

ON MULTIDIMENSIONAL LINEAR MODELLING INCLUDING REAL UNCERTAINTY

Jana NOWAKOVA, Miroslav POKORNY

¹Department of Cybernetics and Biomedical Engineering, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava-Poruba, Czech Republic

jana.nowakova@vsb.cz, miroslav.pokorny@vsb.cz

Abstract. *The theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory. In the paper are defined vague data as specialized fuzzy sets - fuzzy numbers and there is described a fuzzy linear regression model as a fuzzy function with fuzzy numbers as vague regression parameters. To identify the fuzzy coefficients of model the genetic algorithm is used. The linear approximation of vague function together with its possibility area are analytically and graphically expressed. The suitable numerical experiments are performed namely in the task of two-dimensional fuzzy function modelling and the time series fuzzy regression analysis as well.*

Keywords

Complex systems, fuzzy linear regression, fuzzy number, fuzzy set, genetic algorithms, possibility area, vague property.

1. Introduction

Regression models are often used in engineering practice wherever there is a need to reflect more independent variables together with the effects of other unmeasured disturbances and influences. 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 [8].

The suitable theoretical background for abstract formalization of vague phenomenon of complex systems is fuzzy set theory. In the paper are defined vague data as specialized fuzzy sets - fuzzy numbers. Next, a fuzzy

linear regression model as a fuzzy function with fuzzy numbers as vague parameters is identified using the genetic algorithms.

2. Model Definition

2.1. Ordinary Linear Model

The ordinary linear regression model of the investigated system [11] is given by a linear combination of values of its input variables:

$$Y = A_0 + A_1x_1 + \dots + A_nx_n = A_0 + \sum_{i=1}^n A_ix_i, \quad (1)$$

where (x_1, \dots, x_n) are input variables and (A_0, A_1, \dots, A_n) are ordinary regression coefficients.

The conventional regression model is based on the assumption that the system characteristics is defined as sharp, precise and deviations between the observed and estimated values of dependent variables are results of observation errors. The origin of a deviation between the observed and estimated values of dependent variables may not be of significant extent caused by poor local variables of the system structure. The causes of these variations are in a not very sharp nature of the system parameters. Such fuzzy phenomenon must also be reflected in the fuzziness of the corresponding parameters of the model.

2.2. Uncertainty Interval Linear Model

The development of the indeterminate regression model is the development of the model of vagueness, using the formalization of uncertainty rather than nu-

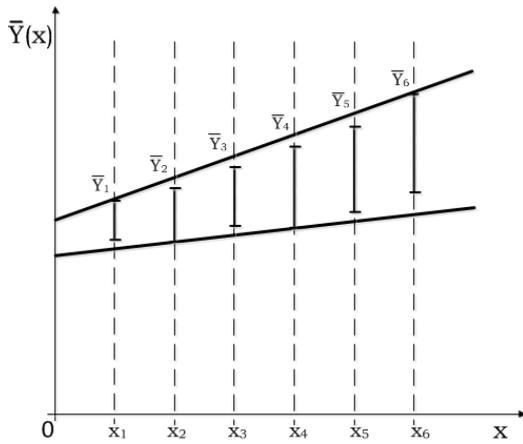


Fig. 1: Ordinary one-dimensional linear interval regression model.

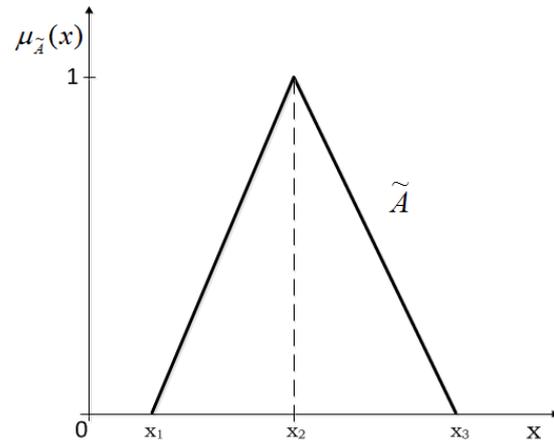


Fig. 2: Triangular membership function of fuzzy number \tilde{A} .

merical intervals [3], [9]:

$$\bar{Y} = \bar{A}_0 + \bar{A}_1x_1 + \dots + \bar{A}_nx_n = A_0 + \sum_{i=1}^n \bar{A}_ix_i. \quad (2)$$

Regression coefficients are numeric interval:

$$\bar{A}(a_i + e_i, a_i - e_i) \equiv \bar{A}_i(a_i, e_i), \quad (3)$$

where a is the middle of the interval is and e is half of its width. For one dimensional function:

$$\bar{Y}(x) = \bar{A}_0 + \bar{A}_1x = (a_0, e_0) + (a_1, e_1)x, \quad (4)$$

the interval regression model is depicted in Fig. 1.

2.3. Interval Model Fuzzification

Regression models reflecting the vagueness of the modelled systems are called fuzzy regression models [10], [1] and [8]. The indeterminate nature of the fuzzy regression model is represented by the fuzzy output values and the fuzzy regression coefficients in the form of specialized fuzzy sets - fuzzy numbers. The shape of fuzzy linear regression model is given by:

$$\tilde{Y} = \tilde{A}_0x_0 + \tilde{A}_1x_1 + \dots + \tilde{A}_nx_n = \sum_{i=0}^n \tilde{A}_ix_i, \quad (5)$$

where $x_0 = 1$ and $(\tilde{A}_0, \tilde{A}_1, \dots, \tilde{A}_n)$ are fuzzy regression coefficients - fuzzy sets. The fuzzy set \tilde{A} is defined as image, which assigns to every element x of universum X number $\mu_{\tilde{A}}(x) \in \langle 0, 1 \rangle$ as a degree of its membership function \tilde{A} [6], [7]:

$$\tilde{A} = \{x, \mu_{\tilde{A}}(x) \mid x \in X\}; \mu_{\tilde{A}}(x) \geq 0, \forall x \in X. \quad (6)$$

At least piecewise continuous function $\mu_{\tilde{A}}(x) = f(x)$ is called membership function, which defined

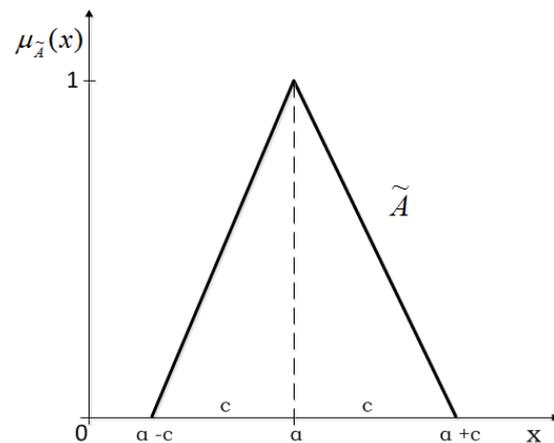


Fig. 3: Membership functions of fuzzy regression coefficients.

fuzzy set \tilde{A} conclusively. The membership function is usually in engineering praxis approximated by broken line (Fig. 2).

Triangular fuzzy set \tilde{A} then formalizes uncertain number (fuzzy number) “about x_2 ”. The degree of uncertainty of number x_2 is defined as the width of the carrier bearing of the fuzzy set \tilde{A} as closed interval $\langle x_1, x_3 \rangle$ (Fig. 2). Parameters of that fuzzy sets constitute structured vector of values of breaking point $[x_1, x_2, x_3]$. Using this vector fuzzy sets are computer-formalized.

In the fuzzy regression model, the fuzzy regression coefficients (fuzzy numbers) \tilde{A} are defined using its triangular shape membership function $\mu_{\tilde{A}}(x)$ (Fig. 3), where α is the mean value (core) of fuzzy number \tilde{A} and c is a half of the width of the carrier bearing $\tilde{A}_i = \{\alpha, c\}$. The output variable \tilde{Y} of fuzzy regression model (Eq. (5)) is fuzzy number defined using the triangular membership function (similar Fig. 3), where β is the mean value (core) of fuzzy number \tilde{Y} and d is a half of the width of the carrier bearing $\tilde{Y} = \{\beta, d\}$.

Fuzzy regression modelling (Eq. (5)) requires operation with fuzzy numbers. For this types of operations it is needed to use relations of fuzzy arithmetic with usage of extensional principle.

3. Fuzzy Arithmetic Application

Extensional principle (principle of extension) allows to transfer operation over ordinary numbers to operation over fuzzy numbers. It allows to create fuzzy arithmetic for computing with imprecise (fuzzy) numbers [2].

Let consider universum U and V and function f , which maps U to V , i.e.:

$$f : U \rightarrow V \tag{7}$$

and fuzzy set $A \subseteq U$. Fuzzy set A then in V induces fuzzy set, whose membership function is defined by relation:

$$\mu_f(v) = \begin{cases} \sup_{f(u)=v} \mu_A(u) & \text{if } \exists u \in U \text{ such that} \\ & v = f(u), \\ 0 & \text{elsewhere} \end{cases} \tag{8}$$

Using the extension principle fuzzy numbers arithmetic can be defined [6]. Take the case of the sum of two fuzzy numbers m ("about m ") and n ("about n "). These relations are needed for calculation of output value \tilde{Y} (Eq. (5)):

$$\mu_{\tilde{m} \otimes \tilde{n}} = \sup_{x,y/z=x \cdot y} \min(\mu_{\tilde{m}}(x), \mu_{\tilde{n}}(y)), \tag{9}$$

$$\mu_{\tilde{m} \oplus \tilde{n}} = \sup_{x,y/z=x+y} \min(\mu_{\tilde{m}}(x), \mu_{\tilde{n}}(y)). \tag{10}$$

4. Fuzzy Model Identification

4.1. Identification Method Description

Fuzzy number Y_j^0 is mentioned of a triangular type. The values d_j can be calculate by the formula:

$$d_j = \frac{1}{2} |y_{j+1}^0 - y_{j-1}^0|, \tag{11}$$

where $j = 1, 2, \dots, m$ is the number of observations. Finding values α_i and c_i as searched parameters of fuzzy regression coefficients \tilde{A}_i (Fig. 3) is defined as an optimization issue.

Fitness of the linear regression fuzzy model to the given data is measured through the Bass-Kwakernaaks's index H , Fig. 4, [4]. Adequacy of the

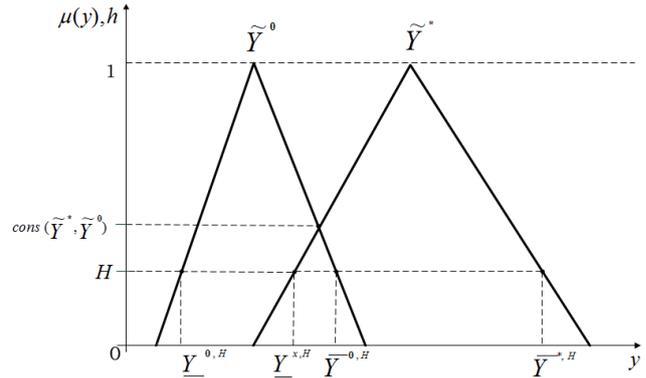


Fig. 4: Adequacy of linear regression model.

observed and estimated values is conditioned by the relation (Eq. (12)), the maximum intersection (consistency) of two fuzzy sets, the estimated \tilde{Y}_j^* and the examined \tilde{Y}_j^0 ; must be higher than the set value H :

$$\begin{aligned} \max_y \{ \mu_{\tilde{y}^0}(y) \wedge \mu_{\tilde{y}^*}(y) \} = \\ = \text{Cons}(\tilde{Y}^0, \tilde{Y}^*) \geq H. \end{aligned} \tag{12}$$

Only if the condition (Eq. (12)) is fulfilled we assume good estimation of the observed output value \tilde{Y}_j^0 . The relation (Eq. (13)) is satisfied under the condition (Fig. 4):

$$\underline{Y}^{*,H} \leq \bar{Y}^{0,H}, \tag{13}$$

$$\underline{Y}^{0,H} \leq \bar{Y}^{*,H}. \tag{14}$$

Consider the determined level H the boundary of intervals $Y^{*,H}$ and relations (Eq. (13), Eq. (14)) can be expressed:

$$\underline{Y}^{*,H} = -(1-H) \sum_{i=0}^n c_i |x_i| + \alpha x_i, \tag{15}$$

$$\bar{Y}^{*,H} = (1-H) \sum_{i=0}^n c_i |x_i| + \alpha x_i. \tag{16}$$

According to the Fig. 3, it can be written:

$$\underline{Y}^{*,0} = y_j^0 (1-H) d^0, \tag{17}$$

$$\bar{Y}^{*,0} = -y_j^0 (1-H) d^0. \tag{18}$$

The conditions (Eq. (13), Eq. (14)) can be written in the shape:

$$\begin{aligned} \sum_{i=0}^n \sum_{j=1}^m a_{i,j} x_{i,j} + \\ + (1-H) \sum_{i=0}^n \sum_{j=1}^m c_{i,j} |x_{i,j}| \geq \\ \geq y_j^0 + (1-H) d^0, \end{aligned} \tag{19}$$

$$\begin{aligned}
 & - \sum_{i=0}^n \sum_{j=1}^m a_{i,j} x_{i,j} + \\
 & + (1 - H) \sum_{i=0}^n \sum_{j=1}^m c_{i,j} |x_{i,j}| \geq \\
 & \geq -y_j^0 + (1 - H) d^0, \\
 & c_{i,j} \geq 0.
 \end{aligned} \tag{20}$$

The requirement on adequacy of the estimated and observed values will be complemented by the requirement on minimum possible total uncertainty of the identified fuzzy regression function:

$$\sum_{i=0}^n \sum_{j=1}^m c_{i,j} \rightarrow \min, i = 0, 1, \dots, n, \tag{21}$$

$$j = 1, 2, \dots, m,$$

where $i = 0, 1, \dots, n$ is the number of input values of the regression function and $j = 1, 2, \dots, m$ is the number of observations.

4.2. Model Parameters Optimization

Then we can set the optimization problem:

- minimization of fuzzy model vagueness (Eq. (21)),
- under the condition (Eq. (12)).

To solve the minimization problem under the condition, many authors use the linear programming method. Nevertheless, in this paper we use the genetic algorithm method to solve this problem [4]. Mainly, the reason is that the authors are oriented to use unconventional methods of artificial intelligence in order to prove their quality and efficiency in solving complex tasks. Genetic algorithms are a representative of evolutionary methods; their higher computational complexity is nowadays eliminated by high-performance computing. They are widely used in the search for optimal solutions. They can be well used for the identification of fuzzy regression models where they deal with the task of finding the optimal fuzzy regression coefficients as triangular fuzzy numbers. The identification of fuzzy regression coefficients – fuzzy numbers: $\tilde{A}_0, \tilde{A}_1, \dots, \tilde{A}_n$, was divided into two tasks:

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

The tasks are solved by using the genetic algorithm in series. First the identification of α_i and then the identification of c_i are done. Thus, the optimization of the fuzzy linear regression model is a two-step process when two genetic algorithms, designated G1 and G2,

are used. For the identification of the mean value (core) α_i of fuzzy number \tilde{A}_i the minimization of the fitness function J_1 is defined in the form:

$$\min J_1 = \min \frac{1}{m} \sum_{j=1}^m (y_j^0 - \beta_j)^2, \tag{22}$$

and the genetic algorithm GA1 is used. For the identification of as a half of the width of the carrier bearing \tilde{A}_i the minimization of the fitness function J_2 is defined in the form:

$$\min J_2 = \min \sum_{j=1}^m \sum_{i=0}^n |c_{j,i}|, \tag{23}$$

and the genetic algorithm GA2 with two constraints (Eq. (21)) is used. Minimization of the fitness function J_2 is based on the previous identification of the role of the mean value (core) α_i and uses the already identified values of α_i for determining the width of the carrier bearing α_i . The value of $H = 0.5$ is expertly determined in the next part of paper.

4.3. Genetic Algorithms Utilization

As mentioned before, the classical method of linear programming used for the identification of fuzzy regression coefficients [11] was substituted by using a genetic algorithm (GA) [4]. Mainly, the reason is that the authors are oriented to use unconventional methods of artificial intelligence in order to prove their quality and efficiency in solving complex tasks. Genetic algorithms are a representative of evolutionary methods; their higher computational complexity is nowadays eliminated by high-performance computing. They are widely used in the search for optimal solutions. They can be well used for the identification of fuzzy regression models where they deal with the task of finding the optimal fuzzy regression coefficients as triangular fuzzy numbers.

The identification of fuzzy regression coefficients – fuzzy numbers: $\tilde{A}_0, \tilde{A}_1, \dots, \tilde{A}_n$, was divided into two tasks:

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

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

As it was mentioned before, the genetic algorithm is an unconventional optimisation method, which is used for minimization of the target optimization function

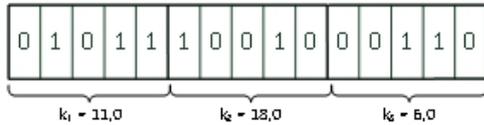


Fig. 5: Binary coded parameters in structure of chromosome.

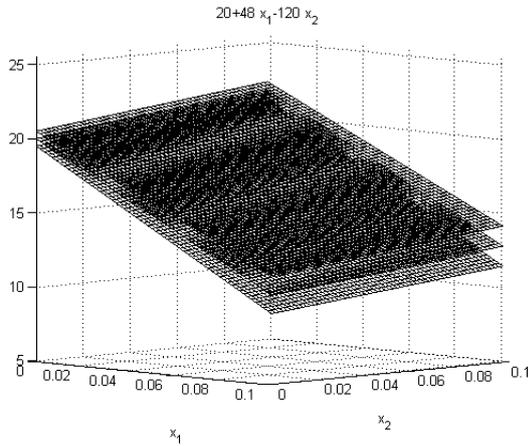


Fig. 6: Adequacy of linear regression model.

(fitness function). It is used instead of conventional methods, such as the linear programming method.

GA is a seeking procedure which looks for the best solutions according to the fitness function based on the processes observed in Nature, on the principle of natural selection and genetic laws, i.e. selection, crossover and mutation. The basis of GA is to use a character string, also called a chromosome, in which parameters of an optimized model are stored. An example of a chromosome which is composed of three parameters k_1 , k_2 and k_3 expressed by three 5-bit binary words is shown in Fig. 5.

Individual bits represent the string of chromosome genes; at the particular optimization step their specific values represent binary codes of three parameters of the model. Each chromosome is evaluated by the size of its fitness function, the value of which determines the distance of a solution (which is represented by a particular chromosome) from the optimal solution.

The set of evaluated n-chromosomes represents one population, the best individuals (solutions) of which are genetic operations of selection and are picked out for follow-up populations. Selected individuals are subjected to genetic operations of crossover, in which two individuals (parents) interchange gene circuits and generate two new chromosomes - offspring with different combinations of k_1 , k_2 and k_3 . The descendants, who were generated this way, then form a new population where individuals (solutions) appear to have better characteristics (better fitness function value) than the best individual in the population of parents. Then, an

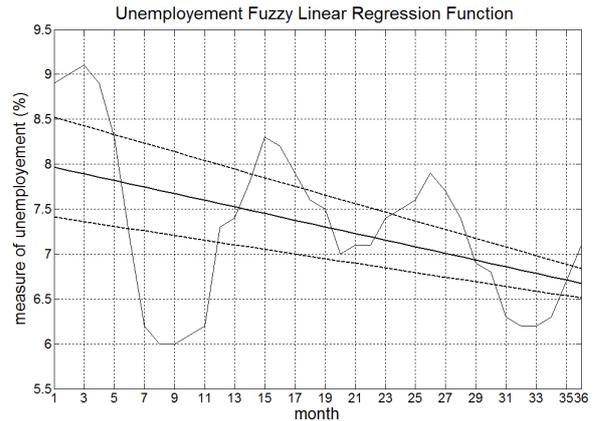


Fig. 7: Fuzzy linear regression function for unemployment in the Czech Republic.

appropriate follow-up offspring population is created (solution step, iteration) and the genetic crossing procedure is repeated. Good convergence for finding an optimal individual (solution) is supported by a genetic operation - mutation. The features of genetic operations as selection, crossover and mutation are defined by setting the internal parameters of the genetic algorithm in the way that the convergence of a solution to optimum is favorable.

The procedure of the genetic algorithm is usually finished by a solution step (population) in which the values of the fitness function of the best current individual and the best individual in the last step vary less than the specified limit (stop-criterion). As an optimal solution is then determined the best chromosome of the last population. Corresponding (coded) parameter values are used in the optimal model.

The main tasks while designing a genetic algorithm are the method of encoding the optimized parameters to a chromosome string and the definition of its fitness function. Optimization of the fuzzy linear regression model is a two-step process when two genetic algorithms, designated G1 and G2, are used.

5. Fuzzy Model Verification

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

$$Y^0 = 20 + 48x_1 - 120x_2, \tag{24}$$

was chosen. The set of Y^0 with ten members using (Eq. (19), Eq. (20)) 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. The result can be seen in Fig. 6, [5].

The usage in economic area is depicted in Fig. 7. As an input data the unemployment rate in the Czech republic for years 2009 to 2011 was used.

6. 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 could be fuzzy set theory, which was shortly described. 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. It is used the effective genetic algorithm instead of commonly used linear programming method for identification of fuzzy regression coefficients of the model. The two-dimensional numerical example and practical economic usage are presented and the possibility area of vague model is graphically illustrated. Next research will be focused on model vague non-linear systems.

Acknowledgment

This work has been supported by Project SP2014/156, "Microprocessor based systems for control and measurement applications", of the Student Grant System, VSB-Technical University of Ostrava.

References

- [1] ABDALLA, A. and J. BUCKLEY. Monte Carlo Methods in Fuzzy Non-Linear Regression. *New Mathematics and Natural Computation*. 2008, vol. 4, iss. 2, pp. 123-141. ISSN 1793-0057. DOI: 10.1142/S1793005708000982.
- [2] HANSS, M. *Applied fuzzy arithmetic: an introduction with engineering applications*. New York: Springer-Verlag, 2005. ISBN 35-402-4201-5.
- [3] ISHIBUCHI, H. and H. TANAKA. Identification of fuzzy parameters by interval regression models. *Electronics and Communications in Japan (Part III Fundamental Electronic Science)*. 1990, vol. 73, iss. 12, pp. 19-27. ISSN 1520-6440. DOI: 10.1002/ecjc.4430731203.
- [4] REEVES, C. R. *Genetic algorithms: principles and perspectives: a guide to GA theory*. Boston: Kluwer Academic, 2003. ISBN 14-020-7240-6.
- [5] MATHWORKS. *MATLAB and Simulink for Technical Computing* [online]. Available at: <http://www.mathworks.com>.
- [6] MOHAN, C. *Fuzzy set theory and fuzzy logic: an introduction*. Tunbridge Wells: Anshan, 2009. ISBN 18-482-9025-X.
- [7] NOVAK, V. and I. PERFILIEVA. *Mathematical Principles of Fuzzy Logic*. Boston: Springer US, 1999. ISBN 978-146-1552-178.
- [8] NOWAKOVA, J. and M. POKORNY. Fuzzy Linear Regression Analysis. In: *12th IFAC Conference on Programmable Devices and Embedded Systems*. Velke Karlovice: VSB-Technical University of Ostrava, 2013, pp. 245-249. ISBN 978-3-902823-53-3. DOI: 10.3182/20130925-3-CZ-3023.00079.
- [9] POLESHCHUK, O. and E. KOMAROV. A fuzzy linear regression model for interval type-2 fuzzy sets. In: *Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*. Berkeley: IEEE, 2012, pp. 1-5. ISBN 978-1-4673-2336-9. DOI: 10.1109/NAFIPS.2012.6290970.
- [10] SHAPIRO, A. F. *Fuzzy Linear Regression*. Penn State University, 2005.
- [11] A. TANSU. *Fuzzy Linear Regression: Fuzzy Regression*. New York: LAP LAMBERT Academic Publishing, 2012. ISBN 978-38-44-38442-0.

About Authors

Jana NOWAKOVA was born in 1987 in Trinec. She received her M.Sc. in Measurement and Control from VSB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science in 2012. Nowadays she continues her studies in Technical Cybernetics. She is interested in addition to fuzzy modelling, also in statistical data processing in cooperation with University Hospital Ostrava.

Miroslav POKORNY was born in 1941. He received his M.Sc. in High-frequency communication technologies from Technical University Brno, Faculty of Electrical Engineering in 1963. His Ph.D. he received in 1994 in Cybernetics and Informatics from Technical University Brno, Faculty of Electrical Engineering and Informatics. Since 1993 he worked as associate professor and since 1998 he has worked as professor at VSB-Technical University of Ostrava,

Faculty of Electrical Engineering and Computer Science, Department of Cybernetics and Biomedical Engineering. Between 1964 and 1971 he worked as researcher in Research institute of metallurgy VZKG

Ostrava-Vitkovice and between 1972 and 1992 as researcher and head of development in VUHZ-Research Institute of metallurgy.