Parameter estimation of a demand forecasting function associated with the behavior of weed using genetic algorithm

Marinna S. Sterzo*[†], Paulo E. Cruvinel*

*Embrapa Instrumentation
Brazilian Agricultural Research Corporation, São Carlos, SP, Brazil
Email: paulo.cruvinel@embrapa.br

[†] Electrical Engineering Department
Federal University of São Carlos, São Carlos, SP, Brazil
Email: ssmarinna@gmail.com

Abstract—This paper presents a computational method used to estimate the parameters of a demand forecasting equation, and its semantic interpretation, that describes the behavior of a generation of weeds in an agricultural field. Its implementation was made with a genetic algorithm (GA) structured by evolutionary logic, which uses the technique of identification of the function parameters from the seeds density of weed on an experimental field. In order to evaluate and validate the developed technique, a seeds density of weed database was generated from the application of original parameters of the weed *Abutilon theophrasti* which can be found in a maize crop (*Zea mays*) area. This database was used, and the GA demonstrated a high efficacy. In 38% of the analyzed cases, the fitness was below 5 for 100 realized experiments. This associated variable, i.e, for each analyzed individual, is related to the mean square error (MSE), which has its existence in the domain $0 \leq MSE < \infty$.

Keywords—*Weed, Genetic Algorithm, Estimation of Parameters.*

I. Introduction

Research and production in the agricultural sector have been benefited by the development of management technologies which consider the diversity of factors that interfere with productivity, including information on the area of planting and management techniques, such as Precision Agriculture. The spatial variability is one of the factors used in the area of soil science, where the use of soil fertility mapping is emphasized for the application of fertilizers and correctives in localized regions, a concept that can be extended to other areas, including the science of weeds.

According to Shiratsuchi and collaborators, weed control has traditionally been done by the uniform application of herbicides using a mean dose throughout the agricultural area. This can cause low control efficiency in some locations and, in others, leach chemicals products that cause environmental pollution [1].

Weeds can undergo high population growth in line with tillage, as certain farming procedures modify the environment of the place where the management takes place. This factor can cause loss of part of the agricultural production because it interferes directly with its quality. However, it is possible to map weed variability using precision farming tools in plantations, as the infestation of these plants usually happens in a non-uniform way, allowing decision making on the best management of the agricultural field. Analyzing the probability of spreading weeds is a way to prevent their infestation, since it a priori allows the control of the application of inputs in each place, eliminating waste and costs, and reducing the environmental impact [2].

This work aims to do the application of a probabilistic method based on the usage of the GA to estimate the parameters of a demand forecasting equation that describe the dynamics of the *Abutilon theophrasti* weed behavior into a maize crop (*Zea mays*) area.

II. Theoretical Bases

A. Weed seed density model

In 1995, Gonzales-Andujar emphasized the importance of combating weeds using high levels of control to achieve eradication. As the parameters considered in his work allow the modeling of weed seed density over the years it is possible to predict future generations of such plants [3].

Such information can be obtained by means of actual experiments made directly with the soil in which there are some of these plants and their surroundings. Additionally, from the use of such parameters it was possible to verify that the control rate of the population acted to reduce the survival of the weed seeds. These observations, added to the mathematical modeling of the problem, gave rise to the equation (1) which considers an initial seed density (N_{t-1}) which will provide, according to the value of the parameters, a value corresponding to the number of seeds per area of the following year (N_t), such as:

$$N_t = (1-c)sgN_{t-1}(1+agN_{t-1})^{-b} \qquad (1)$$

where N is the seed density, t is the generation corresponding to the year of the seed density, s is the number of seeds per plant in the absence of competition, g is the seed germination rate, a is the seed population decline rate, b is the intensity rate of plant density effect on seed production and c is the population control rate [3][2].

The model presented in the equation (1) was based on a nonlinear difference process that describes discrete variables and was developed from differential models of population dynamics of individual species over generations. From this primordial modeling, it is possible to evaluate the descriptive properties of plant and animal populations, in order to allow the appearance of reproduction functions as appropriate descriptors [4][5].

For purposes of validation of this work, the characteristic parameters of the *Abutilon theophrasti* weed presented in the Table I were considered. In this context, the values of the parameters a, b and s were considered to obtain successive seed densities from an initial seed, whereas the parameters c and g were proposed together with the adaptations made by the authors who developed the equation, as found in the references [6].

TABLE I: Parameters of the *Abutilon theophrasti* weed considered to obtain successive seed densities from one initial seed

Parameter	a	b	c	g	s
Value	0.0004	2.77	0.9	0.1	248

By applying the described parameters of the weed in the equation (1), the behavior of its seed density for a considered period of 20 years is obtained, as indicated in Figure 1 [7].
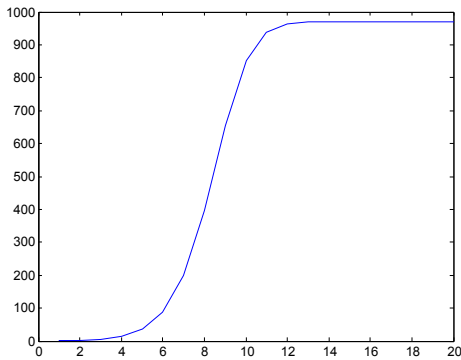

Fig. 1: Seeds density as a function of successive generations of the *Abutilon theophrasti* weed (adapted from [7].)

Note that in the developed research, the values presented in Figure 1 were used to generate the initial database as it can be better observed when detailing and validating the developed method.

The purpose of constructing such models of population dynamics using equations of differences is to enable the examination of potential population-level consequences of the effects of plant density and the factors affecting it. In the absence of a persistent seed bank, that is, a seed bank that does not survive independently of weed germination, the populations show damped or stable oscillations under low germination values, exhibiting non-cyclical oscillations of chaotic behavior or, extinct for higher values of germination due to complete failure of seed fixation at high plant density [8].

B. Demand forecasting

Demand forecasting methods have been developed with the aim of anticipating future states of factors and/or problem modeling variables that influence strategic planning and for which there is no immediate control. The choice of the forecasting method is always based on the characteristics and needs presented by these problem situations since the forecast of demand differs according to the methodology used according to the behavior of its object of study to be influenced by the factors that are around them. The equation (1) represents an exponential model of demand forecasting and fits into the descriptions discussed [9].

There are different methods associated with forecasting demands, which present both feasibility and problems that have motivated the construction of new solutions. These problems associated with demand forecasting models have been improved through the use of artificial intelligence (AI) techniques continuously, as well as using fuzzy logic, neural networks, and GA [6].

C. Genetic Algorithm (GA)

The bioinspired algorithms based on populations and metaheuristics are increasingly used in problem-solving, especially those involving non-differentiable or discontinuous functions since they are effective and have a broad domain of search and optimization of robust solutions. The fundamental principle of these algorithms is to obtain a random initial population through a constructive method and to implement the development technique, i.e., obtaining better individuals to managed in an adequate solution. Its major evolutionary components are populations of individuals, which the individuals of one or more populations compete for limited resources; fitness, which refers to the individuals' ability to survive and reproduce; notion of dynamic changes in populations due to the birth and death of individuals and concepts of variability and heredity, in which the new individuals possess many characteristics of their parents, although they are not identical. Such concepts come from the Synthetic Theory of Evolution or Neo-Darwinism, which identifies the mutation, genetic recombination, and natural selection as the main evolutionary factors [10].

The GAs are included in a subdivision of AI algorithms. The search techniques aim to solve complex problems through alternative methods obtain the solution as close to reality as possible while avoid robust mathematical models

using computational techniques that can consider a greater number of problem variables. Within this are encompassed stochastic methods, which follow rules of probability to search for solutions and, for this, use a defined interval to perform the search. The evolutionary algorithms are contained in the latter, which are methods that use rules of probability to find solutions, from where is derived the GA [11].

The main characteristics of a GA are the facts that operate on a population or set of points, have a space of coded solutions and do not act directly in the search space, thus requiring only information on the value of a fitness function for each member of the population. In this way, they use probabilistic transitions for their execution. In addition, as it is shown in Figure 2, its population size could be fixed or variable and its individuals have binary-encoded strings, factors that depend on of the application of the genetic operator's probabilities, which are chosen after an analysis of the problem to be optimized [11][12][13][14].

Algorithm 1 Standard Genetic Algorithm

1: $t \leftarrow 0$ ▷ start with an initial time
2: initpopulation P(t) ▷ initialize a usually random population of individuals
3: evaluate P(t) ▷ evaluate fitness of all initial individuals of population
4: **while** not done **do** ▷ test for termination criterion (time, fitness, etc.)
5:    $t \leftarrow t + 1$ ▷ increase the time counter
6:    P' $\leftarrow$ selectparents P(t) ▷ select a sub-population for offspring production
7:    recombine P'(t) ▷ recombine the "genes" of selected parents
8:    mutate P'(t) ▷ perturb the mated population stochastically
9:    P $\leftarrow$ survive P, P'(t) ▷ select the survivors from actual fitness

---

Fig. 2: Standard GA (adapted from [13]).

The GA realizes so many generations, whose computational equivalents are the iterations, as it is necessary to obtain the desired ideal solution, which, when it is known, can be stipulated as the algorithm's stop criterion. In contrast, there is no single criterion to define the iteration stop of the algorithm when the global optimum value is not known and it is necessary to use a limit of iterations. This limit represents a way to prevent the algorithm from continuing to run when Premature Convergence occurs, that is, the best individual has its value stagnated in a good location that doesn't represent the better solution. One of its causes is a phenomenon denominated in biology by genetic drift, which consists in the disappearance of certain genes of the population during the generations. When this happens, GA is unable to fully exploit the search space. With appropriate mutation rates, it is possible to maintain a good diversity of population genes in order to counterbalance these gene losses.

One of the advantages of using GA is that its robustness allows generating satisfactory solutions without depending on the choice of the initial parameters. In addition, it is characterized by the use of mutation and recombination operators, making the best solutions to be exploited without harming the exploration of the search space, which becomes multidimensional by preserving the groups of candidates for the solution and bringing about the exchange of information with the population [15].

There is in the literature of the area an optimization method called multi-objective optimization, which is characterized as a problem that needs a solution but has several objectives that must be simultaneously satisfied in the best way possible. To get a single solution, it is necessary to choose from a set of conflicting optimal solutions by considering the help of often subjective extra information called high-level information. If a weighted sum of the objectives is carried out through an importance vector, it is possible to transform a multi-objective problem into a simple goal, and the weight of each objective will be proportional to its importance, allowing a particular solution.This subjective procedure is called a preference-based multi-objective optimization [16].

Chodak and Kwaśnicki have developed a GA for to predict the demand for products and services by obtaining the parameters of their curve using the historical data describing the behavior of the demand and the attributes that this product or service has presented previously. It was done a binary GA that accomplished the search of parameters of the demand forecasting equation which its individuals was formed by the 6 parameters that described the attributes of the demand forecasting, each parameter consists of 10 bits. To analyze the fitness of individuals, first, the demand corresponding to the month considered was subtracted from the value provided when the set of parameters of an individual was applied to the demand forecasting function. In order to classify and identify the best candidates, the relative fitness of the set of values obtained by the first fitness analysis of individuals was calculated [6].

III. METHODOLOGY

The computational environment used to develop this method was the Matlab, with its high-level language and the processing of these experiments were performed on an i7 processor computer.

In order to optimize the dynamics model of weed seed density, it was used the GA to perform the search for parameters for the purpose of determining the values of the parameters of the equation (1), considered as the fitness function. The idea of the development of GA was based on the utilization of real weed seed density data. Therefore, it was considered the opportunity to use time series in order to find elements for the evaluation of parameters related to real information provided of an agricultural field.

The GA developed in this work has been based on the methods applied to demand forecasting proposed by [6] and the preference-based multi-objective optimization introduced by [16]. It was simulated 20-year weed seed density data to compose the database as high-level information for making comparisons, which are referred to here as high-level comparisons.

The binary base has been chosen to be used in this problem because it has the advantage that an individual can represent all 5 parameters at the same time. Thus, it is necessary to define a different maximum and minimum for each of the parameters, that is, an interval that was established according to the necessity of the parameter.

This occurs because it is vital to map them, ie, to force the set of bits extracted from the chromosome to belong to the interval when the binary base is converted to the real base, for which the equation (2) is used. The binary value represented by b_{10} is mapped to a real number within the range $(min < x < max)$ according to the number of bits nb [17], such as:

$$x = min + (max + min)\frac{b_{10}}{2^{nb} - 1} \quad (2)$$

where x is the real corresponding value of the individuals' fitness, min is the minimum value of the domain interval, max is the maximum value of the domain interval, b_{10} is the chromosome, i.e., the binary value that represents the individual and nb is the number of bits of the chromosome.

Then, the population control rate (c) and seed germination (g) are normalized, as they represent a percentage. The seed population decline rate (a) and the intensity rate of the effect of plant density on seed production (b) were defined according to the behavior of the equation (1). The number of seeds per plant in the absence of competition (s) is the only one that presents integer values because it represents seeds, so it is always rounded during the execution of the algorithm. The delimited ranges are indicated in Table II.

TABLE II: Delimited intervals for each parameter

Parameter	Minimum	Maximum
a	0	0,01
b	0	5
c	0	1
g	0	1
s	0	500

In this way, observing the limits of the intervals and the desired precision of the results' values, it was verified the need to define 12 bits for each parameter, resulting in an individual composed of 60 bits. The initial population's size was defined as 100 individuals, being kept fixed throughout the iterations through a simple method of selection, which occurs because the uniform crossover method used in crossover generates additional individuals according to the defined crossover probability [17].

The elitism method was applied in the GA, which the best solution candidates will always be inserted in the next population. For the genetic operators, the uniform crossover and the multipoint crossover methods were tested and the first was apply in the GA because it presents a higher efficiency than the second one. The uniform crossover method utilizes a mask of random generates bits which has the same size of the individuals to generate offspring from the parents, then the offspring contains a mixture of genes from each parent. Common mutation method was used while the roulette wheel selection was implemented for to do the selection of pairs for crossover. The latter consists in generate random numbers between 0 and the sum of all the population members fitness and to return individuals whose fitness corresponding to this value [13].

The probability of crossover and mutation were defined as 40% and 10% of the population, respectively, while the selection remained fixed in the same number of individuals of the initial population. These values were settled with the intention of avoiding the high occurrence of premature convergence of the algorithm, what was done after the performing of tests with other crossover and mutation probabilities. However, it was noticed that using the smaller values, stagnation was delayed and the mutation allowed small changes in the chromosome that reflected in smaller alterations in the fitness of the individuals, benefiting the search for the better solution because of the increasing of the chance of the individuals' fitness. By means of such procedure, the result is improved and the loss of good candidates by mutation is diminished.

The parameters with the defined values and with 1 initial seed per unit of area as seed density were used to generate 20 points by the equation (1) representing 20 years of data collected to compose the actual seed bank in all 100 tests performed. This procedure was necessary to perform a proof of the algorithm because the result that must be obtained is known as the parameters presented in the Table I.

These generations were inserted in the high-level comparison for to compare the converted values of the individuals' parameters among the values of the real database. The equation (2) was proposed by [17] to do the conversion conversion of the binary value of the individuals' to the real parameter values (a, b, c, g and s). This set of value was inserted in the fitness function 19 times because there is the one initial known generation that was used to generate the next generations. Thus, there are 20 generation values that should be analyzed for their fitness, but it is impractical to use more than one fitness value for each individual in a GA when it has only one objective. In this specific step of the algorithm, this objective means to determine the set of points that are closest to the original points inserted at the beginning of the algorithm, because if the points are said to be the closest, their parameters will be considered the best solution. This is the case with fitness, in order to solve the impasse, the MSE technique was used, in which is calculated considering the 20 points and allows to perform a classification of the individuals using only one fitness value to each one.

The performance function, as presented in the equation (3), consists of, for each individual of the population, accomplishment the steps of applying values to the fitness function, realize the high-level comparison and calculating the MSE, such as:

$$f_i(N) = \sum_{t=1}^{n} \frac{(X_t - N_t)^2}{n} \quad (3)$$

where $f_i(N)$ is the performance function for the current individual, N is the seed density generated by the GA resulting from the fitness function, i.e. equation (1), X is the seed density taken from the database, t is the generation corresponding to the year of the seed density and n, whose value is 20, is the total number of generations considered.

Therefore, the value resulting from its execution is the fitness of the individual.

The range of fitness can be set to $0 \leq MSE < \infty$, because theoretically the original points of each generation can be exactly equal to the points of each calculated generation or they may have completely different values, which would give an MSE very high.

The convergence criterion was defined basing on the precision wanted for the solution that was intended to be obtained using the method presented in this paper. It was adopted that the algorithm would consider the solution found if the suitability of the fitness of the best individual presents a value less than 5 since it is desired that the answer is as exact as possible. The iteration limit, used to stop the GA if it did not achieve the value of MSE required was 1000 iterations.

IV. Results and discussions

As explained in the methodology, the GA was executed to perform the 100 experiments with the already cited and described considerations. To aid in the analysis of the results, the Table III presents the range and the percentage of experiments that have the occurrence of the fitness. There was convergence in 38% of cases, but the set of parameters presents very diverse values since the original values of the parameters placed in Table II are not the only solution to the problem that within the proposed range.

TABLE III: Intervals of occurrences by MSE according to the results of the GA of parameter estimation for the 100 experiments performed.

MSE occurrence interval	Percentage of occurrences of MSE
$MSE \geq 500$	0%
$200 \leq MSE \leq 500$	7%
$50 \leq MSE \leq 200$	7%
$10 \leq MSE \leq 50$	23%
$5 \leq MSE \leq 10$	25%
$0 \leq MSE \leq 5$	38%

The Table IV displays the mean and standard deviation of the resulting values of the parameters obtained by the GA on the 100 experiences. The standard deviation reflects the probability of GA producing results whose values occur far from the mean [18], which reflect low dispersions.

TABLE IV: Mean and standart deviation of each parameter obtained by GA.

Parameter	a	b	c	g	s
Mean	0.005	2.896	0.659	0.150	138,640
Standart deviation	0.003	1.207	0.248	0.155	107.881

It can be verified that the results presented most of the occurrences in the values closing to the originals, so they are coherent with the expected ones. As can be seen in Figures 3 and 4, in which the normal distributions are showed, the histograms and normal curves that were produced with the resulting data provided by the GA ensure that the process is not random.
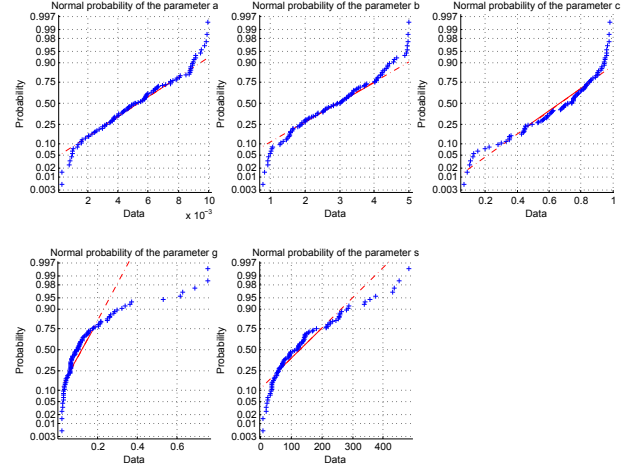

Fig. 3: Normal distributions of the parameters resulting from the 100 experiments done with the GA.
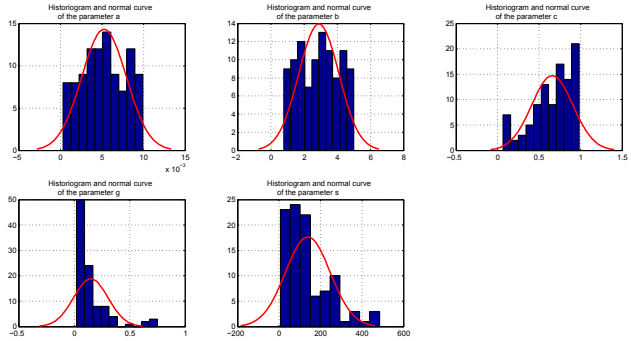

Fig. 4: Histograms and normal curves of the parameters resulting from the 100 experiments done with the GA.

The experiment 27 shown in Figure 5 presented the highest MSE value among the 100 experiments since its fitness was 481.314 when it reached the iteration limit.
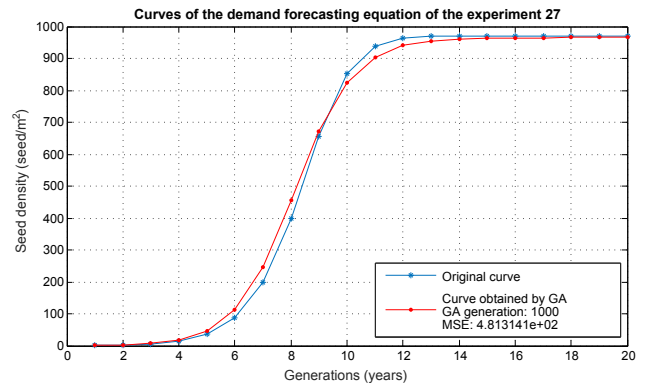

Fig. 5: Curves obtained through the original parameters and through the parameters with the higher fitness value among the 100 experiments that belongs to the experiment 27 in the GA execution.

The Figure 6 presents the result of experiment 13, which had the lowest fitness value achieved during GA execution of the 100 experiments. It took 47 generations to the value of MSE reach 0.937, producing the result with the closest possible curve to the original curve in this analysis.
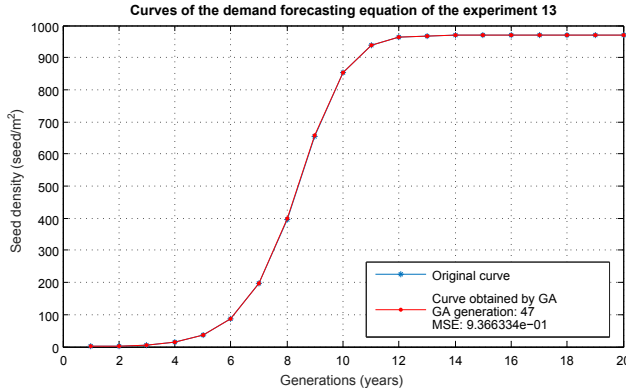

Fig. 6: Curves obtained through the original parameters and through the parameters with the lowest fitness value among the 100 experiments that belongs to the experiment 13 in the GA execution.

It is possible to verify the occurrence of the phenomenon of genetic drift in the GA population of this experiment since the fitness of individuals from 2 to 100 remained stagnant in 12.1502 when the convergence criterion was reached. The mutation of the GA produced small changes to the individuals' chromosome due to the 10% probability, which allowed the production of an individual with a small fitness value, as shown in Figure 7.
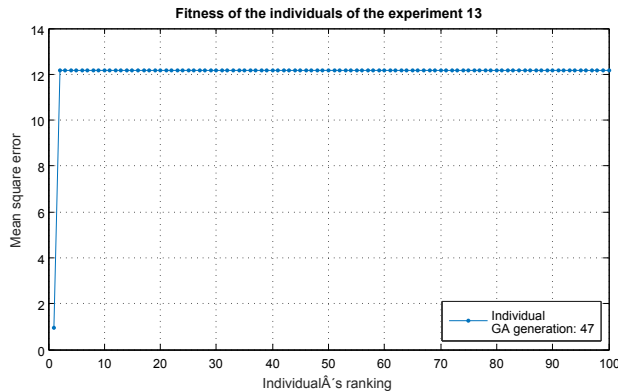

Fig. 7: Graph of the fitness of individuals obtained in experiment 13 resulting from the execution of GA by their rank in which the numbers 1 and 100 represent respectively the best and worst value.

To illustrate the capacity of the algorithm, the curves of the result of experiment 3, which achieve the best fitness value with the least number of iterations among all 100 experiments, are shown in Figure 8. The fitness reached the value of 3.5216 in only 15 iterations, which, besides

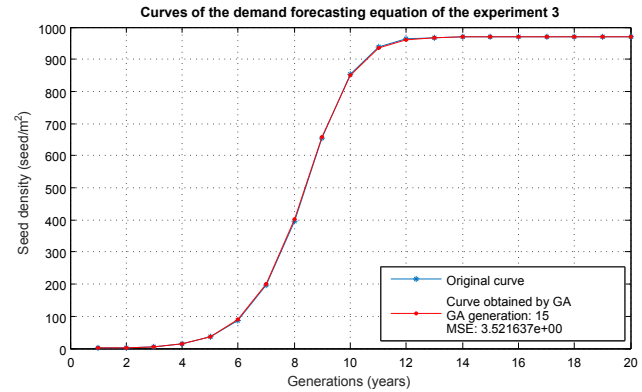indicating a great efficiency, also was among the 38% better results.

Fig. 8: Curves obtained through the original parameters and through the parameters with the quickly convergence among the 100 experiments that belongs to the experiment 3 in the GA execution.

The Figure 9 shows the great variability presented by the fitness values of the GA population of this experiment, i.e., the search space was well represented and explored. There was no appearance of the phenomenon of genetic drift yet, since as no signs of stagnation among the first individuals of the population when the fitness reached a lower value than the criterion of convergence required.
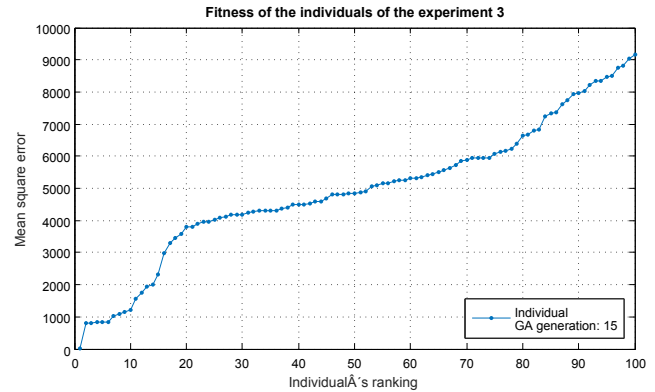

Fig. 9: Graph of the fitness of individuals obtained in experiment 3 resulting from the execution of GA by their rank in which the numbers 1 and 100 represent respectively the best and worst value.

As can be seen in the Figures 5, 6 and 8 that presents curves produced through the equation (1), which the blue one was made with the original parameters and the red one was made with the better parameters of the experiences 27, 13 and 3, respectively, the results have shown that the GA achieve an adequate reproduction of the actual values, which was possible to produce the set of parameters that generate the curve according to the behavior of the weed among the years. This occurred because the curvature of

the curves was corresponding with each other and the values for each year are very close.

V. CONCLUSION

The GA based on evolutionary logic developed to perform the estimation of parameters that describe the dynamics of weed behavior yielded results whose MSE values were below 5 in 38% of the time during the 100 experiments performed, indicating a great accuracy. The adoption of crossover and mutations probabilities of 40% and 10%, respectively, minimized the occurrences of premature convergence of the algorithm and the losses of the best solutions. Besides, providing small variations in the individuals' structures, is possible to make the best solution often, i.e., producing close values with lesser fitness. The resulting parameters values for the same curve had low values of standard deviation because of show low dispersion, as expected. Therefore, it can be affirmed that the developed algorithm allowed to obtain parameters that adequately represent the description of an exponential model that respects the actual behavior of weed seeds.

REFERENCES

[1] L. S. Shiratsuchi, P. J. Cristoffoleti, and J. R. A. Fontes. *Mapeamento da variabilidade espacial das plantas daninhas com a utilização de ferramentas da agricultura de precisão.* PhD thesis, Universidade de São Paulo, 2001.

[2] *Sistema computacional de baixo custo para reconhecimento da variabilidade de plantas daninhas baseado em imageamento estereoscópico*, 2004.

[3] J. L. Gonzalez-Andujar. High control measures cannot produce extinction in weed populations. *Ecological modelling*, 91(1):293–294, 1996.

[4] A. M. Mortimer, J. J. Sutton, and P. Gould. On robust weed population models. *Weed Research*, 29(4):229–238, 1989.

[5] R. Law and A. R. Watkinson. Response-surface analysis of two-species competition: an experiment on phleum arenarium and vulpia fasciculata. *The Journal of Ecology*, pages 871–886, 1987.

[6] G. Chodak and W. Kwaśnicki. Genetic algorithms in seasonal demand forecasting. 2000.

[7] G. M. Bressan, L. V. Koenigkan, V. A. Oliveira, P. E. Cruvinel, and D. Karam. Fuzzy classification system for risk of weed infestation considering spatial variability. *Planta Daninha*, 24(2):359–364, 2006.

[8] FA Bazzaz, DD Ackerly, FI Woodward, and L Rochefort. Co2 enrichment and dependence of reproduction on density in an annual plant and a simulation of its population dynamics. *Journal of Ecology*, pages 643–651, 1992.

[9] F. de O. Lemos. Metodologia para seleção de métodos de previsão de demanda. 2006.

[10] A. B. de S. Serapião. Fundamentos de otimização por inteligência de enxames: uma visão geral. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, 20(3):271–304, 2009.

[11] G. L. Soares. Algoritmos genéticos: estudo, novas técnicas e aplicações. *Belo Horizonte: Universidade Federal de Minas Gerais*, 1997.

[12] J. Tanomaru. Motivação, fundamentos e aplicações de algoritmos genéticos. In *Anais do II Congresso Brasileiro de Redes Neurais*, volume 1, 1995.

[13] K.-F. Man, K.-S. Tang, and S. Kwong. Genetic algorithms: concepts and applications. *IEEE transactions on Industrial Electronics*, 43(5):519–534, 1996.

[14] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[15] J. E. Baker. Reducing bias and inefficiency in the selection algorithm. In *Proceedings of the second international conference on genetic algorithms*, pages 14–21, 1987.

[16] R. L. Escobar. Otimização de parametros de projeto de sistemas mecanicos atraves de algoritmo genetico multi-objetivos. 2007.

[17] E. G. M. de Lacerda and A. C. P. L. F. de Carvalho. Introdução aos algoritmos genéticos. *Sistemas inteligentes: aplicaçoes a recursos hídricos e ciências ambientais*, 1:99–148, 1999.

[18] A. M. F. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics.* International Student edition. McGraw-Hill, 1974.