{n_a}\right] + N\, p_m \frac{1}{n_a} \\ &= E[N_{a(q)}](1 - p_m) + N\, p_m \frac{1}{n_a}. \end{aligned} \tag{27}$$

Therefore, the gene frequencies of the next generation may be approximated by the number of expected alleles and the population size, that is

$$\psi_i(k+1) = \frac{1}{N} E[R_{a(i)}], \quad \forall i = 1, \ldots, n_a, \tag{28}$$

Now, from Eq. (5), it is possible to predict the population diversity replacing p_i by $\psi_i(k+1)$, as follows

$$\Gamma(k+1) = 1 - \sum_{i=1}^{n_a} [\psi_i(k+1)]^2. \tag{29}$$

Eq. (27) shows that the gene frequencies in the next generation, $\psi_i(k+1)$, $\forall i = 1, \ldots, n_a$, are affected by the number of expected individuals, $E[R_{a(i)}]$, $\forall i = 1, \ldots, n_a$, which is a function of the selection pressure, β, and the mutation rate, p_m. The selection pressure affects the selection probability, thus benefiting the best fitness individuals as β increases. If the mutation rate is increasing, the number of expected offspring of a given allele type tends to decrease, thus benefiting the diversity. Thus, by using the selection pressure and mutation rate, it is possible to control the expected number of offspring.

The model proposed describes the gene frequency dynamics as a function of two evolutionary factor parameters, the selection pressure and mutation rate. This model of population dynamics

formalizes, through Eqs. (28) and (29), a model of evolutionary process.

4. Related works

Stark and Spall [21] created a model to compute the rate of convergence for the standard genetic algorithm, modeled as a Markov chain. Nix and Vose [28] showed that the stochastic transition through the genetic operations can be described by the transition matrix for one generation. Despite the authors used binary strings implementation, a discretization technique, such as that of Section 2, may be used to analyze the rate of convergence of the real-coded genetic algorithm. The model proposed by Stark and Spall [21] may be used to predict the population diversity; nevertheless, such model likely demands high computational effort because it is based on a $N_p \times N_p$ Markov chain transition matrix, where N_p is the total number of possible populations. Our model is dependent on the particular fitness of each individual. Thus, the diversity prediction can be calculated only as a function of each fitness, yielding a more efficient performance because the processing time to compute the probabilities increases linearly as the population grows.

Dasgupta et al. [29] proposed a mathematical model of the evolutionary dynamics of a one-dimensional differential evolution population. The fundamental dynamics of each search-agent (individual) employs the gradient-descent type search strategy. The stability and convergence-behavior of the proposed dynamics is analyzed with the stability theorems of Lyapunov [30]. The model of Dasgupta et al. [29] also may be used to predict the population diversity. However, this type of model cannot be used in any environment; for instance, in combinational optimization problems, the gradient-descent technique is not useful. Since our model uses only the fitnesses of the individuals, they can be used with distinct types of representations in different applications, such as computational and biological models.

Prugel-Bennett and Shapiro [22] created a dynamics model of genetic algorithms based on statistical mechanics. The authors studied the distribution of the fitness in a population generation by generation in an iterative way. The evolution of the system distribution allowed Prugel-Bennett and Shapiro [22] calculated the effect of selection, crossover, and mutation on an arbitrary distribution.

Prugel-Bennett and Shapiro [22] studied the diversity of the population fitness through the first four cumulants considering the mean, the variance, the degree of asymmetry of the distribution mean, and an indication if the function decays quicker or slower than a Gaussian. The evolution of the genetic algorithm is evaluated by the changes in time of the cumulants due to the genetic operators. The authors argued that the use of cumulants is suitable because they can be measured, they are more stable than the moments, and they can be related with the way the genetic algorithms work. For example, selection increases the third cumulant of the population. Hence, according to the authors, the extend to which the GA selection and variation operators decrease the third cumulant compared with how much they decrease the mean fitness is a measure of the effectiveness of the operators.

Differently of our model in which the genetic diversity is observed, the cumulants approach describes the distribution of the fitness function. In unimodal problems, the fitness distribution can indicate the location of the population. For instance, a population with a low fitness variance, from the second cumulant, indicates that the whole population is located around the same region. However, in multimodal problems, a low fitness variance cannot reveal this situation, because the population can be located on two or more local optimum with similar fitness values. In our approach, the genetic similarity signifies proximity in the search space, and it can be used to detect if the whole population is located in a particular region within the solution space. For example, by using the genetic diversity conjugated with the mean fitness, it is possible to estimate the spread of the population in the solution space.

Table 3
EA parameters set used in the example problem.

Parameter	Value
Population size	100
Individual size (# of genes)	10
Gene interval	[– 10, 10]
# of alleles	10
Parent selection rate	1.0
Crossover rate	0.7
Crossover method	2-Point
Mutation rate	0.05
Selection method	Roulette wheel
Selection pressure	0.08
Number of generations	100
Population model	Generational

Table 4
Initial and final diversity for f, Eq. (30).

Method	$\Gamma(0)$	$\Gamma(100)$
Euclidean-based	17.7073	2.5888
H_e-based (Gene #1)	0.8736	0.0810
H_e-based (Gene #4)	0.8858	0.0787
H_e-based (Gene #9)	0.8930	0.0895

5. Experiments on the diversity measurement

The basic idea of the population diversity, calculated by Eq. (5), is that the gene heterozygosity may reflect the population diversity. This section presents experiments to suggest that different genes produce similar diversity profiles for the same population. An example, from a maximization problem in which the fitness function has a quadratic form, is as follows

$$f = 100 - \sum_{i=1}^{10} x_i^2, \tag{30}$$

a population of 100 individuals, each one with 10 genes, is created to find the optimum point $X^* = [0, 0, \ldots, 0]^T$ at the origin. Table 3 shows the EA parameters set used in this example. All results presented were based on the mean of 30 simulations. The objective of this example is to analyze the diversity behavior of the population.

Table 4 shows the population diversity after 100 generations calculated by H_e and Euclidean-based methods used in [23,3], respectively. For H_e-based method, three genes, chosen at random, were used to calculate the population diversity. Euclidean and H_e-based methods describe a decreasing tendency for the population diversity. Fig. 1 shows initial and final allele distributions of the genes 1, 4, and 9. The initial distributions are approximately uniform. After 100 generations, the final distributions showed a tendency to decrease in diversity.

Based on the results presented in Fig. 1, Table 5 shows the minimum and maximum of the allele distributions for initial and final populations for the H_e-based method. Simulations produced

Table 5
Statistical data of the H_e-based methods.

Method	Min		Max	
	Initial	Final	Initial	Final
H_e-based (Gene #1)	5	0	14	95
H_e-based (Gene #4)	5	0	14	96
H_e-based (Gene #9)	7	0	13	97

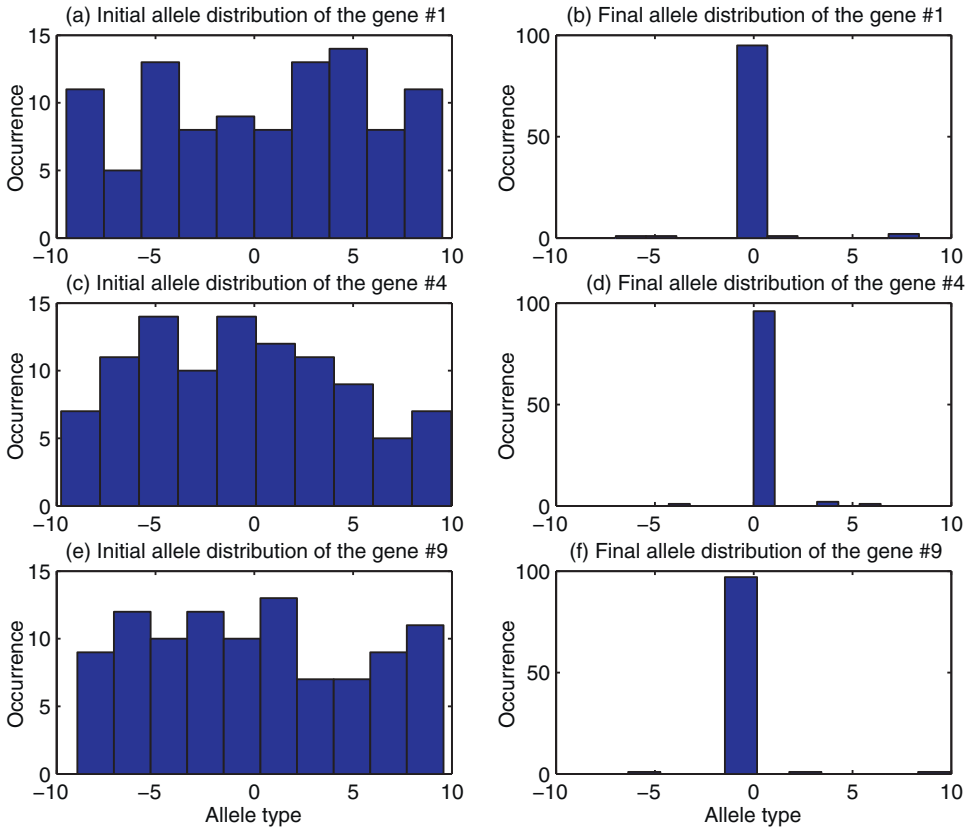

Fig. 1. Initial and final gene distributions for genes #1, #4, and #9.

similar results for genes 1, 4, and 9, as shown in Fig. 1 and Tables 4 and 5, suggesting that any gene produces the same effect with respect to allele distribution and diversity measured by Eq. (5).

6. Experiments on the population dynamics model

The population dynamics model was tested in several simulations with different population sizes and three types of alleles, $n_a = \{2, 10, 100\}$. The analysis was based on the error with respect to an EA with a generational population. Table 6 shows the functions, their parameters, and boundaries used in the experiments. Function f_1 is a continuous quadratic environment. Function f_2 is a non-continuous environment generator [31], in which N_p peaks (optimums) are specified, $N_p = 50$, each one independently specified by its location $x_{ij} \in [-100, 100]$, $\forall i = 1, \ldots, N_p$, $\forall j = 1, \ldots, L$, height $H_i \in [0, 300]$, and slope $R_i \in [2, 7]$. Function f_3 is the combinatorial NP-complete knapsack problem [32], the target of which is to find the best choice of objects that can fit into one knapsack to be carried on a single trip. Given a set of n types of objects, $n = 30$, each one with a weight, $w_i \in [10, 20]$, and a value, $v_i \in [0, 10]$, determine the number of objects, $x_i \in \{0, 1\}$, which are present in a collection so that the total weight, $\sum_{i=1}^{n} x_i w_i$, is less than a given limit, $C = 200$, and the total value, $\sum_{i=1}^{n} x_i v_i$, is as large as possible.

Table 6
Test functions used to validate population dynamic model.

Function	Boundaries
$f_1(X) = a - b\sum_{i=1}^{L} x_i^2$, for $a = 100$ and $b = 10^{-5}$	$x_i \in [-100, 100]$
$f_2(X) = \max_{i=1,\ldots,N_p}\left[H_i - R_i\sqrt{(x_1 - x_{1i})^2 + \cdots + (x_n - x_{ni})^2}\right]$	Variable
$f_3(X) = \sum_{i=1}^{n} x_i v_i$, subject to $\sum_{i=1}^{n} x_i w_i < C$	$x_i \in \{0, 1\}$

Table 7 shows the parameter values used in the experiments. The population size varied from 10 to 400 individuals and the parent selection rate was 1.0, i.e., the whole population could be selected to crossover. The performance analysis was based on the following statistics obtained during the evolutionary process for the mean results of 30 runs for each population size:

- Min and Max: minimum and maximum error.
- Mean: mean error, $\bar{e}$.
- SD: standard deviation.
- Mean (%): mean percentual error.

The population dynamics model was created for a large population, where the expected allele frequencies are error-free. For the quadratic function, f_1, and a non-continuous environment generator, f_2, $n_a = \{10, 100\}$ were used. The objective was to analyze the impact of different numbers of alleles upon the predicted diversity. For the knapsack problem, f_3, $n_a = 2$ because only one object

Table 7
Parameters used in the experiments.

Parameter	Value
Population size (N)	10, 50, 100, and 400
Individual size (# of genes, L)	10
Number of alleles (n_a)	{ 2, 10, 100 }
Parent selection rate	1.0
Crossover method	2-Point
Crossover rate (p_c)	0.7
Mutation method	Uniform
Mutation rate (p_m)	0.05
Selection method	Roulette wheel
Selection pressure (β)	0.01
Number of generations	100
Population model	Generational

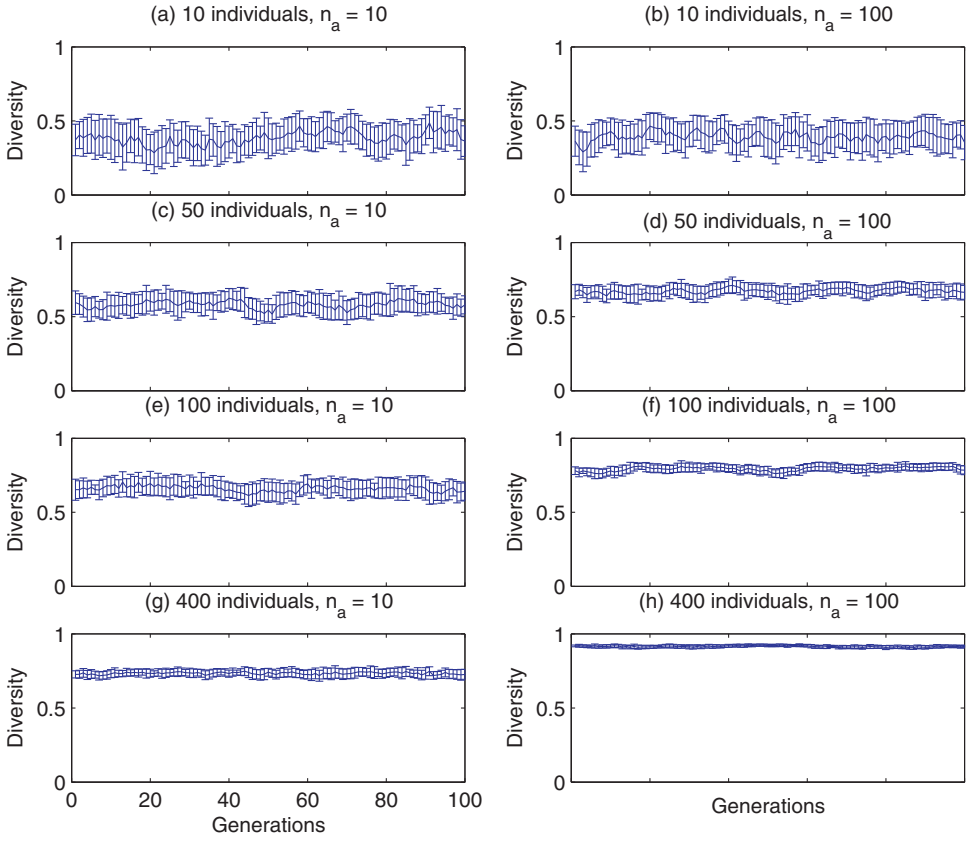

Fig. 2. Error bar for quadratic function environment, f_1.

Table 8
Overall results of quadratic function, f_1.

Population size	Min	Max	Mean	SD	Mean (%)
10-Allele					
10	0.0723	0.1549	0.1158	0.0171	30.62
50	0.0516	0.1113	0.0754	0.0117	12.98
100	0.0313	0.0824	0.0640	0.0097	09.70
400	0.0169	0.0419	0.0295	0.0053	04.02
100-Allele					
10	0.0738	0.1384	0.1072	0.0145	26.87
50	0.0312	0.0692	0.0465	0.0075	6.89
100	0.0173	0.0385	0.0273	0.0045	3.43
400	0.0054	0.0148	0.0101	0.0019	1.11

of each type can be carried in the knapsack on the trip, resulting in a binary-coded representation. The following subsections present the experiment results and analyses of the experiment.

6.1. Quadratic function

Fig. 2 shows the error bar of the population diversity for the quadratic function environment, f_1. As expected, the diversity error[1] tends to decrease when the population size increases. The overall results of the quadratic function, f_1, are shown in Table 8. The mean relative error for a population size of 10 individuals was very expressive. Nevertheless, the proposed population dynamics model may be used for a small population (e.g., 10 individuals) when high accuracy is not required, e.g., for tendency analysis.

[1] The diversity error is the difference between the population diversity and the predicted error calculated by the proposed model.

For populations with more than 50 individuals, the diversity prediction errors and the standard deviations decrease with respect to smaller populations. The standard deviations for these population sizes suggested robustness since the error variations were small. The 100-allele case produced results that were quantitatively similar to those of the 10-allele case. Fig. 2 shows the error bar of the population diversity for the quadratic function environment, f_1. All diversity prediction errors can be seen in Table 8.

For both 10- and 100-allele cases, the statistical results were very close to each other. All population sizes in both 10 and 100-allele cases reached low standard deviations, thus suggesting the robustness of the proposed model of population dynamics.

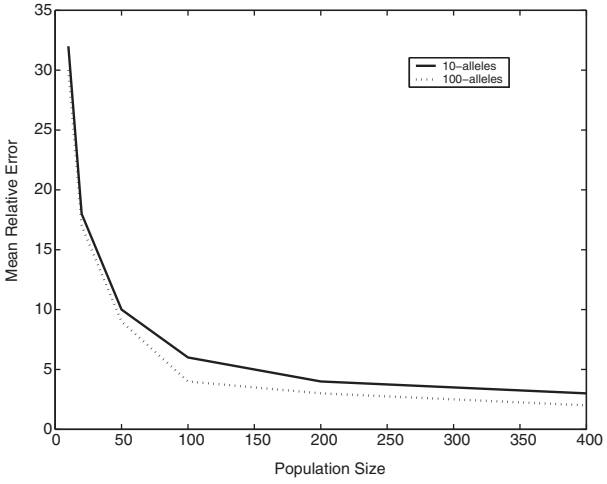

Fig. 3. Mean relative error for both 10 and 100-allele case.

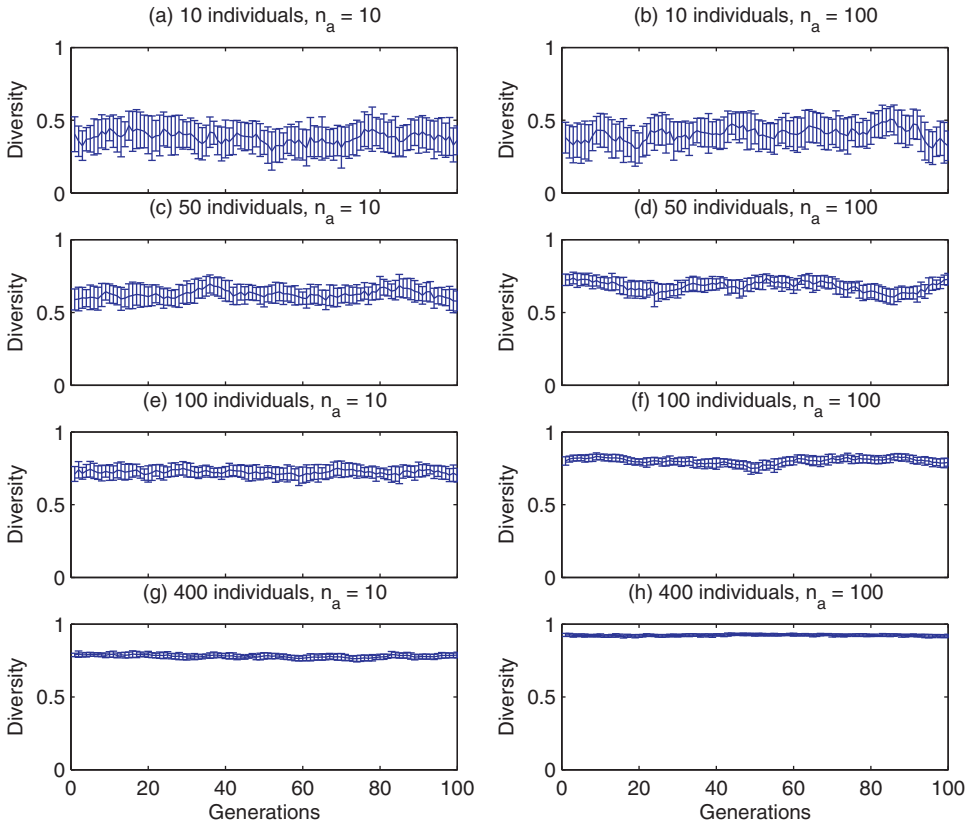

Fig. 4. Error bar for moving peaks generator, f_2.

Fig. 3 suggests that the mean relative error, for all populations in both 10- and 100-allele cases, tends to decrease quickly for the populations of 50 individuals.

6.2. Moving peaks generator

Fig. 4 shows the error bar of the population diversity for the moving peaks generator, f_2. For both 10- and 100-allele cases, the error decreases as the population size increases. A decrease in error as the population size increases is also seen in the 100-allele case. For the highest populations, 100 and 400 individuals, the diversity did not decrease as it did for the smallest populations, as a result of the generational model and, especially, of the low selection pressure.

Table 9 shows the statistical results, which confirm that the errors decrease as the population size increases. The standard deviations also confirm, especially for large populations, that the error is kept almost constant throughout the evolutionary process, as can be seen in Fig. 4.

Table 9
Overall results of moving peaks generator, f_2.

Population size	Min	Max	Mean	SD	Mean (%)
10-Allele					
10	0.0854	0.1601	0.1119	0.0153	29.18
50	0.0381	0.0974	0.0671	0.0118	10.84
100	0.0376	0.0890	0.0611	0.0115	9.14
400	0.0109	0.0262	0.0178	0.0032	2.34
100-Allele					
10	0.0648	0.1499	0.1083	0.0162	25.46
50	0.0256	0.0719	0.0415	0.0071	5.69
100	0.0155	0.0342	0.0228	0.0039	2.79
400	0.0057	0.0135	0.0093	0.0017	1.01

6.3. Knapsack problem

Fig. 5 shows the error bar of the population diversity for the knapsack problem, f_3. Analogous to the previous experiments, the error decreases as the population size increases in the knapsack problem. Table 10 shows the statistical results, and confirms the experiments mentioned above. The decrease in errors and standard deviations decrease as the population sizes increase, and their value ranges were similar to the previous experiments.

7. Analyses and comparisons

The experiments presented in Section 6 showed that the error prediction of the population diversity tends to decrease as the population size increases. Such results reinforced the accuracy of the population dynamics model, Section 3, proposed in this paper.

This population dynamics model can be used in different domains and approaches. Particularly in evolutionary algorithms, the model can be used to predict the population diversity and to analyze its sensitivity to variation of some evolutionary parameter, such as selection pressure and mutation rate. Thus, the designer can choose the appropriate parameter to be adjusted considering a given problem. Also, several works proposed a diversity-based EA [33] without regarding the diversity sensitivity [3,15–18]. In this

Table 10
Overall results of knapsack problem, f_3.

Population size	Min	Max	Mean	SD	Mean (%)
10	0.0721	0.1841	0.1293	0.0207	41.65
50	0.0272	0.0740	0.0472	0.0106	11.93
100	0.0169	0.0485	0.0288	0.0060	6.76
400	0.0077	0.0202	0.0132	0.0022	3.05

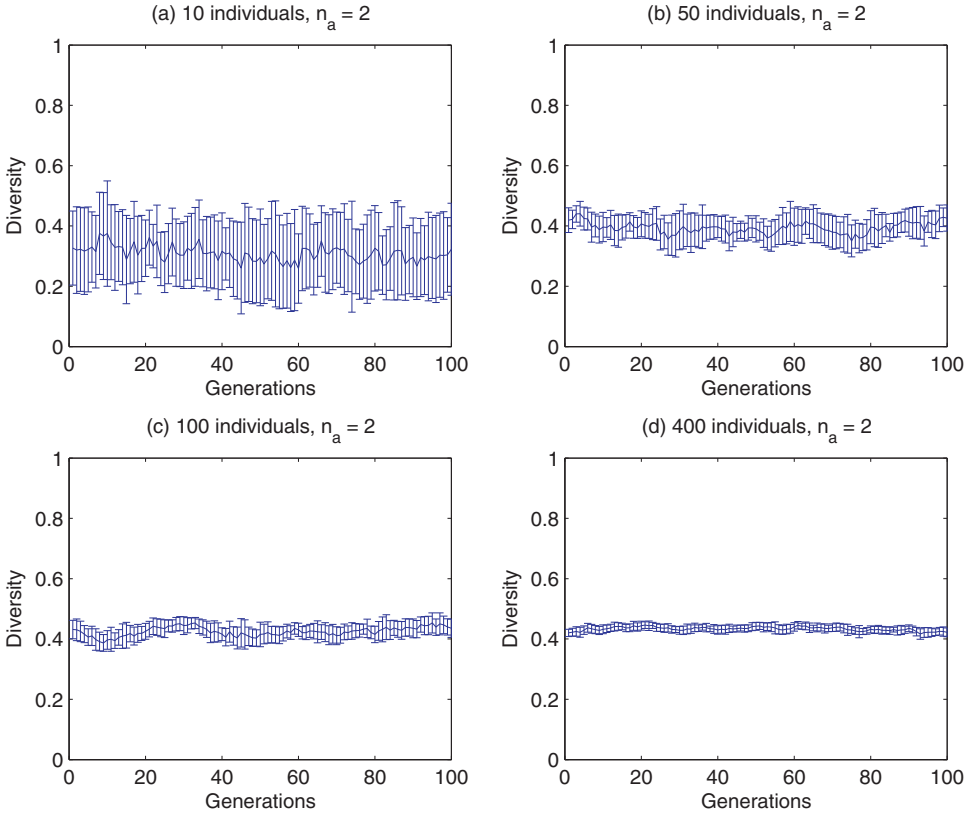

Fig. 5. Error bar for knapsack problem, f_3.

case, given our proposed model, an evolutionary parameter can be adjusted according to both the error between the population and the aimed diversities, and the measure of its sensitivity [34]. In nature, to maintain genetic diversity is the central concern of conservation biologists [35,36]. In a biological population, diversity is required for adaptation to environmental changes. Taking into account a population of individuals, loss of diversity is related with fitness decreasing. This is a deleterious consequence of the increased homozygosity at loci, a degradation factor for the fitness [37]. There are studies proposing diversity prediction methods in order to study the health and longevity of biological populations [38–40]. If the species mutation rate and type of selection (e.g., at random) are known, and the individual fitness with respect to a specific gene can be defined, the population dynamics model may predict the population diversity knowing only the current generation data.

Since the evolutionary process is naturally stochastic, for finite populations, there is an inherent error which increases as the prediction is performed for several generations ahead. Nevertheless, for populations larger than 100 individuals, as shown in Section 6, we believe that the prediction error for several generations ahead may be kept around the mean value obtained in the experiments for one generation ahead. Moreover, a prediction for only one generation ahead is very important for the study of population diversity and for its sensitivity analysis. The former can be used by conservation biologists and the later can be used to adjust some evolutionary parameter in a diversity-based evolutionary algorithm, as discussed above.

The population dynamics model proposed was designed for an infinite population; thus, the prediction error is zero, because the selection probability, Eq. (7) or (10), is computed according to a deterministic fitness. However, if the fitness function is stochastic, even for an infinite population, there will be a prediction error. For example, in the roulette wheel, the selection probability for the ith individual can be defined by Eqs. (7) and (9), and its expected number of offspring is computed by Eq. (19) without an error for an infinite population. For a stochastic fitness function, the selection probability for the ith individual, Eq. (7), will be now computed as

$$p_{s_1(i)} = \frac{f_i + \epsilon_i}{\sum_{j=1}^{N} f_j + \epsilon_j}, \tag{31}$$

where ϵ_i is the stochastic portion of the ith individual fitness. Therefore, the prediction error of the population diversity for a stochastic optimization problems is always present, even for the infinite population case, and it increases as ϵ_i grows. On the other hand, if the mean value of the portion ϵ_i, $\forall i$, is known, we may use it to minimize the prediction error. For example, Lin and Horng [41] created a method to minimize the average number of overkills per wafer, $E[V]$. Thus, this objective function is based on a mean value which provides an error whether $E[V]$ deviates from the correct value of V.

8. Conclusions

In this paper, a new model to describe the population diversity dynamics was presented. The model was originally designed for a generational population. The population dynamics proposed is based on current allele frequencies and two evolutionary factor parameters: mutation rate and selection pressure.

Different test functions and numbers of alleles were simulated. In all of them, the diversity errors were close for the same population sizes. For all types of test functions and numbers of alleles, the error decreases as the population size increases in similar proportions. As shown in Figs. 2–5, the error tends to decrease for populations larger than 50 individuals. The results for both 10- and

100-allele cases suggested the proposed model is reliability and robustness for large populations (100 and 400 individuals), with small errors and standard deviations. For small and medium populations (10 and 50 individuals), the diversity error was significantly higher. Nevertheless, if in applications to a small population, high precision in prediction is not essential, the model may be used to indicate tendencies.

A population dynamics model aims to contribute to applications such as parameter control, diversity control, and convergence rate analysis. For example, it is possible to calculate the approximate gradient of the expected population diversity with respect to some genetic parameters, such as the mutation rate or selection pressure. This direction may be used to adjust a genetic parameter in a diversity control problem [18]. In these applications, the knowledge of the evolutionary process is essential not only to understand its dynamics, but also to define strategies and adjustment directions.

Another important application of our model may be for real population analysis. The population dynamics model can be used in predicting loss of diversity, especially because real populations have large sizes, thus allowing accurate predictions. Our model may be used to estimate the gene frequencies of a population or species, and to calculate their deviation from the Hardy–Weimberg equilibrium. Thus, an evolutionary factor with significant impact upon the evolutionary process could be foreseen before the next generation comes.

Acknowledgement

The authors gratefully acknowledge the National Council for Scientific and Technological Development (CNPq) which supported this research study.

References

[1] M.R. Sherif, I.W. Habib, M. Nagshineh, P. Kermani, Adaptive allocation of resources and call admission control forwireless atm using genetic algorithms, IEEE Journal on Selected Areas in Communications 18 (2) (2000) 268–282.
[2] M.A.C. Lima, A.F.R. Araújo, A.C. César, Adaptive genetic algorithms for dynamic channel assignment in mobile cellular communication systems, IEEE Transactions on Vehicular Technology 56 (5) (2007) 2685–2696.
[3] R.K. Ursem, T. Krink, M.T. Jensen, Z. Michalewicz, Analysis and modeling of control tasks in dynamic systems, IEEE Transactions on Evolutionary Computation 6 (4) (2002) 378–389.
[4] I.L.L. Cruz, L. Van Willigenburg, G. Van Straten, Efficient differential evolution algorithms for multimodal optimal control problems, Applied Soft Computing Journal 3 (2) (2003) 97–122.
[5] J. McCall, Genetic algorithms for modelling and optimisation, Journal of Computational and Applied Mathematics 184 (1) (2005) 205–222.
[6] A.E. Eiben, R. Hinterding, Z. Michalewicz, Parameter control in evolutionary algorithms, IEEE Transactions on Evolutionary Computation 3 (2) (1999) 124–141.
[7] J.E. Smith, T.C. Fogarty, Operator and parameter adaptation in genetic algorithms, Soft Computing 1 (2) (1997) 81–87.
[8] M. Srinivas, L.M. Patnaik, Adaptive probabilities of crossover and mutation in genetic algorithms, IEEE Transactions on Systems, Man and Cybernetics 24 (4) (1994) 656–667.
[9] R. Hinterding, Z. Michalewicz, A.E. Eiben, Adaptation in evolutionary computation: a survey, in: Proceeding of the 4th IEEE Conference Evolutionary Computation, IEEE Press, Indianapolis, USA, 1997, pp. 65–69.
[10] M. Sebag, M. Schoenauer, Controlling crossover through inductive learning, in: Y. Davidor (Ed.), Proceeding of the Third Conference on Parallel Problem Solving from Nature, Springer-Verlag, 1994, pp. 209–218.
[11] W.M. Spears, Adapting crossover in a genetic algorithm, in: J.R. McDonnell, R.G. Reynolds, D.B. Fogel (Eds.), Proceedings of the Fourth Annual Conference on Evolutionary Programming, MIT Press, Cambridge, MA, 1995, pp. 367–384.
[12] J. Arabas, Z. Michalewicz, J. Mulawka, Gavaps – a genetic algorithm with varying population size, in: Proceedings of the First IEEE International Conference on Evolutionary Computing, IEEE Press, 1994, pp. 73–78.
[13] D.S. Voosen, H. Mühlenbein, Adaptation of population sizes by competing subpopulations, in: Proceedings of the 1996 IEEE Conference on Evolutionary Computation, IEEE Press, Piscataway, NY, 1996, pp. 330–335.
[14] F. Herrera, M. Lozano, Two-loop real-coded genetic algorithms with adaptive control of mutation step sizes, Applied Intelligence 13 (3) (2000) 187–204.
[15] P. Monsieurs, E. Flerackers, Reducing population size while maintaining diversity, in: C. Ryan, T. Soule, M. Keijzer, E. Tsang, R. Poli, E. Costa (Eds.), Genetic Programming, Proceedings of EuroGP'2003, Vol. 2610, Springer, Berlin/Heidelberg, Essex, UK, 2003, pp. 142–152.
[16] Y.-Y. Wong, K.-H. Lee, K.-S. Leung, C.-W. Ho, A novel approach in parameter adaptation and diversity maintenance for genetic algorithms, Soft Computing 7 (8) (2003) 506–515.
[17] N. Chaiyaratana, T. Piroonratana, N. Sangkawelert, Effects of diversity control in single-objective and multi-objective genetic algorithms, Journal of Heuristics 13 (2007) 1–34.
[18] M.M. Gouvêa Jr., A.F.R. Araújo, Diversity-based model reference for genetic algorithms in dynamic environment, in: in: 2007 IEEE Congress on Evolutionary Computation, CEC'07, IEEE Press, Singapore, South Korea, 2007.
[19] N.J. Radcliffe, Equivalence class analysis of genetic algorithms, Complex Systems 5 (2) (1991) 183–205.
[20] M.D. Vose, G.E. Liepins, Punctuated equilibria in genetic search, Complex Systems 5 (1) (1991) 31–44.
[21] D.R. Stark, J.C. Spall, Rate of convergence in evolutionary computation, in: in: 2003 IEEE American Control Conference, IEEE Press, 2003, pp. 1932–1937.
[22] A. Prugel-Bennett, J.L. Shapiro, An analysis of genetic algorithms using statistical mechanics, Physical D 104 (1997) 75–114.
[23] A.E. Magurran, Measuring Biological Diversity, Blackwell, Oxford, UK, 2004.
[24] C.R. Rao, Diversity and dissimilarity coefficients: a unified approach, Theoretical Population Biology 21 (1982) 24–43.
[25] A. Solow, S. Polasky, J. Broadus, On the measurement of biological diversity, Journal of Environmental Economics and Management 24 (1) (1993) 60–68.
[26] S. Champely, D. Chessel, Measuring biological diversity using Euclidean metrics, Environmental and Ecological Statistics 9 (2) (2002) 167–177.
[27] C.E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423, 623–656.
[28] A. Nix, M.D. Vose, Modelling genetic algorithms with Markov chains, Annals of Mathematics and Artificial Intelligence 5 (1991) 79–88.
[29] S. Dasgupta, S. Das, A. Biswas, A. Abraham, On stability and convergence of the population-dynamics in differential evolution, AI Communications 22 (1) (2009) 1–20.
[30] K.S. Narendra, A.M. Annaswamy, Stable Adaptive Systems, Dover Publications, 2005.
[31] R.W. Morrison, Designing Evolutionary Algorithms for Dynamic Environments, Springer, Berlin, 2004.
[32] R.E. Smith, D.E. Goldberg, Diploidy and dominance in artificial genetic search, Complex Systems 6 (3) (1992) 251–285.
[33] J. Branke, Evolutionary Optimization in Dynamic Environment, Kluwer Academic Publishers, Norwell, Massachusetts, USA, 2002.
[34] M. M. Gouvêa Jr., Algoritmo evolucionário adaptável em problemas multimodais dinâmicos, Ph.D. thesis, Recife, PE, Brasil, 2009 (in Portuguese).
[35] R. Frankham, Effective population size/adult population size ratios in wildlife: a review, Genetical Research 66 (1995) 95–107.
[36] R. Frankham, Genetics and conservation biology, Comptes Rendus Biologies 326 (2003) S22–S29.
[37] D. Charlesworth, B. Charlesworth, The genetic basis of inbreeding depression, Genetics Research 74 (1999) 329–340.
[38] M.J.E. Charpentier, C.V. Williams, C.M. Drea, Inbreeding depression in ring-tailed lemurs (lemur catta): genetic diversity predicts parasitism immunocompetence and survivorship, Conservation Genetics 9 (6) (2008) 1605–1615.
[39] J.N. Griffin, V. Mendez, A.F. Johnson, S.R. Jenkins, A. Foggo, Functional diversity predicts overyielding effect of species combination on primary productivity, OIKOS 118 (1) (2009) 37–44.
[40] J.R. Zaneveld, C. Lozupone, J.I. Gordon, R. Knight, Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives, Nucleic Acids Research 38 (12) (2010) 3869–3879.
[41] S.Y. Lin, S.C. Horng, Application of an ordinal optimization algorithm to the wafer testing process, IEEE Transactions on Systems, Man and Cybernetics: Part A – Systems and Humans 36 (6) (2006) 1229–1234.