sciendo

LATVIAN JOURNAL OF PHYSICS
AND TECHNICAL SCIENCES
2021, N 4

DOI: 10.2478/lpts-2021-0034

METHODS OF FUZZY SET IN SIMULATION FOR PREDICTING UNOBSERVED STATES OF THE ECOLOGICAL AND GEOENGINEERING SYSTEMS

I. Yeremeyev[1], A. Dychko[2*], V. Kyselov[1], N. Remez[2],
S. Kraychuk[3], N. Ostapchuk[3]

[1] Taurida National V.I. Vernadsky University
33 Ivana Kudri, Kyiv, 04000, UKRAINE
[2] Institute of Energy Saving and Energy Management,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
37 Peremohy Ave., Kyiv, 03056, UKRAINE
[3] Department of Economic Cybernetics,
Rivne State University of Humanities
12 Stepana Bandery Str., Rivne, 33000, UKRAINE
*e-mail: aodi@ukr.net

The present paper provides the ways of implementing methods of fuzzy set approach, which contributes to an increase in the accuracy, efficiency and functional flexibility of the complex control and recognition monitoring systems for the environmental and geoengineering system simulation. They are based on data mining methods and may be implemented with the help of intellectual technologies, including the combination of model pluralism, membership functions, methods of nearest neighbour, results of fractal and chaos theories, methods of ensuring robustness and retrospective analysis of the decision tree for success in decision making in similar situations.

It is proposed to use the model pluralism to explain a particular information process, which uses a number of adequate models, describing the behavior of objects in the case where each of the model reflects its behavior objectively, but under different circumstances, which are difficult to consider a priori in real time when choosing an adequate model. It is shown that the method of the nearest neighbour should be used if it necessary to identify causal relationships and predict further development of the environmental safety events.

Keywords*: Fuzzy set, environmental simulation, geoengineering system, nearest neighbour method, predicting unobserved states.*

1. INTRODUCTION

Modern complex control and recognition systems of the environmental safety management and geoengineering systems monitoring, as a rule, operate in the context of incomplete and fuzzy information, which often affects their effectiveness. In addition, changes and fluctuations in the performance of systems that are probabilistic and fuzzy must be taken into account in the risk management [1]. To improve the efficiency, accuracy and functionality of such systems, different software and hardware solutions are used, that is, ensuring the sufficiency, completeness, timeliness and reliability of information on the basis of which those or other solutions are proposed by generating additional data from data mining methods [2]–[5]. For the identifying, processing, and data analyzing, the intelligence systems are used. Data, generated during the course of operations, including data generated from processes and the additional data, can be structured, semi-structured, or unstructured depending on the nature and conditions of the data use. Due to the amount of data generally generated during the course of operations, intelligence systems are commonly built on top of and utilize a data warehouse [6]. All these make the process of decision making more complicated, unclear and uncertain with appropriate consequences.

The method of "the nearest neighbour" [7] should be used if it is necessary to identify causal relationships and predict further development of the events. The method is based on the estimation of the states of the "nearest neighbours", which are within the accepted limits of the confidence interval 2σ, i.e., the uncertainty interval. Characteristics of the researched object at the point of interest are compared with the data (characteristics) in the nearest (in time or in space) neighbouring points of the object. If a consistent change in the status indicators is observed at all adjacent (right and left) points within the standard deviation, one can assume that there is a certain trend that can be trusted.

The k-nearest neighbour rule is used for an adaptive process monitoring to solve the problems arising from nonlinearity, insufficient training data, and time-varying behaviours [8]. To simplify the process of calculating and making online monitoring possible the rule of a distance-based update is developed.

As it is rightly stated in [9], one cannot consider information without considering any situations of uncertainty. Analysis of the potential impact of the uncertainty of input variables on the performance of the wastewater volume forecasting model shows that a significant influence of the uncertainty of the input variables is demonstrated by water consumption, humidity, rainfall, while duration of sunshine, rainfall depending on certain conditions have a relatively weak impact of uncertainty of input variables [10].

Machine Learning models, including Deep Neural Networks, Convolutional Neural Networks, naïve Bayes and k-Nearest-Neighbour, are proposed to forecast biological species behaviour based on traits, and infer trait connections responsible for species interactions. It is demonstrated that such models are more flexible and informative compared to usual linear models used in ecological research [11].

A model-based diagnosis framework in which a Bayesian approach is used is proposed in Fault Augmented Model Extension work [12]. Fault diagnosis using a Bayesian approach is based on computing

a set of probability density functions, a process that is usually intractable for any reasonably complex system. Approximate Bayesian Computation helps bound the numerical and computational complexity. Such an approach gives a possibility to create probability distributions of possible outcomes and then compare those distributions against actual observations to perform parameter estimation.

The possible imperfections of the proposed methods of fuzzy set approach used for analysis of complex systems, processes and ecological data include the development of the only, more perfect model for simulation.

The above-mentioned aspects require a multi-vector analysis of problems, coverage of these problems from different sides in order to facilitate understanding (for example, based on analogies) and finding more successful ways of informational (including visual) provision of management procedures based on system analysis and data mining. Moreover, the application of such methods in practice is limited by economic and management sectors, or data processing for some ecological research. However, environmental issues are the most complex, fuzzy and uncertain that demand the innovative system approach for their decision.

With the aim of the problem elimination, it is proposed to use the intellectual technologies, including the combination of model pluralism with methods of fuzzy sets, membership functions, methods of nearest neighbours, results of fractal and chaos theories, methods of ensuring robustness and retrospective analysis of the decision tree for successful (by results) choices/decisions to be used in similar situations.

2. METHODS OF FUZZY SET APPROACH

Model pluralism, as one of the data mining methods, to explain (understand) a particular information process uses a number of adequate (relevant) models (by nature, usually empirical or semi-empirical) to describe the behaviour of objects in the case where each model reflects its behaviour objectively, but under conditions, difficult to consider a priori in real time.

The effect of uncertainty on the result of process can be demonstrated by the example of a biochemical wastewater treatment or anaerobic digestion system [5]. Such a biochemical technology is a classic model of a system operating under uncertainty: its states (indicators of the degree of contamination of wastewater and treated water by various pollutions) are usually determined not in real time, as well as there is a close, but unambiguous, relationship among external factors (temperature of the environment, atmospheric pressure, rainfall intensity, time during which purification processes take place, etc., and there is also a significant time lag between different events and changes in states of the system). Management of such a system should involve:

- monitoring of the states of the system and the degree of possibility of such states;
- determination of conditions that are impossible according to additional information (for example, wastewater indicators cannot be better or the same as natural water indicators);
- prediction of the states that are not actually observed, but fundamentally possible.

The values of the amplitudes of the actual output data of monitoring of the ecosystem or complex geoengineering system (%) can be presented as a function of time for the three selected models-standards of the response to the calibrated input parameter (for example, a single pulse) (Fig. 1).

Such a model is characterised by its distance (Euclidean metrics) d_E from the distribution of really determined values of the amplitudes y_i and the distribution of values y_i^m corresponding to the m-th model:

$$d_E = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} y_i m} \quad (1)$$

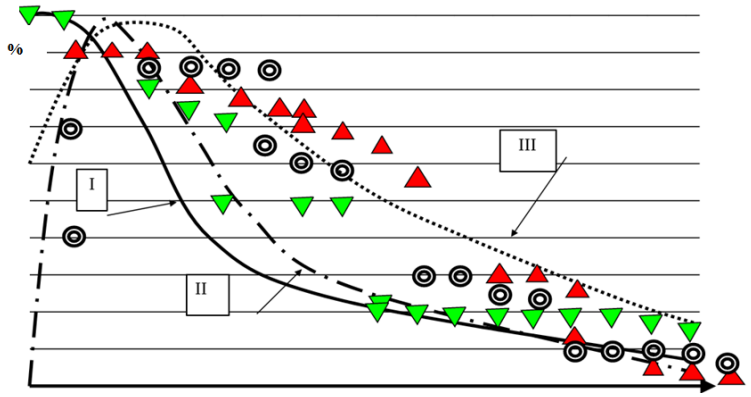

Fig. 1. Values of the amplitudes of output data (%) as a function of time for the three selected models-standards of response to the input parameter: n – the number of points at which the amplitudes are determined, m – the model number (I, II, III), ◎, ▲, ▼, – the value of the actually measured parameter (respectively, for the three variants of the real systems).

The use of multiple models instead of the only one allows supplementing the existing "reality" with those "nuances" inherent in alternative models, and provides the more comprehensive assessment of the situation (including behavioral motives) and informed decision making as a rule.

The choice of the optimal model M_{opt} corresponds to the condition:

$$M_{opt} = M\ (min\ \{d^{I}_{E},\ d^{II}_{E},\ d^{III}_{E}\}). \quad (2)$$

The following heuristics is proposed to be used to estimate the trend by the "nearest neighbours" method:

$$IF\ [\ \underline{MSTAB}\ ]\ AND\ [\ \underline{SIGN\Delta ST}_{i \pm j}\ EQ],\ THEN\ [\ \underline{SOT}\ ] \quad (3)$$

where *MSTAB* – the situation when the measured value does not exceed the limits of the standard deviation; $SIGN\Delta ST_{i \pm j}\ EQ$ – the observation result, indicating that the sign of the state change at the current measurement at all points to the left and right of *i*, i.e., at points from *i* to *i-j* and from *i* to *i+j*, is the same relative to the state at the same points during the previous measurement or in the spatial distribution; *SOT* – a certain trend is observed (see Fig. 2).

Increase of the state Δx_i in case of inconsistent movement of the state indicators at points $x_1 \ldots x_5$ has index 1 (that is Δx_{i1}), and in case of the consistent movement – index 2 (that is Δx_{i2}).

The method of comparing the characteristics of the nearest neighbours should not be considered a tool for improving reliability, but it allows estimating at least plausible trends in areas where more exact information is missing (or rather, hidden due to being in the range of uncertainty). However, such a technique in itself broadens boundaries of "vision" of the problem to some extent.

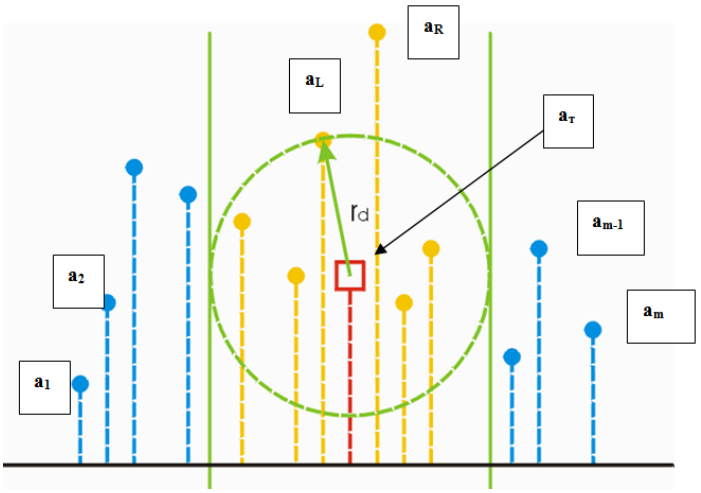

Fig. 2. Graphical interpretation of the nearest neighbour method:
a_1, a_2, a_{m-1}, a_m – the measured values of the parameter, r_d – the radius of the boundaries containing the trend values, a_l – the measured value at the trend boundary, a_R – the measured value beyond trend, a_T – the defined value of the parameter.

3. PREDICTING UNOBSERVABLE STATES

Predicting unobservable states is one of the main goals of data mining. For the prediction of the states of the ecosystem that are not actually observed, but fundamentally possible, a nonzero degree of possibility $f_M(c)$, less than the minimum degree of possibility $f_M(\alpha)$ calculated for the observed states, is given.

For example,

$$f_M(c) = 0.5 \min_{\alpha} f_M(\alpha), \tag{4}$$

or

$$D_p\left[f_M(c), f_M(\alpha)\right] = \left\{ \frac{1}{n-1} \sum_{c \in C} \left[f_M(c) - f_M(\alpha)\right]^p \right\}^{1/p} \subseteq 0.5 \min_{\alpha} f_M(a) \tag{5}$$

where p – a parameter of the distance function D_p (for the Euclidean distance $p = 2$).

The fact that there is a possibility of predicting unobservable states assumes the connection among the event, the phenomenon and the state of the system. Such a connection does exist, but it is not straightforward and allows only evaluating at a qualitative level the possible states of the system, with a significant (up to a few tens of percent) error.

If we compile a table containing:

- the strata (boundaries within which wastewater pollution indicators may be observed, with appropriate estimates of the probability of observation);
- the set of states capable of observing a measure of zero (boundaries beyond which the observed variables never reach);
- the computed possible states (but not such that can actually be observed);
- the corresponding degrees of feasibility of the state realisation and the probable consequences of this realisation, it is possible to create the conditions for efficient management of processes based on heuristics.

We propose to determine the augmentation of the Euclidean distance (1) between pairs of actually observable states that occur nearby each other at a given time interval, as well as the probability and the possibility of transition from one state to another and the driving forces (internal and external) contributing to it.

The definition of the fundamentally possible states that cannot be recorded online but that may affect the system and its indicators globally, as well as the evaluation their feasibility and expected consequences are proposed further. Here, in the first step, it is necessary to make the definitions on the basis of available data (data from observation of the current process at a certain interval of monitoring). Then, at different levels of refinement, the best hypotheses regarding the feasibility estimates of those or other states of the generalised system are determined. The assumptions about the effect of these hypotheses on the real properties of the researched variables are formed then (these assumptions are formed on the basis of relevant experimental characteristics and specific functions). Finally, the given generalised limit is supplemented (or replaced) by the limits reconstructed with the help of better hypotheses, and each is associated with a degree of confidence. When using only the information contained in the data, the proposed approach allows including in the estimated uncertainty (generalised limit) some characteristics that cannot be established with the real observable data, i.e., it is possible to predict or update, with an estimated degree of certainty, the states of variables not included in the forecast or updated in the observation data.

If the measure of increasing the confidence *MB* to the hypothesis *h* based on the observation of the output *e*:

$$MB[h, e] = \{P(h \mid e) - P(h)\}/(1 - P(h)), \quad (6)$$

where $P(h|e)$ – conditional probability *h* with known *e*, and $P(h)$ – expert evaluation of probability for the specified time interval, then the degree of confidence increasing *MD* relatively *h* may be presented as

$$MD[h, e] = \{P(h) - P(h \mid e)\}/P(h), \quad (7)$$

and the factor of uncertainty *CF* may be presented as

$$CF[h, e] = MB[h, e] - MD[h, e]. \quad (8)$$

The values *MB*, *MD* and *CF*, obtained for every specific event, are placed in the table, which may be used to form the system control heuristics for operation under uncertainty conditions. These heuristics

make it possible to improve the quality of biochemical sewage treatment under conditions of uncertainty and action of the factors which are evaluated weakly.

The above-mentioned situation may be illustrated using the data concerned with quality of wastewater treatment at Zhytomir (Ukraine) treatment facilities (see Fig. 3 and Table 1, created on the basis of Fig. 3).

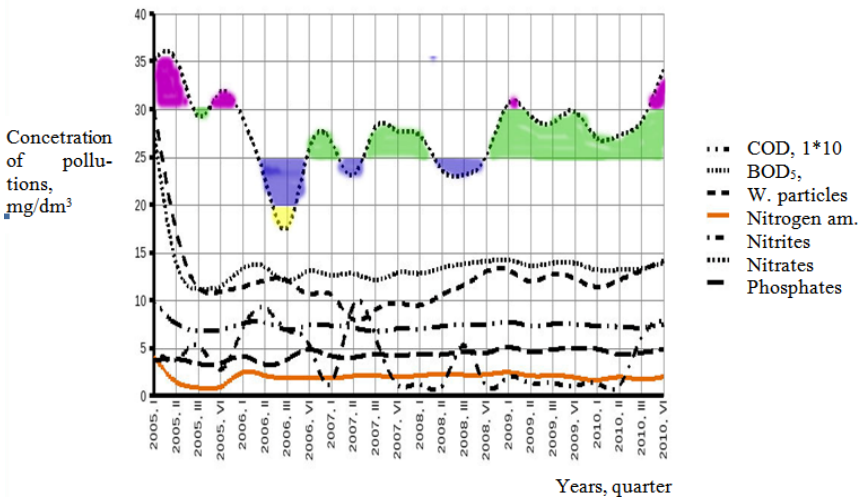

Fig. 3. Quality of wastewater treatment at the treatment facilities.

Decision about control action is made on the basis of selection (with the Monte-Carlo approach) of some contamination states (limits of existence, which in linguistic form may be presented as "Great value" (G), "Medium value" (M), "Low value" (L) and "Natural value" (N) (when no control actions are required) accounting the possibility of probability of its realisation. The spectrum of virtual contaminants obtained thus allows for the appropriate regulatory actions to be taken to minimise these contaminants. Thus, if the required ratio of biogenic elements in the aerotank is BOD_{total}:N:P=100:5:1, heuristics for the implementation of the feeding of active sludge with nitrogen and phosphorus compounds has the following form:

IF [{(BOD_M) AND (CNSW_M) AND (CPSW_M)},
OR {(BOD_M) AND (CPSW_M)}], THEN {RAS_N}, (9)

IF{(BOD_G) AND (CNSW_M) AND (CPSW_M)}, THEN {RAS_M}, (10)

IF{(BOD_G) AND (CNSW_L)}, THEN {RAS_G}, (11)

where *BOD* – values of biological oxygen demands (BOD_{Total}) in wastewater (*SW*), *CNSW* – content of nitrogen, *CPSW* – content of phosphorus, *RAS* – replenishment of active sludge, while the probability P_e of heuristic effective operation is calculated as

Table 1. The Boundaries of Annual Monitoring of Contamination at Treatment Facilities

Type of contamination	Maximum value	Minimal value	Stratum	Realisation possibility
Biological oxygen demand (BOD_5)	370 mg/l	175 mg/l	G (370–250)	0.71
			M (250–200)	0.20
			L (200–150)	0.09
Suspension particles	30 mg/l	8 mg/l	G (30–15)	0.08
			M(15–10)	0.75
			L(10–5)	0.17
Nitrates	27.5 mg/l	12 mg/l	G(30–20)	0.03
			M (20–15)	0.03
			L (15–10)	0.94
Chemical oxygen demand	100	70	M(100–50)	1.0
Nitrites	1.0	0.1	G(1.0–0.5)	0.3
			M (0.5–0.1)	0.7
Phosphates	5	3	M (5–1)	1.0
Nitrogen ammonium	5	1	M (5–1)	1.0
pH	10.6 O_2/l	8.7 O_2/l	G(11–10)	037
			M(10–9)	0,46
			L(9–8)	0.17

$$P_e = (1/K) \sum_{k=1}^{K} p_{jk}, \tag{12}$$

where p_{jk} – the probability of realisation of j-th strate ($j=\overline{[1, J]}$) of k-th parameter (contamination), K – the quantity of the parameters (contamination).

Heuristics for the use of the enhancement process may be formulated as:

IF [{(VSWP_M) AND (BOD_M) AND (CAS_M) AND (IAS_M)} OR {(VSWP_L) AND (BOD_G) AND (CAS_G) AND (IAS_M)}], THEN (NCP_N), (13)

IF{(VSWP_M) AND (BOD_G) AND(CAS_G) AND (IAS_M),THAN(NCP_M), (14)

IF {(VSWP_G) AND (BOD_M) AND (CAS_M) AND (IAS_G), THAN(NCP_G), (15)

where *VSWP* – velocity of *SW* passing; *CAS* – active sludge concentration; *IAS* – the index of active sludge (reflects its properties); *NCP* – necessity of treatment process enhancement.

Practicability of preliminary adjusting of pH level of sewage may be formulated as:

IF {(LPH_M) AND (VSWP _M)}, THEN (NAL_N), (16)

IF{(LPH_L) AND (VSWP_G)}, THEN (NAL_G), (17)

IF{(LPH_G) AND (VSWP_L)}, THEN (NAL_M), (18)

where *LPH* – the level of *pH*, *NAL* – necessity of level of *pH* adjustment.

4. CONCLUSIONS

1. Methods of fuzzy set can be implemented for the environmental simulation with the help of intellectual technologies, including the combination of model pluralism with methods of fuzzy sets, membership functions, methods of nearest neighbours, results of fractal and chaos theories, methods of ensuring robustness and retrospective analysis of the decision tree for successful decision making.
2. The use of multiple models of state of the ecosystem or complex geoengineering system instead of the only one allows supplementing the existing data with unobserved one in alternative models, and provides a more comprehensive assessment of the situation (including behavioural motives) and informed decision making.
3. The augmentation of the Euclidean distance between pairs of actually observable states of the geoengineering system allows including in the estimated uncertainty some characteristics that cannot be established with the real observable data, i.e., it is possible to predict or update, with an estimated degree of certainty, the states of variables not included in the forecast or updated in the observation data.
4. The calculation of the boundaries of annual monitoring of contamination at treatment facilities with definition of the stratum and the realisation possibility demonstrates the trends of the process and separate periodic variable processes from catastrophic ones.
5. The use of the created approach allows for effective environment control and management under condition when there is insufficient or fuzzy information about the real state, outward and interior disturbances, and also about their deviations.

REFERENCES

1. Michna, J., Ekmanis, J., Zeltins, N., Zebergs, V., & Siemianowicz, J. (2012). Innovation Risk Management in the Rational Energy Use (Part 2). *Latvian Journal of Physics and Technical Sciences*, *49* (1), 3–15. doi: https://doi.org/10.2478/v10047-012-0001-9
2. Fayyad, U.M., Candel, A., Ario de la Rubia E., Pafka, S., Chong, A., Lee, J-Y. (2017). Benchmarks and Process Management in Data Science: Will We Ever Get Over the Mess? In *Proceedings of the 23rd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining (pp. 31–32). doi: https://doi.org/10.1145/3097983.3120998

3. Mika, M. (2017). Interoperability Cadastral Data in the System Approach. *Journal of Ecological Engineering, 18* (2), 150–156. https://doi.org/10.12911/22998993/68303
4. Dychko, A., Yeremeyev, I., Kyselov, V., Remez, N., & Kniazevych, A. (2019). Ensuring Reliability of Control Data in Engineering Systems. *Latvian Journal of Physics and Technical Sciences*, *56* (6), 57–69. doi: https://doi.org/10.2478/lpts-2019-0035
5. Dychko, A., Remez, N., Kyselov, V., Kraychuk, S., Ostapchuk, N., Kniazevych, A. (2020). Monitoring and Biochemical Treatment of Wastewater. *Journal of Ecological Engineering*, *21* (4), 150–159. doi: https://doi.org/10.12911/22998993/119811
6. Johnston, M., & Kazemzadeh, E. (2018). U.S. Patent No. 10,114,612. Washington, DC: U.S. Patent and Trademark Office.
7. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier.
8. Zhu, W., Sun, W., & Romagnoli, J. (2018). Adaptive k-Nearest-Neighbor Method for Process Monitoring. *Industrial & Engineering Chemistry Research, 57* (7), 2574–2586. doi: 10.1021/acs.iecr.7b03771
9. Diduk, N.N. (2014). The Measures of Internal and External Information (on Example of Probabilistic Situations of Uncertainty). Part IV. *System Research and Information Technologies, 1*, 113–129.
10. Jurasz, J., Piasecki, A., & Kaźmierczak, B. (2019). Sewage Volume Forecasting on a Day-Ahead Basis – Analysis of Input Variables Uncertainty. *Journal of Ecological Engineering*, *20* (9), 70–79. doi: https://doi.org/10.12911/22998993/112507
11. Pichler, M., Boreux, V., Klein, A., Schleuning, M., & Hartig F. (2019). Machine Learning Algorithms to Infer Trait-Matching and Predict Species Interactions in Ecological Networks. *Methods in Ecology and Evolution, 11,* 281–293. doi: 10.1111/2041-210X.13329
12. Minhas, R., Kleer, J., Matei, I., Bhaskar, S., Janssen, B., Bobrow, D.G. & Kurtoglu, T. (2014). Using fault augmented modelica models for diagnostics. In *Proceedings of the 10th International Modelica Conference 2014* (pp. 437–445), 10-12 March 2014, Lund, Sweden. Linkping University Electronic Press. doi: https://doi.org/10.3384/ecp14096437