ORIGINAL PAPER

MATHEMATICAL MODEL OF SICKNESS DUE TO ALTITUDE DECOMPRESSION USING GENETIC PROGRAMMING. A META-ANALYSIS PERSPECTIVE

MARIUS TURNEA[1], MARIANA ROTARIU[1], MIHAI ILEA[1*], DRAGOS AROTARITEI[1]

Manuscript received: 17.06.2019; Accepted paper: 23.08.2019;
Published online: 30.09.2019.

Abstract. *Altitude decompression sickness can be mathematically modeled by analytic expressions but these models are actually guessed initially by the authors based on experience and data analysis. A systematic alternative approach is developed in this paper, a solution based on evolutionary approach, the genetic programming. The measure of fitness is evaluated by maximum likelihood estimation and experimental results are presented.*

Keywords: *meta-analysis, mathematical model, graphic user interface, genetic programing, machine learning algorithms.*

1. INTRODUCTION

Altitude decompression sickness (DCS) is usually defined as the evolution of gas bubbles that take place inside the human body due to decreasing the barometric pressure (hypobaric DCS) [1]. In normal atmospheric condition, the primary gas from human body is an inert gas, nitrogen. Depending of altitude, the barometric pressure decrease and quantities of nitrogen diffuses into blood, tissues and lung and sometime are expulsed in expired air. Henry's law postulate that the quantity of gas dissolved in a solution not chemically combined with the solution is a function of partial pressure and saturation constants [1]. The supersaturation due to bubbles of nitrogen with minor amounts of oxygen, carbon dioxide, and water vapor can produce decompression sickness (DCS).

In [2], the authors proved that solution for decompression sickness that occurs at altitude can be resolved at lower altitudes using hyperbaric therapy in hyperbaric chambers. The altitude-related DCS (we refer there about Type II) presents a wide spectrum of symptoms and sometime it is difficult to be diagnosed [3]. A number of 133 cases have been studied in [3] and the main manifestations associated have been classified: visual disturbances, joint pain (sometime associated with headache), and limb paresthesia. Other symptoms are investigated also and the treatment by hyperbaric oxygen was a fully successful method in 97.7% of the cases [3]. Physical exercises (dynamic arm or leg, isometric arm or leg) can induce a high DCE rate among the subjects (31%-50%) with a variation of VGE incidence (precordial venous gas emboli) as 47-66% among subjects [4].

In [5], the authors proposed a study related to O_2 pre-breathing on symptom and bubble incidence during decompression. The incidence of symptoms and bubbles are inverse correlated with pre-breathing time [5]. Susceptibility to altitude DCS can be influenced by many factors. In [6], in a study on a total of 291 human subjects, 197 men and 94 women, no

[1] "Gr.T. Popa" University of Medicine and Pharmacy, Department of Medical Biosciences, 700115 Iasi, Romania. E-mail: ileamihai2004@yahoo.com; mihail.ilea@umfiasi.ro.

significant results have found related to gender. Counterpart, the individual's age can be an influencing factor when the subject is between 18-21 or 26-41 years old, but no influence has been detected when the subject is in 26-41 years old [7].

The threshold altitude when 1 hour of oxygen pre-breathe is used is studied in [8]. The main conclusion of the study is that is a threshold DCS in value of 5% from 51 male subjects below 6858 m after 1 h of pre-breathe [8]. Variability in individual susceptibility to DCS is studied in [9].

The variables analyzed are anthropometric (age, height, body mass index, age, weight, aerobic capacity, and percent body fat) along with physiologic ones. The results are rather negatives; it seems not to be a correlation between anthropometric and physiologic variables for individual susceptibility [9].

Some papers deal with simple models to determine some thresholds for DCS incidence using variables like denitrogenation, exercise, rest and period of exposure to altitude. The most simple one is logistic regression analysis applied to altitudes below 9,144 m subject to time exposure, no-prebreathe and physical exercise[10].

DCS is a complex problem that involves many variables [11]. In [11] after intensive research, the authors proposed 14 models based on the same logistic hazard function. The models are in fact extensions of the same approach, combination of log logistic models that include exercises [11]. There are also approaches much more complexes that introduce new probability distributions (e.g. hypertabastic) in [12], and the model has a very complicated formula including hyperbolic secant and hyperbolic cotangent.

2. MATERIALS AND METHODS

2.1. MLE (Maximum Likelihood Estimation) and LSE (Least Square estimation) for mathematical models

Mathematical models for DCS require a large amount of data, an adequate definition for DCS, physical and physiological variables that must be taken into account and an analytical approach. Most common representation for time survival from collected data is life table, but other parametric and non-parametric statistic based representations are used: Kaplan-Meyer curves, survival functions and hazard functions. Survival models usually fall in one of the following categories: Cox proportional hazards regression, parametric survival models, survival trees and a machine learning based method, survival random tree forest. Censoring (missing data problem) can be present or not in these analysis.

By probabilistic point of view, survival function S(t), cumulative distribution function F(t) denoted by CDF, hazard function h(t), cumulative hazard function H(t) and probability density function f(t), denoted by PDF are the main representation used in survival analysis [11]. Practical, all these forms are different description on the same survival analysis that is if you know one of them, you can determine the others. The table of survival time (censored or not) is the basis for analysis of survival time from experimental collected data.

We denote by $\Pr$ the probability, T the survival time (a random variable) and t a specific value of T (time) and let's suppose that $F(t)$ is derivable. The connections among these expressions are:

$$S(t) = \Pr(T > t) \tag{1}$$

$$F(t) = \Pr(T \le t) = 1 - S(t) \tag{2}$$

$$f(t) = F'(t) = \frac{d}{dt} F(t) \tag{3}$$

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)du = 1 - F(t) \tag{3.1}$$

The hazard function focus on fall of one individual meanwhile the survival function focus on survival of one individual. The hazard function is related to survival function and the relationship between them is:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = -\frac{dS(t)/dt}{S(t)} \tag{4}$$

$$S(t) = \exp\left[-\int_0^t h(u)du \right] \tag{5}$$

As usually method for survival model construction based on experimental data is to find a function S(t), the probability to be alive at time t by guessing that a known distribution function is close enough from experimental data. Any distribution defined for [t, +∞) can be used for survival function. Some of the most important parametric distributions which can be used are: exponential, Weibull, Gompertz-Makeham, Gamma, Generalized Gamma, Log-Normal, Generalized F, Coale-McNeil Model, inverse Gausssian, and Pareto ([13-16]). The form can be a simple one but sometime the form is more complicated, e.g. exponential power:

$$h(t) = \alpha \lambda^{\alpha} t^{a-1} e^{(\lambda t)^{\alpha}}, \quad \alpha, \lambda > 0, t \ge 0 \tag{6}$$

$$S(t) = e^{(1 - e^{(\lambda t)^{\alpha}})} \tag{7}$$

$$f(t) = \alpha e^{\lambda^{\alpha}} t^{a-1} e^{(\lambda t)^{a} - e^{e^{(\lambda t)^{a}}}} \tag{8}$$

The basic strategies for model fitting to survival models are in order of complexity, parametric approach, semi-parametric approach and non-parametric approach. In order to determine the parameters (constants) of a chosen model to fit to experimental data, two basic estimators of fitting are used: LM (least-squares estimation) and MLE (maximum likelihood estimation) [18].

Let's denote by $w=[w_1, w_2, ...w_m]$ the parameter vector and $x=[x_1,x_2,..., x_n]$ the data vector for a random PDF (probability density function) denoted by f(x|w). The principle of MLE was developed initially in 1929 by Fisher and states that the desired probability distribution is one that is closest of the real one in terms that maximize the likelihood (conventionally log-likelihood) function $L(w|x)$ where w and x roles are reversing [19].

Once the distribution for is guessed, the next step is to find the parameter w, usually by gradient methods [19]. LSE and MSE estimates can differ in situations when probability density function in not known by analytical formulas (it is supposed only). Counterpart, MSE refers in generally to probabilities meanwhile LSE are most common to fit a function given pairs of (x,y) n-dimensional vector data.

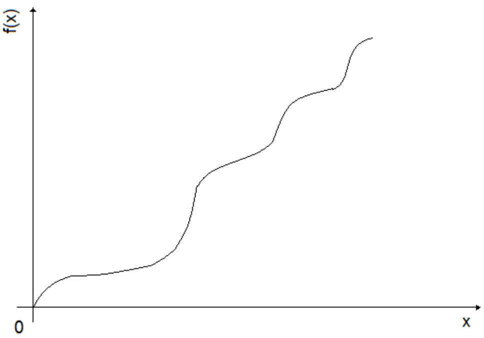

Figure 1. A fictitious function CDF.

Association with a known probability function is task that depend on distribution of the points (if the visual inspection can suggest a known one), the most common situation, but if the curve has many "S", as in fictitious curve (Fig. 1), the initial guess of distribution can be extremely difficult. In this case, new methods can give a solution for these type of problems, and one of them is based on evolutionary computation, genetic programming described by shortly in the next section.

2.2. Genetic Programming (GP)

Bernoulli and Poison schemas are the most common used to model probability distributions but the models are not efficient in the case of DCS when many parameters and variables are involved. Other complex random distributions are difficult to find out and alternative solutions can application of a systematic method to find a mathematical for of these. Because any mathematical curve has many different analytical forms than fit with a good precision for one experimental dataset, there can is various mathematical expressions found by genetic programming algorithms.

Genetic Programming belongs to a larger class of algorithms that are based on evolutionary mechanism of evolution (evolution by natural selection). In evolutionary algorithms, a population (fixed or variable) of individuals where each individual competes with the others in order to reach a user defined objective (or more objectives, possible in opposition), so namely fitness functions [19]. Two common ways are usually proposed measure of performance for fitness function f_{fit}: maximization of f_{fit} or minimization f_{fit}. The most used one is the minimization meanwhile the maximization is given simple by minimizing the f_{fit} with negative sign, which is the function $-f_{fit}$.

As in genetic algorithms (GA), the main operations are selection, crossover and mutations. Meanwhile in GA the individuals are coded in chromosomes represented by a linear succession of allele, in Genetic Programming, the individuals are coded in trees (that represent functions) as chromosomes in inverse Polish notation. All the operation as selection, crossover and mutations are made using tree or subtrees from chromosome.

The representation of a simple program that codes a function is given in figure 2. The crossover and mutation are presented in one example in Figs. 2-3. The operation of selection is made using elitism; the best two chromosomes are selected to produce two offspring with a single cut of chromosomes. The mechanism of selection was selected among the most popular one: roulette, tournament, and stochastic sampling.

The initial population was generated using a random generator (in our case, uniform distribution), and the population is considered fixed.

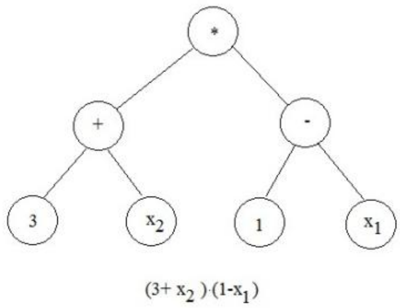

Figure 2. A simple program in Genetic Programming.

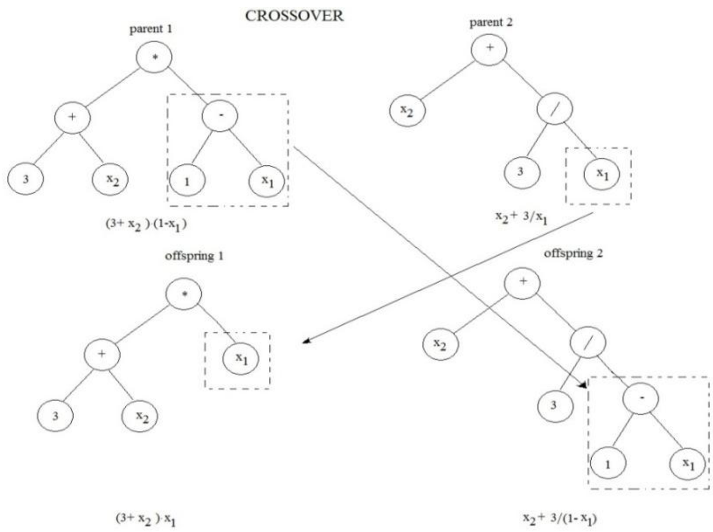

Figure 3. The Crossover operator in Genetic Programming .

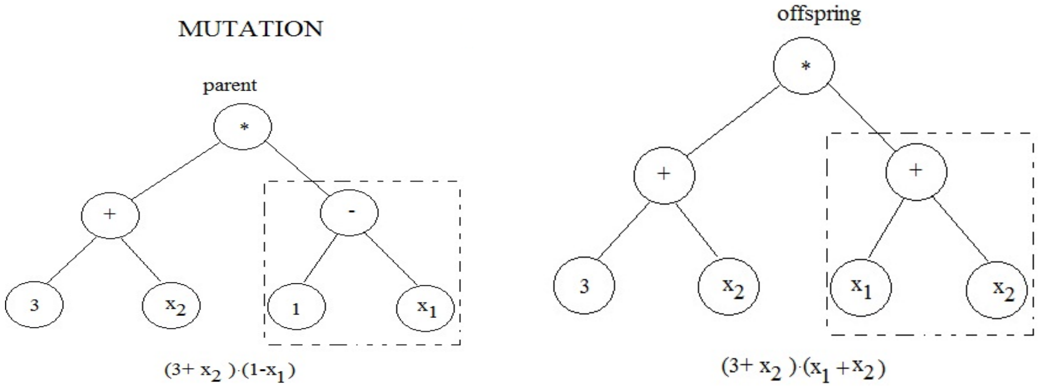

Figure 4. Mutation in Genetic Programming .

The crossover operation are used to create new individuals (new trees) in order to optimize the fitness function. The less performant individuals (two from group of two parents that are subject of crossover and the corresponding offspring) are discarded.

The operation of mutations is performed with a small predefined probability in order to prevent the premature convergence and elitism disadvantages. The population is evaluated in a loop until stop conditions are fulfilled. The stop condition can be a predefined number of iterations or a number of steps that cannot produce an improvement of fitness for the best individual from population. If the convergence is premature and the goal is not achieved, the method used is to restart the procedure with a new random population.

An extension of Genetic Programming is proposed in [20]: individuals are made by a combination of few genes corresponding to trees in a linear way and optimization of them in two stages: optimization of each gene and optimization of the linear combination that represents a linear regression. The optimal weights obtained by least mean squares fit the genes with output data [20]. In our application, we used the software GPTIPS 2.0 [20]. Practically, a nonlinear model is modelled by a combination of pseudo-linear models with nonlinear genes.

3. RESULTS AND DISCUSSION

In [20], decompression sickness (DCS) and venous gas emboli (VGE) consequences are modeled based on experimental data collected from a database. We will compare our result with the performance goodness-of-fit for PDF function as example of power of the usage of GP algorithms.

A cumulative DCS incidence after 4-h exposure is considered to be modeled at 4.25 psia (pounds per square inch absolute). Because the shape of data seems to be somewhat like in [20], following the same reasons, a similar function is proposed as model:

$P(DCS)_t=1-\exp(-\ln[1+a_1\cdot(t\cdot a_2)^{a_3}])$ where $A=[a_1\ a_2\ a_3]$ are constants.

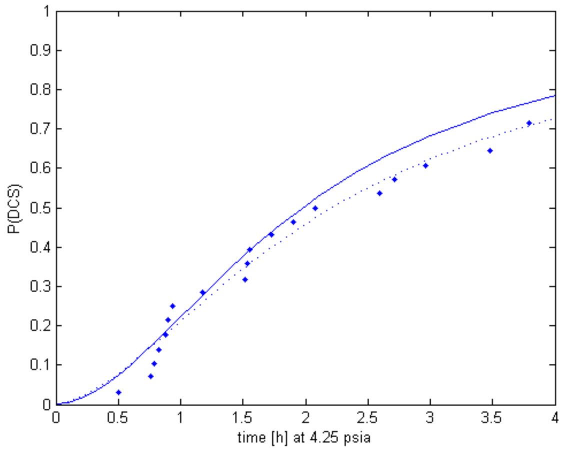

Figure 5. Continuum line – model using MLE (A = [0.9671, 0.515. 1.835]); dotted line model using LSE (A = [0.1931 1.22 1.6546]); dots are the PDF of experimental data.

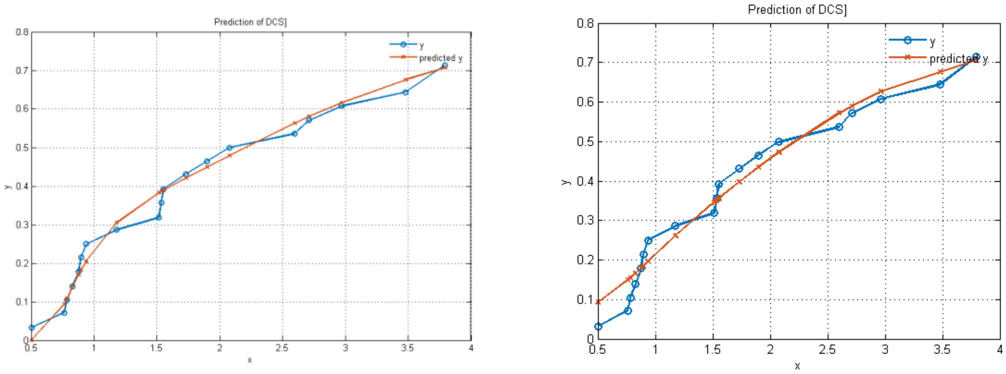

Figure 6. Two formulations in GP terms for LSE evaluation for experimental data from Fig. 5. Formula (9) in the left and figure and formula 10 in the right figure.

$$\exp(\exp(\frac{-t}{\exp(t)-1.0}) - \frac{1}{t \cdot t^{\exp(1-\exp(t)) \cdot \left(\exp(t)-t+\frac{\exp(t)}{t}\right)}} \tag{9}$$

$$\frac{t(t+\exp(t^{t-2.718})-1.0)}{4.437t+5.437\exp(t^{t-2.718})-6.437} \tag{10}$$

In order start the Genetic Programming algorithm, we use a set of terminals (T) and operators (F) [20]. When we use a multigene Genetic Programming, that the formula is given by a sum of individual genes that evolves separately in a genetic tree, the terminals operators must be defined commonly for all the trees.

For a single variable t, T = {*t*, *RA*}. In terminal set, RA is a random value in the range [–10.0, +10.0]. For set *F = {+,−, ∗, /, exp, power}* the results are showed in Fig. 6, two instances for the same set of operators, and for set *F = {+,−, ∗, /, sin, cos}* the results are showed equation (11).

$$\sin\left(\sin\left(\sin\left(\frac{1}{t}\right)\right) - \cos\left(\frac{\sin\left(\sin\left(\sin\left(\left(\frac{1}{t}\right)\right)\right)-1.0\right)}{t}\right)\right) \tag{11}$$

In [20], the best value goodness-of-fit is reasonable p=0.69, and in our experiments, the maximum value is very close to this, p =0.685 and p =0.712 for situation from Fig. 6.

A multigene option is testes also, with a limitation to maximum three numbers of genes. After 8 restarts, the best solution is given in equations (12)-(15).

$$g_1 = 4.781 + 1.366t^{1/2} - 0.07406 \tag{12}$$

$$g_2 = -1.701 \cdot (t \cdot (8.849 \cdot t + 3.0))^{1/2} \tag{13}$$

$$g_3 = -0.028478\cos(t^3 + 8.787t^2 + t) \tag{14}$$

$$g = g_1 + g_2 + g_3 \tag{15}$$

The fit in this case is very impressive, $R^2 \approx 0.99$, a very good value (Fig. 7). The evolutionary algorithm start with a random value allocated to genes and this start could drive to a non-satisfactory solution if the start point is to far from a minimum (maximum) fitness from solution hypersurface. The usually solution to this problems are the restart with new random values or constraints and additional operators, more complicated which operated on chromosomes.

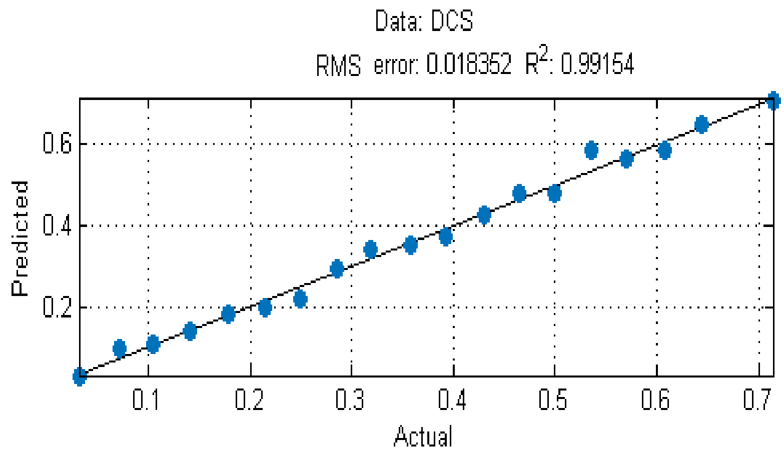

Figure 7. Fitness of the model for multigene regression (using [20]).

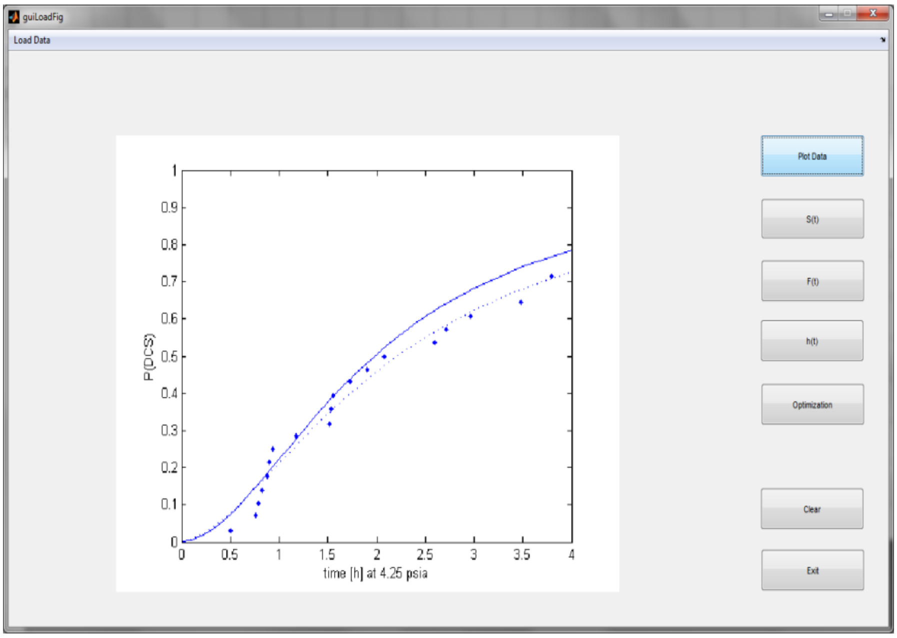

Figure 8. P(DCS) on graphical interface .

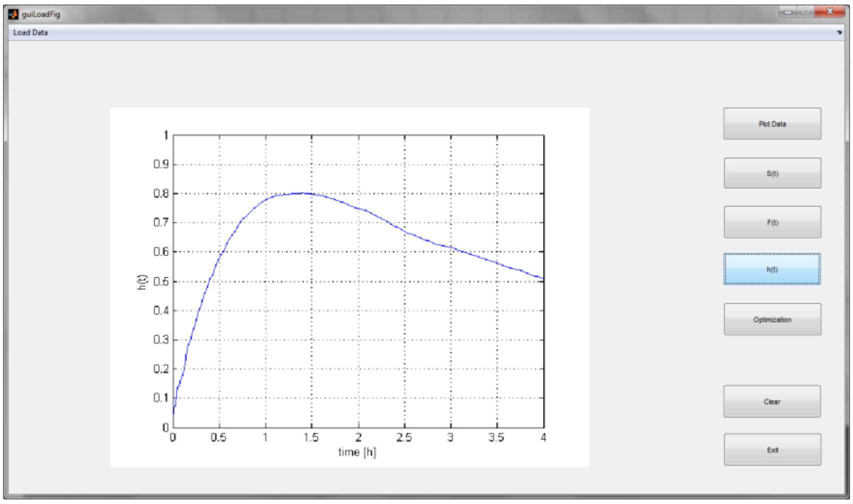

Figure 9. Hazard function h(t).

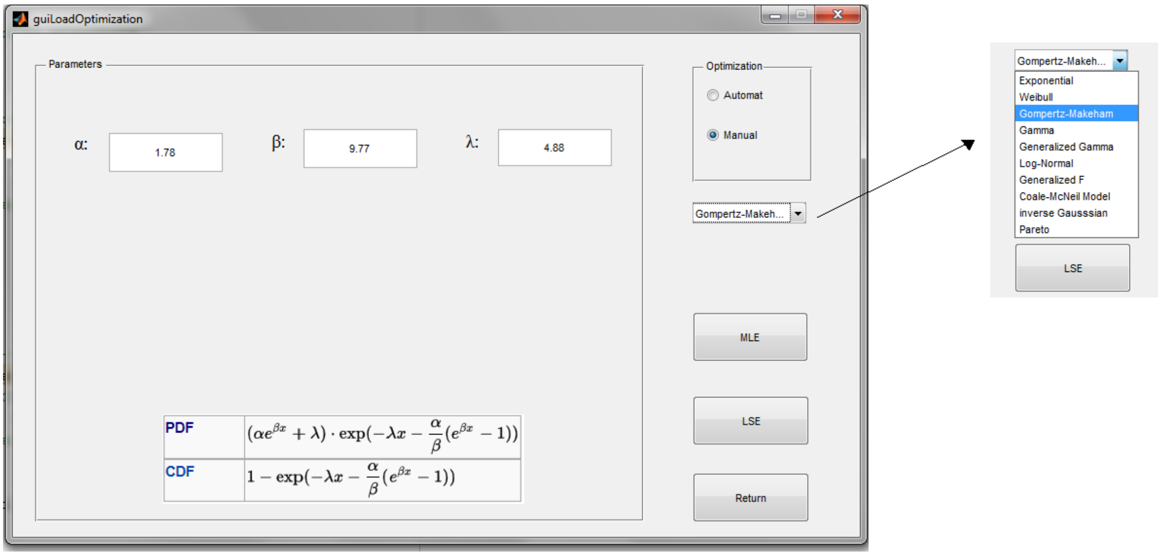

Figure 10. Manual optimization, second window from graphical interface.

A graphic user interface (GUI) is proposed to help the user to find a model for experimental data stored in a database. The graphical interfaces are intuitive and permit to find a model manual or automatic using Genetic Programming package.

4. CONCLUSIONS

A systematic method based on Genetic Programming is proposed in this paper. The proposed solution proved to be an option for very complicated probability distribution function or when we are looking for combination of known distribution (multigene GP).

The formulas can be very complicated, so a driven Genetic Programming, a solution that has some constraints (inspired by known distribution) is the subject of future research.

The DCS has a strong relation with physiology and perception. Neurons are primary responsible for perception and sensation so a modeling using neural networks can be a solution that reflects the physical reality.

Modeling by neural networks (e.g. pairs of cortical neurons neuron-antineuron for tuning and respectively noise) are difficult to apply because actually, the link between activity in neurons and sensory perception are still a challenging problem in neuroscience, but can be one of the directions for future research.

REFERENCES

[1] Davis, J.R., Johnson, R., Stepanek, J., Fogarty, J.A., *Fundamentals of Aerospace Medicine*, LWW, 4th Edition, 2008 .
[2] Davis J.C. et. al., *Aviat. Space Environ. Med*., **48** (8), 722,1977.
[3] Wirjosemito, S.A., Touhey, J.E, Workman, W.T., *Aviat. Space Environ. Med*., **60**(3), 256, 1989 .
[4] Pilmanis, A.A, Olson, R.M., Fischer, M.D., Wiegman, J.F., Webb, J.T., *Aviat. Space Environ. Med*., **70**(1), 22,1999 .
[5] Waligora, J.M., Horrigan, D.J, Conkin, J., *Aviat. Space Environ. Med*., **58**(9), 110, 1987.
[6] Webb, J.T., Kannan, N. , Pilmanis, A.A., *Aviat. Space Environ. Med*., **74**(1), 2, 2003.
[7] Sulaiman, Z.M., Pilmanis, A.A., O'Connor, R.B., *Aviat. Space Environ. Med*., **68**(8), 695, 1997.
[8] Webb, J.T., Pilmanis, A.A, *Aviat. Space Environ Med*., **76(**1),34, 2005.
[9] Webb, J.T. et. al., *Aviat. Space Environ. Med*., **48**(8), 722, 2005.
[10] Kumar, K.V., Waligora, J.M., Calkins, D.S., *Aviation, Space, and Environmental Medicine*, **61**(8),685, 1998.
[11] Kannan, N., Raychaudhuri, A., Pilmanis A.A., *Aviation, Space, and Environmental Medicine*, **69**(10), 965,1998 .
[12] Tabatabai, M.A., Bursac, Z., Wiliams, D.K., Singh, K.P., *Theoretical Biology and Medical Modelling*, **4,** 40, 2007.
[13] Rodriguez, G., *Parametric Survival Models. Technical report*, Princeton, NJ, Princeton University, 2010.
[14] Kleinbaum, D.G., Klein, M., *Survival Analysis. Statistics for Biology and Health*, 2nd edition, Springer, 2005 .
[15] Cox, D.R., Oakes, D. , *Analysis of Survival Data*, Chapman & Hall, London, 1984.
[16] Lawless, J.F. , *Statistical Models and Methods for Lifetime Data*, Wiley-Interscience, Hoboken, 2002.
[17] Klein, J.P., Moeschberger, M.L., *Survival analysis Techniques for censored and truncated data*, Springer-Verlag, New York, 2003 .
[18] Myung, I.J., *Journal of Mathematical Psychology*, **47**, 90, 2003 .
[19] Koza, J.R., *Genetic Programming*, MIT Press, Massachusetts, 1992.
[20] Conkin, J. , *Aviat Space Environ Med*., **82**, 589, 2011.