12th IFAC Conference on Programmable Devices and Embedded Systems
The International Federation of Automatic Control
September 25-27, 2013. Velke Karlovice, Czech Republic

Fuzzy Linear Regression Analysis ⋆

Jana Nowaková ∗ **Miroslav Pokorný** ∗∗

∗ *VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Department of Cybernetics and Biomedical Engineering, 17. listopadu 15/2172, 708 33 Ostrava Poruba, Czech Republic (e-mail: jana.nowakova@vsb.cz).*

∗∗ *VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science, Department of Cybernetics and Biomedical Engineering, 17. listopadu 15/2172, 708 33 Ostrava Poruba, Czech Republic (e-mail: miroslav.pokorny@vsb.cz).*

Abstract: The theoretical background for abstract formalization of vague phenomenon of the complex systems is fuzzy set theory. In the paper vague data as specialized fuzzy sets - fuzzy numbers are defined and it is described a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague parameters. Interval and fuzzy regression technologies are discussed, the linear fuzzy regression model is proposed. To identify fuzzy regression coefficients of model genetic algorithm is applied. The numerical example is presented and the possibility area of vague model is illustrated.

Keywords: Regression model, non-specificity, interval model, vagueness, fuzzy model, fuzzy number, genetic algorithms, possibility area.

1. INTRODUCTION

Regression models are used in engineering practice wherever there is a need to reflect more independent variables together with the effects of other unmeasured disturbances and influences Bardossy (1990), Shapiro (2006). In classical regression, we assume that the relationship between dependent variables and independent variables of the model is well-defined and sharp. In the real world, however, hampered by the fact that this relationship is more or less non-specific and vague. This is particularly true when modelling complex systems which are difficult to define, difficult to measure or in cases where it is incorporated into the human element Shapiro (2006).

The theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory Novák (1990). In the paper vague data as specialized fuzzy sets - fuzzy numbers are defined and a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague parameters is described.

2. FUZZY REGRESION ANALYSIS

Linear regression model of investigated system Shapiro (2006) is given by a linear combination of values of its input variables

$$Y^{*}(x_j) = A_0 + A_1 x_{1j} + \ldots + A_{nj} x_{nj}. \tag{1}$$

Conventional regression model is based on the assumption that the system characteristic is defined by sharp, precise and deviations between observed and estimated values of the dependent variables are the result of errors of observation.

However, the statistical regression models based on principles of probability theory are correct only if a number of preconditions is met Pokorný (1993), Shapiro (2006). The most common practical problems are

- a small number of observations , the sample is too small,
- we can not guarantee a normal distribution of error,
- difficult to define the relationship (vagueness) between the input and output variables.

These problems do not occur when the creation of regressions utilize possibility theory and regression dependence is identify as a fuzzy function. The origin of the deviation between the observed and estimated values of the dependent variables may not be significant extent caused by poor local variables of system structure. These variations can be caused by in not very sharp nature of the system parameters. Such fuzzy phenomenon must also be reflected in fuzziness of the corresponding parameters of the model.

If we consider the fuzzification of regression model, we can consider two cases (but which are not mutually exclusive). First of all, we consider that the input data are crisp and uncertain is in the definition of the model. In this case, the vagueness is reflected by fuzzy nature of the regression coefficients as model parameters. In the second case, we can consider the system as a well-defined and

⋆ This work has been supported by Project SP2013/168, "Methods of Acquisition and Transmission of Data in Distributed Systems", of the Student Grant System, VŠB - Technical University of Ostrava.

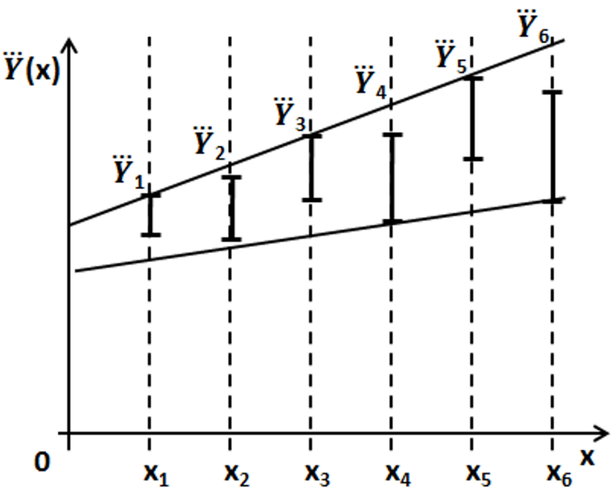

Fig. 1. One-dimensional linear interval regression model

fuzzy character have the measured data. Then the carrier of uncertainty of the model are vague input data. The paper analyzes the situation where both cases are applied.

2.1 Interval Linear Regression Model

The first step in creating a blurred regression model is the work Buckley (1990), who developed the technology of interval regression models

$$\ddot{Y}(x_j) = \ddot{A}_0 + \ddot{A}_1 x_{1j} + \ldots + \ddot{A}_j x_{nj}. \tag{2}$$

Where $\ddot{Y}(x_j)$ is the estimated value of the output variable as a closed numerical interval representing the uncertainty of the non-specific system and $\ddot{A}$ are the regression coefficients of the model again in the form of vague closed numerical intervals. To identify the intervals of regression coefficients the method of linear programming used in Kacprzyk (1992), for algebraic calculations with interval numbers simple interval arithmetic is developed Moore (1979). Example of a one-dimensional linear interval regression model is shown in Figure (1).

2.2 Fuzzy Linear Regression Model

The next step in the development of indeterminate regression model is the development of models of vague, using the formalization of uncertainty rather than numerical intervals using the fuzzy intervals. Regression models reflect the vagueness of the modelled systems are called fuzzy regression models Kacprzyk (1992), Poleshchuk (2012), Shapiro (2006). The indeterminate nature of fuzzy regression model is represented by the estimated fuzzy output values $\widetilde{Y}^*(x_j)$ and the fuzzy regression coefficients $\widetilde{A}$ in the form of specialized fuzzy sets - fuzzy numbers. Shape of fuzzy linear regression model Buckley (2008), Heshmaty (1985), Tanaka (1982) is given by The next step in the development of indeterminate regression model is the development of models of vague, using the formalization of uncertainty rather than numerical intervals using the fuzzy intervals. Regression models reflecting the vagueness of the modelled systems are called fuzzy regression models Kacprzyk (1992), Poleshchuk (2012), Shapiro (2006).

The indeterminate nature of fuzzy regression model is represented by the estimated fuzzy output values $\widetilde{Y}^*(x_j)$ and the fuzzy regression coefficients $\widetilde{A}$ in the form of specialized fuzzy sets - fuzzy numbers. Shape of fuzzy linear regression model Buckley (2008), Heshmaty (1985), Tanaka (1982) is given by

$$\widetilde{Y}^*(x_j) = \widetilde{A}_0 + \widetilde{A}_1 x_{1j} + \ldots + \widetilde{A}_j x_{nj} = \widetilde{\mathbf{A}}.\mathbf{x}' \tag{3}$$

where $\mathbf{x}'$ is a transposed column vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\widetilde{\mathbf{A}}$ is a parameter vector whose elements are fuzzy numbers. In the fuzzy regression function $\widetilde{\mathbf{A}}$ is the multidimensional fuzzy set (fuzzy relation) as the Cartesian product of fuzzy sets of fuzzy parameters

$$\widetilde{\mathbf{A}} = \widetilde{A}_0 \times \widetilde{A}_1 \times \ldots \times \widetilde{A}_n \tag{4}$$

with membership function in the form

$$\mu_{\widetilde{A}}(\mathbf{a}) = \bigcup_{i=1}^{n} \left\{ \mu_{\widetilde{A}_i}(a) \right\}, \ \mathbf{a} = (a_1, a_2, \ldots, a_n). \tag{5}$$

The shape of the membership function of fuzzy numbers output value of fuzzy linear regression model (1) is calculated by Zadeh's extensional principle Novák (1990) in the form

$$\mu_{\widetilde{Y}}(y) = \begin{cases} \bigcup\limits_{\mathbf{a}|\mathbf{a}t'=y} \mu_{\widetilde{A}}(\mathbf{a}) & ; \ \{\mathbf{a}|\mathbf{t}' = y\} = 0, \\ 0 & ; \ elsewhere. \end{cases} \tag{6}$$

Membership function $\mu_{\widetilde{A}_i}(a_i)$ is approximated in the form of triangular fuzzy numbers Ghorsray (1997), Novák (1990)

$$\mu_{\widetilde{A}_i}(a_i) = \begin{cases} 1 - \dfrac{|\alpha_i - a_i|}{c_i} & ; \ \alpha_i - c_i \le a_i \le \alpha_i + c_i, \\ 0 & ; \ elsewhere, \end{cases} \tag{7}$$

where α_i is the mean value (core) of fuzzy number $\widetilde{A}_i$ and c_i is half of the width of the carrier bearing $\widetilde{A}_i = \{\alpha_i, c_i\}$. The term of membership functions for the output fuzzy sets (3) can be written in the form Kacprzyk (1992)

$$\mu_{\widetilde{Y}}(y) = \begin{cases} 1 - \dfrac{|y - \alpha.x'|}{\sum_{i=1}^{n} c_i |x_i|} & ; \ \alpha.x' - \sum\limits_{i=1}^{n} c_i |x_i| \le y \le \alpha.x' + \\ & \quad + \sum\limits_{i=1}^{n} c_i |x_i|, \\ 0 & ; \ elsewhere, \end{cases} \tag{8}$$

2.3 Identification of Fuzzy Linear Regression Model

Fitness of linear regression fuzzy model to the given data is measured through the Bass-Kwakernaakss index H see Figure 2.

In the procedure of model identification the optimization procedure minimizes the vagueness of global fuzzy function through the minimization of sum of fuzzy regression coefficients vagueness

$$min \ J = min \sum_{i=1}^{n} c_i \tag{9}$$

under condition

$$h_j \le H, \ j = 1, 2, \ldots, m. \tag{10}$$

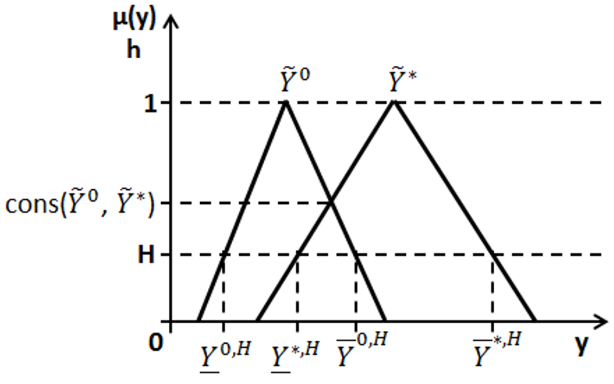

Fig. 2. Adequacy of linear regression fuzzy model

The fitness of estimated value to sampled value is done using α cut and α level set at the fitness $h = H$ (see Fig. 2)

$$\begin{aligned} Y_j^{0,H} &= \left[\underline{Y}_j^{0,H}, \overline{Y}_j^{0,H}\right], \\ Y_j^{*,H} &= \left[\underline{Y}_j^{*,H}, \overline{Y}_j^{0,H}\right]. \end{aligned} \tag{11}$$

We assume the good estimation of output value under the condition is fulfilled

$$max_y \left\{\mu_{\widetilde{Y}^0}(y) \wedge \mu_{\widetilde{Y}^*}(y)\right\} = Cons\left(\widetilde{Y}^0, \widetilde{Y}^*\right) \geq H. \tag{12}$$

The relation (12) is satisfied under the condition (Fig. 2)

$$\begin{aligned} \underline{Y}_j^* &\leq \overline{Y}_j^0, \; j = 1, 2, \ldots, m, \\ \underline{Y}_j^0 &\leq \overline{Y}_j^*, \; j = 1, 2, \ldots, m. \end{aligned} \tag{13}$$

Boundary of intervals $Y_j^{*,H}, j = 1, 2, \ldots, m$ we can express

$$\begin{aligned} \underline{Y}_j^{*,H} &= -(1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| + \alpha^T x_j, \\ \overline{Y}_j^{*,H} &= \quad (1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| + \alpha^T x_j. \end{aligned} \tag{14}$$

Next we can set the optimization problem for using of genetic algorithm

(1) minimization of fuzzy model vagueness

$$min\ J = min\sum_{i=1}^{n} c_{ij}, \; i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, m, \tag{15}$$

(2) subject to

$$\begin{aligned} \alpha^T x_j + (1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| &\geq y_j^0 + (1-H)\Delta y_j^0, \\ -\alpha^T x_j + (1-H)\sum_{i=1}^{n} c_{ij}\left|x_{ij}\right| &\geq -y_j^0 + (1-H)\Delta y_j^0. \\ c_{ij} &\geq 0. \end{aligned} \tag{16}$$

Quantification of models vagueness is formalized by calculating fuzzy intervals of fuzzy numbers estimated output values $\widetilde{Y}^*(x)$. Width of fuzzy numbers carriers are the interval in which the values of the output variables may lie with a defined grade of membership. On Figure 3 is graphically illustrated the course of a one-dimensional fuzzy linear regression function together with the appropriate possibility area of estimated fuzzy output $\widetilde{Y}^*$.

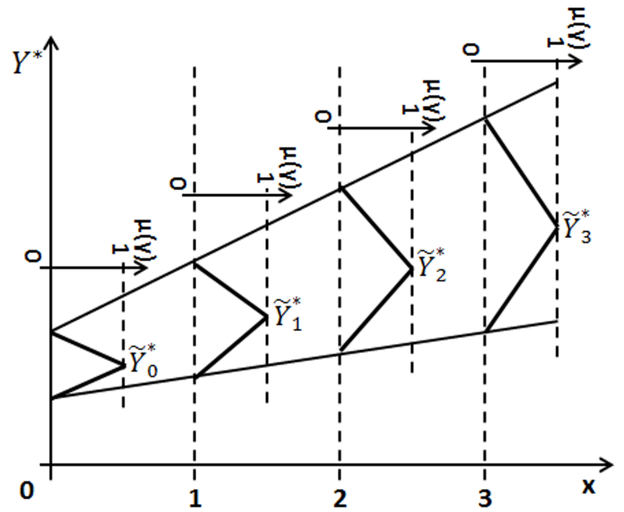

Fig. 3. One-dimensional fuzzy linear regression model

3. USAGE OF GENETIC ALGORITHM

As it was mentioned the classical used method of linear programming for identification of fuzzy regression coefficients was substituted by using of genetic algorithm (GA). The identification of fuzzy regression coefficients $\widetilde{A}_0, \widetilde{A}_1, \ldots, \widetilde{A}_n$, where $\widetilde{A}_i = \{\alpha_i, c_i\}$, was divided into two tasks

(1) the identification of the mean value (core) α_i of fuzzy number $\widetilde{A}_i$ and
(2) the identification of c_i as a half of the width of the carrier bearing $\widetilde{A}_i = \{\alpha_i, c_i\}$.

The tasks are solved by using genetic algorithm in series. First the identification of α_i and then the identification of c_i are done.

The sharp observed values y^0 are fuzzificated

$$\Delta y^0 = a y^0, \tag{17}$$

where $a \in \langle 0.02; 0.1\rangle$ or another value, but the value of a is defined by the expert. Then the fuzzy observed value is defined as

$$\widetilde{Y}^0 = \left\{y^0, \Delta y^0\right\}, \tag{18}$$

and the estimated fuzzy value $\widetilde{Y}^*$ analogously

$$\widetilde{Y}^* = \{y^*, \Delta y^*\}. \tag{19}$$

3.1 *Identification of the Mean Value (Core) α_i*

For the identification of the mean value (core) α_i of fuzzy number $\widetilde{A}_i$ the minimization of fitness function

$$min\ J_1 = min\frac{1}{J}\sum_{j=1}^{J}\left[y^0(x_j) - y^*(x_j)\right]^2, \tag{20}$$

by genetic algorithm is used.

3.2 *Identification of the Half of the Width of the Carrier Bearing c_i*

For the identification of c_i as a half of the width of the carrier bearing $\widetilde{A}_i$ the minimization of fitness function (9)

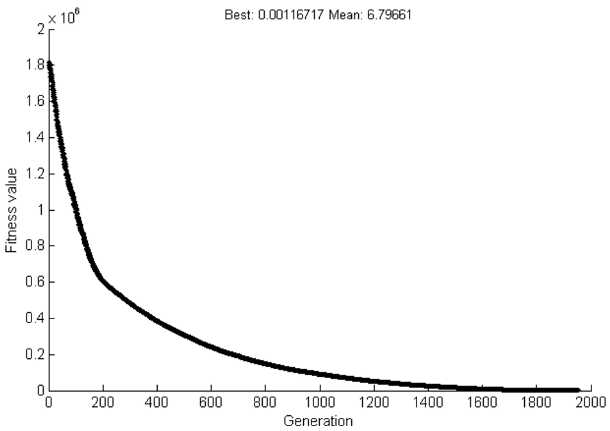

Fig. 4. Course of GA convergence

$$min\ J_2 = min \sum_{i=1}^{n} |c_i| \tag{21}$$

by genetic algorithm with three constraints (16) is used.

4. CASE STUDY

For proving of efficiency of proposed method, the two dimensional linear function in form

$$Y^0 = 1000 - 250x_1 + 430x_2 \tag{22}$$

was chosen. The set of Y^0 with ten members using (22) was created. For creating the set of Y^0 the values of x_1 and x_2 were chosen randomly from the standard uniform distribution on the open interval $(0, 1)$ but multiplied by random integer. For fuzzification of observed value $a = 0.1$ was used.

Then the minimization of fitness function J_1 (20) by embedded function of genetic algorithm in Optimtool in Matlab environment was used. The parameters of GA were elected as

- population type - double vector
- population size - 100
- scaling function - rank
- selection - stochastic uniform
- mutation function - constraint dependent
- crossover function - scattered
- migration - forward
- stop criterion - no changes in fitness function

The shape of convergence of values of minimization of fitness function J_1 is depicted in Figure (4). The outputs of the minimization by described GA are the estimated values of the mean values (cores) α_0, α_1 and α_2 of $\widetilde{A}_0$, $\widetilde{A}_1$ and $\widetilde{A}_2$.

The next step was to determine the c_0, c_1 and c_2 of $\widetilde{A}_0$, $\widetilde{A}_1$ and $\widetilde{A}_2$. For this task the minimization of fitness function J_2 (21) by GA was used with the same parameters as in task of determining of α_i.

As we now have the complete information to assemble the estimated fuzzy numbers $\widetilde{A}_0$, $\widetilde{A}_1$ and $\widetilde{A}_2$ we can define

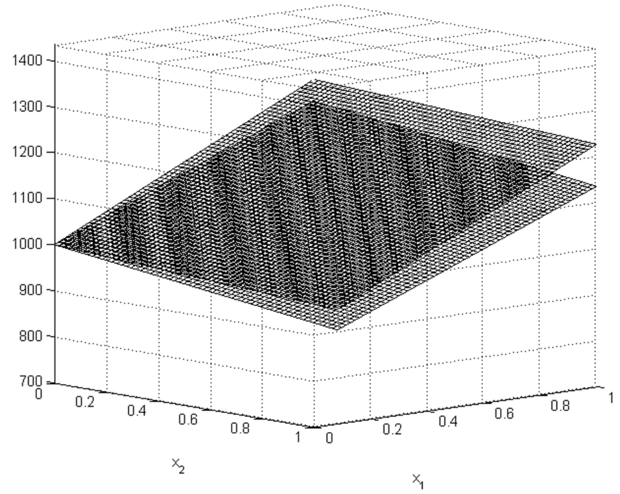

Fig. 5. Possibility area of two-dimensional fuzzy regression model

$$\begin{aligned} Y^*\left(y^*, \Delta y^*\right) &= \widetilde{A}_0\left(\alpha_0, c_0\right) + \widetilde{A}_1\left(\alpha_1, c_1\right) x_1 + \widetilde{A}_2\left(\alpha_2, c_2\right) x_2, \\ y^* &= \alpha_0 \qquad + \alpha_1 x_1 \qquad + \alpha_2 x_2, \\ \Delta y^* &= c_0 \qquad + c_1 x_1 \qquad + c_2 x_2. \end{aligned} \tag{23}$$

With knowledge of (23) we are able to create the surfaces, which are defined as the upper and lower boundary

$$\begin{aligned} \overline{Y}^* &= y^* + \Delta y^*, \\ \underline{Y}^* &= y^* - \Delta y^*. \end{aligned} \tag{24}$$

The area between the created lower and upper surface boundary could be called possibility area. For chosen linear regression function (22) the determined possibility area is shown in Figure (5).

5. CONCLUSION

Abstract mathematical models of complex systems are often not very adequate because they do not accurately reflect the natural uncertainty and vagueness of the real world. The suitable theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory. In the paper vague data as specialized fuzzy sets - fuzzy numbers are defined and it is described a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague parameters. Interval and fuzzy regression technology are discussed, the linear fuzzy regression model is proposed. To identify fuzzy regression coefficients of model instead of commonly used linear programming method Çetintav (2013) the effective genetic algorithm is applied Goldberg (1989). The two-dimensional numerical example is presented and the possibility area of vague model is graphically illustrated. Next research will be focused on development of fuzzy non-linear regression model with fuzzy output value Pokorný (1993) to have possibility to investigate and model vague non-linear systems.

ACKNOWLEDGEMENTS

This work has been supported by Project SP2013/168, "Methods of Acquisition and Transmission of Data in Distributed Systems", of the Student Grant System, VŠB - Technical University of Ostrava.

REFERENCES

A. Bardossy. Note on fuzzy regression. *Fuzzy Sets and Systems*, volume 37, pages 65-75, 1990.

J.J. Buckley, L.J. Jowers. Fuzzy Linear Regression I. *Studies in Fuzziness and Soft Computing*, volume 22, ISBN 978-3-540-76289-8. Springer, 2008.

B. Çetintav, F. Özdemir. LP Methods for Fuzzy Regression and a New Approach. E.Krause, Ed. *In Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, volume 22, ISBN 978-3-642-33041-4. Springer, 2013.

S. Ghorsray. Fuzzy linear regression analysis by symmetric fuzzy numbxer coefficients. *IEEE International Conference on Engineering Systems INES97*, 1997, ISBN 0-7803-3627-5.

D. Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning. ISBN 978-0201157673. Reading, MA: Addison-Wesley Professional, 1989.

B. Heshmaty, A. Kandel. Fuzzy Linear Regression and Its Application to Forecasting in Uncertain Environment. *Fuzzy Sets and Systems*, volume 15, pages 159-171.

H. Ishibushi, H. Tanaka. Identifications of Fuzzy Parameters by Interval Regression Model. *Electronics and Communications in Japan*, 73:volume 12, pages 19-27.

J. Kacprzyk, M. Fedrizzi (Ed.). Fuzzy Regression Analysis. *Studies in Fuzziness and Soft Computing*, ISBN-13: 978-3790805918. Publisher: Physica-Verlag HD, 1992.

R.E. Moore. Methods and Applications of Interval Analysis. *SIAM (Society for Industrial and Applied Mathematics)* Philadelphia, 1979.

V. Novák. Fuzzy množiny a jejich aplikace (in Czech). *Studies in Fuzziness and Soft Computing*, ISBN 80-03-00325-3. SNTL Praha, 1990.

O. Poleshchuk,E. Komarov. A fuzzy linear regression model for interval type-2 fuzzy sets. *NAFIPS 2012*, ISBN 978-1-4673-2336-9. Fuzzy Information Processing Society, 2012.

M. Pokorný. Fuzzy nelineární regresní analýza. *Doctoral Thesis (in Czech)* VUT Brno, FEL, Brno, 1993.

A.F. Shapiro. Fuzzy regression models. *In http://www.soa.org/library/research/actuarial-research-clearing-house/2006/january/arch06v40n1-ii.pdf*, (10.4.2013).

H. Tanaka, S. Uejima, K. Asai. Linear regression analysis with fuzzy model. *IEEE Transactions and Systems, Man and Cybernetics*, 12:6, 1982.