BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research

Volume 5 | Issue 1 | January 2023 | ISSN 2630-5151（Online）

BILINGUAL
PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Volume 5 Issue 1 • January 2023 • ISSN 2630-5151 (Online)

Journal of Computer Science Research

BILINGUAL PUBLISHING GROUP

Pioneer of Global Academics Since 1984

BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Volume 5 | Issue 1 | January 2023 | Page1-45
Journal of Computer Science Research

Contents

Articles

BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Research on Precipitation Prediction Model Based on Extreme Learning Machine Ensemble

Xing Zhang, Jiaquan Zhou*, Jiansheng Wu, Lingmei Wu, Liqiang Zhang

Faculty of Mathematics and Computer Science, Guangxi Normal University of Science and Technology, Laibin, Guangxi, 546100, China

ABSTRACT

Precipitation is a significant index to measure the degree of drought and flood in a region, which directly reflects the local natural changes and ecological environment. It is very important to grasp the change characteristics and law of precipitation accurately for effectively reducing disaster loss and maintaining the stable development of a social economy. In order to accurately predict precipitation, a new precipitation prediction model based on extreme learning machine ensemble (ELME) is proposed. The integrated model is based on the extreme learning machine (ELM) with different kernel functions and supporting parameters, and the submodel with the minimum root mean square error (RMSE) is found to fit the test data. Due to the complex mechanism and factors affecting precipitation change, the data have strong uncertainty and significant nonlinear variation characteristics. The mean generating function (MGF) is used to generate the continuation factor matrix, and the principal component analysis technique is employed to reduce the dimension of the continuation matrix, and the effective data features are extracted. Finally, the ELME prediction model is established by using the precipitation data of Liuzhou city from 1951 to 2021 in June, July and August, and a comparative experiment is carried out by using ELM, long-term and short-term memory neural network (LSTM) and back propagation neural network based on genetic algorithm (GA-BP). The experimental results show that the prediction accuracy of the proposed method is significantly higher than that of other models, and it has high stability and reliability, which provides a reliable method for precipitation prediction.

Keywords: Mean generating function; Principal component analysis; Extreme learning machine ensemble; Precipitation prediction

*CORRESPONDING AUTHOR:
Jiaquan Zhou, Faculty of Mathematics and Computer Science, Guangxi Normal University of Science and Technology, Laibin, Guangxi, 546100, China; Email: wjsh2002168@163.com

ARTICLE INFO
Received: 6 December 2022 | Revised: 14 January 2023 | Accepted: 25 January 2023 | Published Online: 11 February 2023
DOI: https://doi.org/10.30564/jcsr.v5i1.5303

CITATION
Zhang, X., Zhou, J.Q, Wu, J.Sh., et al., 2023. Research on Precipitation Prediction Model Based on Extreme Learning Machine Ensemble. Journal of Computer Science Research. 5(1): 1-12. DOI: https://doi.org/10.30564/jcsr.v5i1.5303

1. Introduction

In recent years, due to the intensification of some natural factors and human activities, the global climate has changed severely, resulting in the frequent occurrence of various extreme natural disasters. For example, the rainstorm in Henan on July 20, 2021 affected 13.3198 million people in 1573 townships and towns in 150 counties of Henan Province, and the death toll reached 71 [1]. Precipitation data usually implicate rich information. Through the analysis of precipitation data, we can get the development law of data, and then predict the future precipitation in the precipitation area, so as to enormously reduce the critical harm of precipitation anomaly to society and people [2].

The traditional statistical methods have their own limitations, need to collect a large number of precipitation data and have high requirements for the quality of data, what is more, the complexity of precipitation causes makes its data nonlinear, which makes it difficult to predict [3,4]. However, the mathematical and statistical models used in traditional methods require complex computing power [5], and may be time-consuming and have little impact. With the development of science and technology, data acquisition methods are gradually diversified. The traditional precipitation prediction model can not meet the development needs of current precipitation prediction.

With the rapid progress of computer technology, machine learning technology is favored by many scholars. The application of artificial intelligence to the field of meteorology is also rising after more than 10 years of silence [6]. A neural network is widely used in various fields because of its fairly good adaptive learning ability and nonlinear mapping ability. Yu Xiang et al. applied ensemble empirical mode decomposition to decompose the original rainfall time series into a batch, and then used Support Vector Regression (SVR) to predict the short-term component intrinsic model function [7]. Yuanhao Xu et al. proposed particle swarm optimization (PSO) to optimize the super parameters of extended short-term memory (LSTM) neural network [8]. The real-time target detection method based on the convolutional neural network (CNN) classifier proposed by V.R.S. Mani et al. has achieved ideal results [9]. Zihao Zhang et al. proposed a variable weight neural network to solve a multivariable, strongly nonlinear, dynamic and time-varying problem [10].

Thanks to it being based on a least square algorithm, an extreme learning machine (ELM) has strong computing power and good generalization performance. Yong Ping Zhao et al. set up one-stage transfer learning ELM (OSTL-ELM) and two-stage transfer learning ELM (TSTL-ELM). OSTL-ELM makes use of one stage to extract information from two domains, while TSTL-ELM uses two stages to realize the separate adaptation of the target domain. The network weights of these two methods are generated by calculation rather than iteration. Only a small amount of target domain data is needed to acquire high diagnosis accuracy [11], CNN is combined with ELM, and the network is optimized based on the developed metaheuristic algorithm [12]. By transforming the structure of the ELM hidden layer, the threshold network can pass through adaptive stochastic resonance, and find the appropriate generalization performance of the threshold network by using the fast learning algorithm of ELM [13]. Xiao et al. employed regularized extreme learning machine (RELM) to distinguish fault types and identify faulty components. At the same time, LU decomposition was used to solve the output matrix of rELM, so as to shorten the training time of RELM [14]. Yang Ju generates an extreme learning machine classifier with large differences by randomly assigning hidden layer input weights and biases [15]. Chen Yang changed the distribution of hidden layer node parameters and randomly selected input weights for each ELM. Meanwhile, he searched for the optimal number of hidden nodes for each base learner and averaged the output consequences of all base learners [16].

At present, ensemble learning technology has received great attention from scholars. Ensemble learning is a technology to create and combines multiple machine learning models to produce an optimal prediction model. The most common is an ensemble classifier based on neural network technology and

using bagging, boosting and random subspace combination technology [17]. The algorithm proposed by Luká š Klein is based on a new combination of stack integration and basic learners. Wide and deep neural networks are used as meta-learners. The research results show that the algorithm achieves satisfactory results [18]. Madhurima Panja proposes an integrated wavelet neural network (XEWNet) model with exogenous factors. Compared with statistical, machine learning and deep learning methods, XEWNet performs better in 75% of the short-term and long-term predicted cases of dengue fever incidence rate [19]. A neural network ensemble method considering parameter sensitivity is proposed to solve the problem of convergence and relatively low accuracy of training [20].

Huang proposed that ELM has significant characteristics such as fast learning speed and excellent generalization performance in both regression and classification tasks [21]. However, due to the weights and deviations between the input layer and the hidden layer is randomly generated, the generated model is different each time. Ensemble learning can combine the advantages of ELM and make up for its disadvantages. In order to improve the accuracy and stability of ELM training and retain the advantages of ELM learning, a new ensemble model based on an extreme learning machine is proposed in this paper. The ensemble model is based on the extreme learning machine with different kernel functions and supporting parameters, and the submodel with the minimum root mean square error is found to fit the test data.

Owing to the complex mechanism and factors affecting precipitation change, the data have strong uncertainty and significant nonlinear variation characteristics. Therefore, in this paper, firstly, the mean-generating function method is used to extend the precipitation sequence, and the principal component analysis is used to reduce the dimension of the extension matrix. The processed data are used as the independent variable and the original precipitation sequence is used as the dependent variable to establish the extreme learning machine ensemble precipitation prediction model. The research structure of this paper is shown in **Figure 1**.

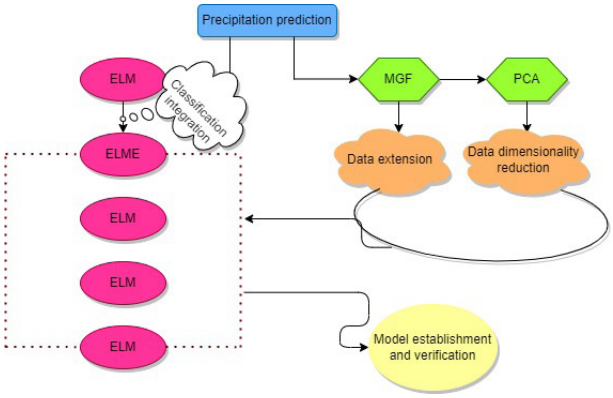


Figure 1. The structure of this study.

2. The proposed methodology

2.1 Mean generating function

For the sake of solving the problem the predicted value tends to be close to the average value of the series when multi-step prediction is carried out on time series data [22].

Wei Fengying and other scholars enriched the concept of arithmetic mean in mathematical statistics, and proposed the algorithm of mean generation function (MGF) [23].

Assuming the precipitation data series as $\{y_t, t = 1,2,\cdots,N\}$. The mean value of $\bar{y} = \frac{1}{N}\sum_{j=1}^{N} y(i)$ is $y(t)$. The MGF is calculated as follows:

$$y_l(i) = \frac{1}{n_l}\sum_{j=0}^{N_l-1} y(i+jl), i = 1,2,\cdots,l, 1 \le l \le Q \tag{1}$$

where $N_l = INT\left(\frac{N}{l}\right), Q = INT\left(\frac{N}{2}\right), l$ is the period of the mean generating function, Q is the maximum length of the cycle, INT is rounded.

The periodic extension sequence is obtained by periodic extension calculation.

$$Y_l(t) = y_l\left[t - l \cdot INT\left(\frac{t-1}{l}\right)\right], \quad t = 1,2,\cdots,N+p \tag{2}$$

where p is the number of steps to forecast the future, thus the extended mean generating function sequence matrix can be obtained.

$$Y^* = \begin{bmatrix} Y_1(1) & Y_1(2) & \cdots & Y_1(N+P) \\ Y_2(1) & Y_2(2) & \cdots & Y_2(N+P) \\ \vdots & \vdots & & \vdots \\ Y_Q(1) & Y_Q(2) & \cdots & Y_Q(N+P) \end{bmatrix}_{Q\times(N+P)} \tag{3}$$

Then the first column in the extensive matrix of the MGF is marked as y_1, the second column is recorded as y_2,..., the Q column is recorded as y_Q.

2.2 Principal component analysis

PCA is a dimensionality reduction algorithm favored by various scholars. That is, high-dimensional data are mapped to low-dimensional space through some linear projection, so as to maximize the amount of data information in the projection dimension and to achieve the purpose of using fewer data and retaining more source data [24]. The main flow of principal component analysis is shown in **Figure 2**.

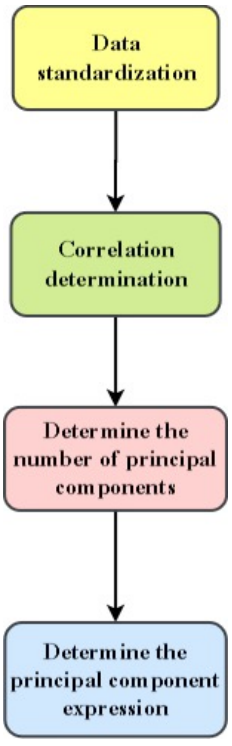


Figure 2. The flow of principal component analysis.

Assuming that there are m samples $\{X^1, X^2, \cdots, X^M\}$, each sample has n-dimensional features $X_i = (x_1^i, x_2^i, \cdots, x_N^i)^T$. Every feature x_j has its own eigenvalue. Centralize all features.

$$\overline{x_n} = \frac{1}{M}\sum_{i=1}^{M} x_n^i \tag{4}$$

Using matrix science, the relationship between eigenvalues λ of the covariance matrix C and its corresponding eigenvectors u is gained.

$$Cu = \lambda u \tag{5}$$

The primitive feature is projected onto the selected characteristic vector. For each sample X^i, the original feature is $(x_1^i, x_2^i, \cdots, x_N^i)^T$, and the new aspect obtained after projection is $(y_1^i, y_2^i, \cdots, y_k^i)^T$. The computing formula of the new feature is:

$$\begin{bmatrix} y_1^i \\ y_2^i \\ \vdots \\ y_k^i \end{bmatrix} = \begin{bmatrix} u_1^T * (x_1^i, x_2^i, \cdots, x_n^i)^T \\ u_2^T * (x_1^i, x_2^i, \cdots, x_n^i)^T \\ \vdots \\ u_k^T * (x_1^i, x_2^i, \cdots, x_n^i)^T \end{bmatrix} \tag{6}$$

For each and every specimen X^i, the dimension is reduced from the original N features of $Xi = (x_1^i, x_2^i, \cdots, x_N^i)^T$ to the new K properties, and the purpose of dimension reduction is achieved.

2.3 Extreme learning machine

The learning process of the ELM algorithm can be summarized as given a regression objective function or classification objective function, as long as the size of hidden nodes in a feedforward neural network is nonlinear and continuous, it can randomly generate the connection weight and threshold between the input layer and phase hidden layer without adjusting the size of hidden nodes, It can approach the target continuous function randomly or classify the classified targets, which improve the counting rate and model prediction accuracy. The structure of ELM is shown in **Figure 3**.

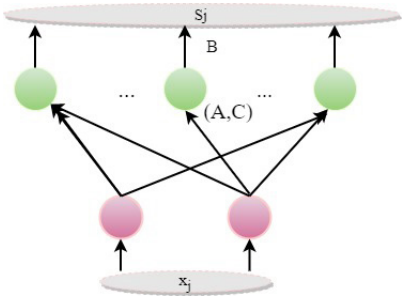


Figure 3. The structure of ELM.

ELM consists of an input layer, a hidden layer and an output layer. Assuming that the neurons in the

input layer be n, the neurons in the hidden layer be r, the neurons in the output layer be m, and the training set be $\{x_j, s_j \mid x_j \in R, s_j \in R, j = 1,2,\cdots,Q\}$.

In the ELM model, the connection weight between the input layer and the hidden layer and the threshold of the hidden layer neuron is emerged randomly [25], and the connection weight A is set as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rn} \end{bmatrix}_{r\times n} \tag{7}$$

where a_{ij} represents the connection weight of the i th neuron in the hidden layer and the j th neuron in the input layer. Set the connection weight B between the hidden layer and the output layer as:

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rm} \end{bmatrix}_{r\times m} \tag{8}$$

where b_{jk} represents the connection weight of the j th neuron in the hidden layer and the k th neuron in the input layer. If the deviation of hidden nodes is C, that is, the threshold of hidden layer neurons, there is

$$\mathrm{C} = [c_1, c_2, \cdots, c_r]'_{1\times r} \tag{9}$$

In general, the first step of ELM training is to use a stochastic-created fastened quantity of neuron nodes to construct the hidden layer. The activation function may be whatever nonlinear function. The commonly used activation functions include the sigmoid function, tanh function, relu function, etc. Let the activation function of hidden layer neurons be $g(X)$. Then from the figure, the output S of the network can be expressed as:

$$S_j = \begin{bmatrix} s_{1j} \\ s_{2j} \\ \vdots \\ s_{mj} \end{bmatrix}_{m\times Q} = \begin{bmatrix} \sum_{i=1}^{r} b_{i1} g\left(a_i x_j + c_i\right) \\ \sum_{i=1}^{r} b_{i2} g\left(a_i x_j + c_i\right) \\ \vdots \\ \sum_{i=1}^{r} b_{im} g\left(a_i x_j + c_i\right) \end{bmatrix}_{m\times 1}, \quad j = 1,2,\cdots,Q \tag{10}$$

where $a_i = [a_{i1}, a_{i2}, \cdots, a_{in}]; x_j = \left[x_{1j}, x_{2j}, \cdots, x_{nj}\right]^T$.

It also can be expressed by the following formula:

$$HB = S' \tag{11}$$

where H is the hidden layer output matrix of ELM. Because the connection between weight A and the threshold C of the hidden layer is generated randomly and preserves constant during the training process. Therefore, the connection weight between the hidden layer and the output layer B can be obtained by solving the least square solution of the following equations:

$$\min \left\| HB - S^T \right\| \tag{12}$$

The solution to Formula (12) is:

$$\hat{B} = H^+ S^T \tag{13}$$

where, H^+ is the Moore Penrose generalized inverse matrix of the matrix H [26].

2.4 Extreme learning machine ensemble

As weights and offsets between the ELM input layer and hidden layer are generated randomly, the models created are diverse at every turn, and their performance is also extremely discrepant. In order to surmount the problem of low precision of a single ELM model and instability results caused by randomly setting input weights, an extreme learning machine ensemble method is proposed in this paper to enhance the degree of accuracy and stability of precipitation prediction. Its structure is shown in **Figure 4**.

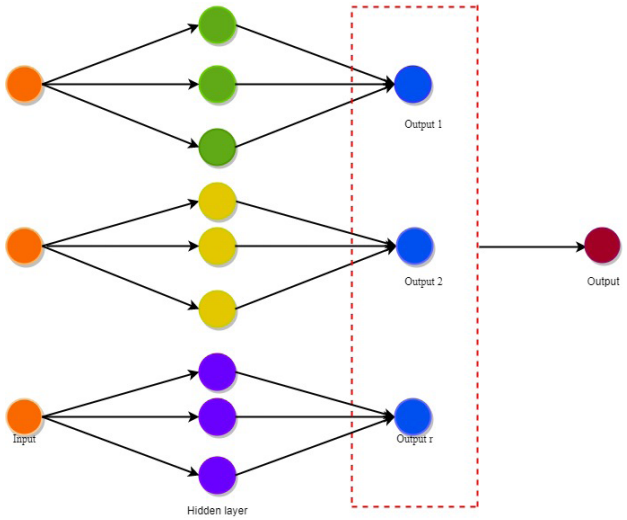


Figure 4. Network structure of ELME.

In order to ensure high accuracy and good stability of the results obtained by the ensemble model,

this paper will select the model with the minimum average absolute percentage error among the ELM model trained by different kernel functions,

$$\begin{aligned}&\min M\hat{E}(\hat{s}_j)=\frac{100}{\mathrm{n}}\sum_{i=1}^{j}\frac{\left|s_j-\hat{s}_j\right|}{\left|s_j\right|}\\&\text{Let}\min M\hat{E}(\hat{s}_j)=\frac{100}{\mathrm{n}}\sum_{i=1}^{j}\frac{\left|s_j-\hat{s}_j\right|}{\left|s_j\right|}=w\end{aligned}\tag{14}$$

At this point, the optimization problem is:

$$\begin{cases}\min\dfrac{1}{2}\left\|w^2\right\|\\ s.t.y_j\left(w\hat{s}_j+c\right)\geq 1\end{cases}\tag{15}$$

For solving the constrained optimization problem, the solution of the initial problem and the optimal problem can be gained by solving the dual problem.

The Lagrange multiplier $\alpha_j\geq 0$ is introduced into inequality (15), and the Lagrange multiplier method takes advantage of solving the above quadratic programming problem, then the above posture can be written as:

$$L(w,c,\alpha)=\frac{1}{2}\left\|w^2\right\|-\sum_{j=1}^{n}\alpha_j\left(y_j\left(w\hat{s}_j+c\right)-1\right)\tag{16}$$

Let $\theta(w)=\max_{\alpha_j\geq 0}L(w,c,\alpha)$

On the basis of the duality of Lagrange, the dual problem of the original optimization problem can be transformed by the minimax problem:

$$\begin{aligned}\min_{w,c}\theta(w)&=\min_{w,c}\max_{\alpha_j\geq 0}L(w,c,\alpha)\\&=\max_{\alpha_j\geq 0}\min_{w,c}L(w,c,\alpha)\end{aligned}\tag{17}$$

Find the partial derivatives of B and C respectively.

$$\begin{aligned}&\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{j=1}^{n}\alpha_j y_j\hat{s}_j\\&\frac{\partial L}{\partial c}=0\Rightarrow\sum_{j=1}^{n}\alpha_j y_j=0\end{aligned}\tag{18}$$

Bring the outcome into Equation (16) to obtain:

$$\begin{aligned}L(w,c,\alpha)&=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\hat{s}_j^T\hat{s}_i-\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\hat{s}_j^T\hat{s}_i-b\sum_{j=1}^{n}\alpha_j y_j+\sum_{\mathrm{j}=1}^{\mathrm{n}}\alpha_{\mathrm{i}}\\&=\sum_{j=1}^{n}\alpha_j-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\hat{s}_j^T\hat{s}_i\end{aligned}\tag{19}$$

Thus, in light of the restraint condition, it is transformed into a convex quadratic programming problem:

$$\begin{aligned}&\max\sum_{j=1}^{n}\alpha_j-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\hat{s}_j^T\hat{s}_i\\&s.t\alpha_j\geq 0\\&\sum_{j=1}^{n}\alpha_j y_j=0\end{aligned}\tag{20}$$

According to the above conditions, the unique solution α_j^* of quadratic programming can be acquired, and the optimal decision function form can be obtained after sorting:

$$f(s)=sign\left(\sum_{j=1}^{n}\alpha_j y_j\left(\hat{s}_j,\hat{s}_i\right)+c\right)\tag{21}$$

The pseudo of ELME is shown in Algorithm 1.

Algorithm 1

Begin

Input $x_i=[x_1,x_2,\cdots,x_k,x_{k+1},\cdots,x_j]^T$

Output: $f(s)$

Generate randomly: a and c;

for i=1 **to** j.

Set activation function：sine、sigmoid、hardlim;

Calculate S and B;

end for

for i=1 to k

Sort min $\mathrm{M\hat{E}}(s)$;

Set $y_j\left(w\hat{s}_j+c\right)\geq 1$;

Introducing Lagrange multiplier;

Calculate α_j;

End

3. Empirical research

3.1 Modeling data

Liuzhou is the largest industrial base in Guangxi. The sustained and stable economic development of Liuzhou is of great importance to the development of Guangxi. Therefore, the real data on precipitation in Liuzhou from Guangxi Meteorological Bureau are selected in this paper. The aggregate data are 213 data from 1951 to June, July and August 2021. A total of 180 data from 1951 to June, July and August 2010 are used as the training data set to establish the precipitation fitting model, and the data from June,

July and August 2011 to 2021 are used as the test data set to optimize the verification model.

The precipitation data used in this paper first employs MGF method to extend the monthly precipitation series of Liuzhou from 1951 to 2021 in June, July and August, and takes the value that the cumulative contribution rate of principal component variance reaches 90%, so as to further reconstruct the original data. Then, take 10 steps of extension, establish the mean generating function extension matrix for the reconstructed succession data, and receive the mean generating function extension matrix, and then employ PCA to reduce the dimension of the data obtained by MGF and extract effective data properties.

3.2 Model performance evaluation

In order to directly perceived through the senses observe the effect of model fitting, training data and test data are made use of in this paper to drill and test the model, and the indicators in **Table 1** are used to measure the quality of the model.

Table 1. Performance evaluation metrics.

Number	Metric	Value
1	Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2}$
2	Symmetric Mean Absolute Percentage Error (MAPE)	$sMAPE = \frac{100\%}{\mathrm{n}}\sum_{i=1}^{n}\frac{\left\|\hat{y}_i - y_i\right\|}{\left(\left\|\hat{y}_i\right\| + \left\|y_i\right\|\right)/2}$
3	Pearson correlation coefficient (PCC)	$PCC = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)\sum_{i=1}^{n}\left(\hat{x}_i - \bar{\hat{x}}_i\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)^2\sum_{i=1}^{n}\left(\hat{x}_i - \bar{\hat{x}}_i\right)^2}}$

where x_i indicates the observed value, and $\hat{x}_i$ represents the fitting value. $\bar{x}_i$ represents the mean of monthly precipitation observation value, and $\hat{x}_i$ represents the equal value of model output value.

Evaluation indexes 1 and 2 can be used to measure the deviation between the actual observation value and the fitting value of precipitation. The smaller the value, the smaller the deviation between them. Evaluation index 3 can perceive whether the model can correctly predict the precipitation trend. The greater its value, the more accurate the model can predict the future precipitation trend.

3.3 Result analysis

In order to verify the quality of ELME model. In this paper, the proposed ELME model is compared with representative machine learning methods such as ELM, GA-BP and LSTM. For the ELM model, this paper trained a total of 15 cases in which five activation functions of "Sigmoid", "Sine", "Hardlim", "Radbas" and "Tribas" were combined with three hidden neurons of 10, 20 and 30. The parameter combination with the best training effect was selected, which was the activation function "sine" with hidden neuron 30. The parameters were employed in the training of ELM. For GA-BP model, the parameters of the genetic algorithm are set as follows: crossover probability is 0.3, mutation probability is 0.1. For the LSTM model, the parameters are set as follows: Solver is "Adam", gradient threshold is 1, and the initial learning rate is 0.01. After 125 rounds of training, the learning rate is reduced by multiplying factor 0.2. The fitting results of the four models for 60 training data of precipitation in June, July and August in Liuzhou are presented in **Figure 5**, **Figure 6** and **Figure 7**. The data in **Table 2** specifically illustrate the fitting precision and fitting effect of the four models on the training data.

As can be observed in **Figure 5**, **Figure 6** and **Figure 7**, the fitted values and real values of ELM, LSTM, GA-BP and ELME on Precipitation in Liuzhou in June, July and August have roughly the same trend. Among them, there is a section with a good fitting effect and a section with relatively considerable fitting error, and the fitting effect is consistent with the general experimental data fitting situation. Obviously, the fitting effect of LSTM, GA-BP and ELME model is closer to the real value in June and July. In August, it can be seen that the ELME model still has the same trend with the real value and the difference between each real value and the fitted value is not gigantic.

As can be seen from **Table 2**, the correlations of

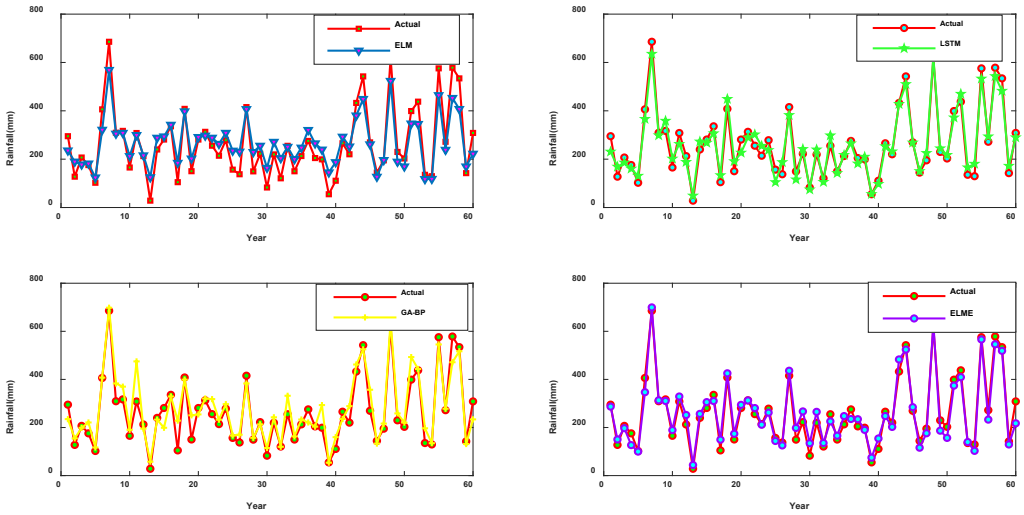


Figure 5. Fitting effect of training data of four models in June.

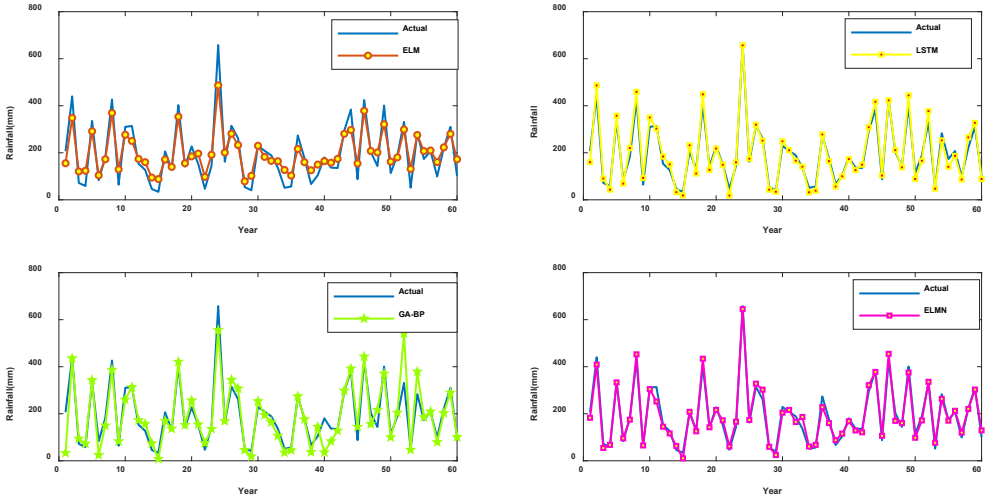


Figure 6. Fitting effect of training data of four models in July.

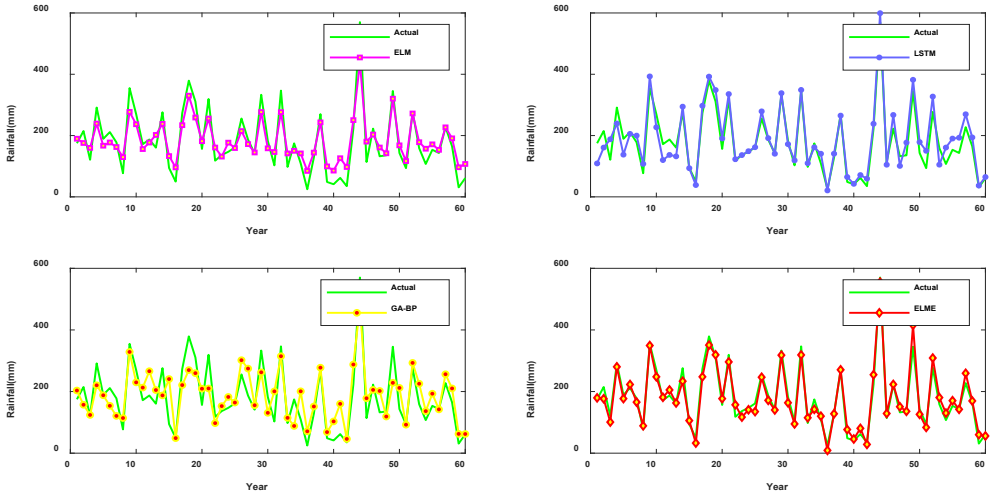


Figure 7. Fitting effect of training data of four models in August.

the four models in June, July and August are highly correlated, indicating that the four models can make correct predictions on the precipitation trend. In addition, the RMSE value of ELM model was 59.418 and sMAPE value was 0.220 in June, the RMSE value of LSTM model was 31.566 and sMAPE value was 0.129, and the RMSE value of GA-BP model was 50.811 and sMAPE value was 0.158, while RMSE value of ELME model is 30.253, sMAPE value is 0.127. In modeling the factor under the same conditions, ELME the precision of the model, relative to the ELM model LSTM model, GA-BP model increased by 42.272%, 1.550% and 19.620% respectively. Meanwhile, in the precipitation data in July ELME the precision of the model, relative to the ELM model LSTM model, GA - BP model increased by 53.169%, 10.135% and 46.800%, respectively. August precipitation data model of ELME the precision of the model, relative to the ELM model LSTM model, GA - BP model increased by 48.031%, 20.482% and 54.007% respectively. The above data show that the fitting accuracy of ELME model based on precipitation data in different months is significantly better than that of ELM model, LSTM model and GA-BP model in training data.

One aspect of evaluating a model is its fitting effect, but more vital is its prediction effect, that is, the generalization ability of the model. Based on the above training model, the test data of precipitation in Liuzhou city in June, July and August are fitted, and the fitting results are shown in **Figure 8**, **Figure 9** and **Figure 10**. The data in **Table 3** specifically illustrate the fitting precision and fitting effect of the four models on the test data.

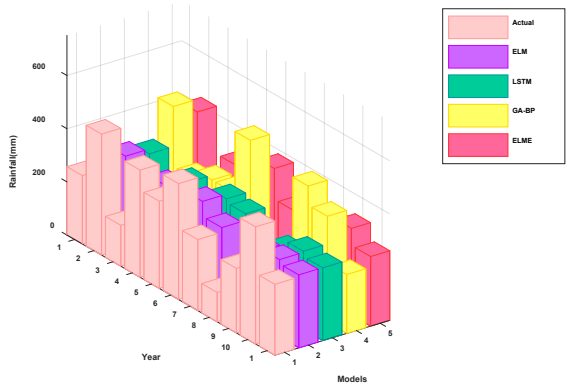


Figure 8. Four models were tested for data fitting in June.

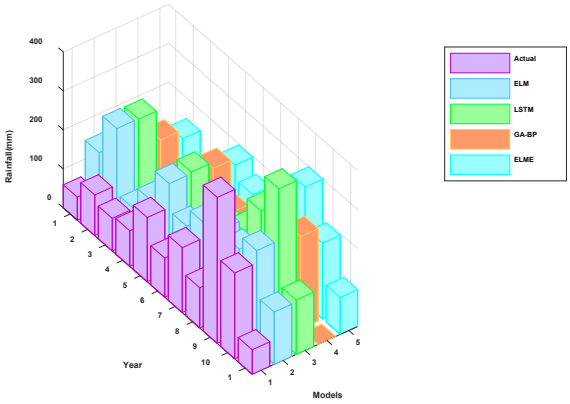


Figure 9. Four models were tested for data fitting in July.

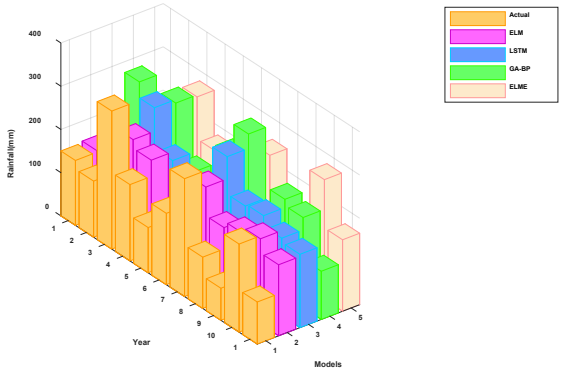


Figure 10. Four models were tested for data fitting in August.

Table 2. Evaluation indexes of training data fitting effect of four models.

Models	June			July			August		
	RMSE	sMAPE	PCC	RMSE	sMAPE	PCC	RMSE	sMAPE	PCC
ELM	59.418	0.220	0.946	50.850	0.284	0.970	42.532	0.254	0.967
LSTM	31.566	0.129	0.977	22.940	0.148	0.989	32.072	0.166	0.958
GA-BP	50.811	0.158	0.942	51.071	0.250	0.925	58.558	0.287	0.826
ELME	30.253	0.127	0.978	21.061	0.133	0.986	21.133	0.132	0.979

4. Discussion

As can be seen from **Table 3**, the correlation of GA-BP model established by precipitation data in July is only 0.458 in test data, indicating that the correlation of GA-BP model in this month is not very content. Meanwhile, the correlation of GA-BP model established by precipitation data in June and August is 0.677 and 0.666 respectively. The correlation strength of the model is moderate. In addition, the correlation between ELM model and LSTM model based on precipitation data in July was 0.516 and 0.676 respectively, and the correlation between LSTM model based on precipitation data in August was 0.710, indicating that the correlation strength of models verified by this test data was relatively general. As for ELME model, the correlation coefficients in June, July and August are 0.896, 0.873 and 0.847 respectively, indicating that the model has a greater correlation with the precipitation data in any month.

For precipitation test data in different months, it can be seen from **Table 3** that the values of RMSE and sMAPE of ELME model are smaller than those of ELM, LSTM and GA-BP models. Among them, the sMAPE value of ELM, LSTM and GA-BP in June was 0.187, 0.251 and 0.352 respectively, while the sMAPE value of ELME model was 0.144, which compared with the other three models improved by 22.995%, 42.629% and 59.091% respectively. The accuracy of July model ELME was improved by 42.647%, 41.542% and 59.912% compared with model ELM, LSTM and GA-BP, respectively. The accuracy of the August model ELME was improved by 31.154%, 39.322% and 45.427% compared with model ELM, LSTM and GA-BP, respectively. The above data show that the fitting accuracy of ELME model based on precipitation data in different months is significantly better than that of ELM model, LSTM model and GA-BP model.

Based on **Table 2** and **Table 3**, it can be seen that under the same construction pattern, the ELME model has a significantly better fitting effect on precipitation data than LSTM model and ELM model, regardless of training data or test data. In addition, compared with ELME model, it can be found that ELME model has superior fitting stability in precipitation data.

5. Conclusions

In the modeling research of monthly precipitation forecast in atmospheric science, the one-dimensional time series observation data of various meteorological elements or climate elements can provide the most notable forecast information source. With the rapid development of machine learning technology, every machine learning prediction technology can provide crucial and useful forecast information. In this paper, MGF is used to extend the precipitation series, and PCA is used to reduce the dimension of the extended series, so as to establish the ensemble precipitation prediction model of an extreme learning machine.

A novel extreme learning machine ensemble is put forward in this paper. The ensemble model is based on the extreme learning machine with different kernel functions and supporting parameters, and the submodel with the minimum root mean square error is found to fit the test data. Consequently, the ELME model proposed in this paper reduces the complexity of the model and achieves better performance. In this paper, the precipitation data of Liuzhou from 1951 to 2021 in June, July and August were utilized to train the model, and the model was

Table 3. Evaluation indexes of fitting effect of test data of four models.

Models	June			July			August		
	RMSE	sMAPE	PCC	RMSE	sMAPE	PCC	RMSE	sMAPE	PCC
ELM	72.374	0.187	0.898	88.627	0.476	0.516	53.077	0.260	0.788
LSTM	84.686	0.251	0.822	88.671	0.467	0.676	55.663	0.295	0.710
GA-BP	109.454	0.352	0.677	88.910	0.681	0.458	69.072	0.328	0.666
ELME	57.049	0.144	0.896	45.685	0.273	0.873	40.119	0.179	0.847

compared with ELM, LSTM and GA-BP models. Experimental results show that the proposed ELME achieves accurate prediction in the field of precipitation, and the model has a simple structure, which can be used as an alternative to reduce the complexity of the model. This shows that ELME can be used in a variety of machine learning domains and has some general applicability, and the proposed algorithm can be verified on a variety of data sets in the future. However, the three activation functions in this paper are randomly set. At present, this structure cannot automatically select the three most appropriate activation functions. In the future, we can consider how to select the activation functions that are suitable for this structure at one time.

Author Contributions

All the authors have made significant contributions to the work of the report. Xing Zhang is mainly responsible for the construction of the idea of this article, the simulation experiment and the writing of the paper. Jiaquan Zhou is mainly responsible for controlling the full text; Jiansheng Wu is mainly responsible for providing ideas; Lingmei Wu is mainly responsible for simulation experiments, and Liqiang Zhang is mainly responsible for obtaining data.

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Funding

This research was funded by Scientific Research Project of Guangxi Normal University of Science and Technology, grant number GXKS2022QN024.

Acknowledgement

The author thanks all the people who have provided technical support for this paper. Thanks to the teacher who provided meteorological data for this experiment.

References

[1] Ma, Y., Zhang, J.L., Li, L.W., 2022. Maintenance mechanism of "21·7" Torrential rain in henan province. Meteorology and Environmental Science. 45(4), 1-12.

[2] Du, Y., 2018. Characteristic analysis and prediction of hydrological time series—Take precipitation in Nanning as an example [Master's thesis]. Nanning: Guangxi University.

[3] Fang, W., Pang, L., Wang, N., et al., 2020. A review of the application of artificial intelligence in short approaching precipitation forecast. Journal of Nanjing University of Information Science & Technology. 12(4), 406-420.

[4] Wu, C.L., Chau, K.W., 2013. Prediction of rainfall time series using modular soft computing-methods. Engineering Applications of Artificial Intelligence. 26(3), 997-1007.

[5] Singh, P., Borah, B., 2013. Indian summer monsoon rainfall prediction using artificial neural network. Stochastic Environmental Research and Risk Assessment. 27(7), 1585-1599.

[6] Wang, T., Liu, Y.P., Dong, C., 2019. A review of the methods and applications of short impending precipitation forecast. The Electronic World. 41(10), 11-13.

[7] Xiang, Y., Gou, L., He, L.H., et al., 2018. A SVR-ANN combined model based on ensemble EMD for rainfall prediction. Applied Soft Computing. 73(9), 874-883.

[8] Xu, Y.H., Hu, C.H., Wu, Q., 2022. Research on particle swarm optimization in LSTM neural networks for rainfall-runoff simulation. Journal of Hydrology. 608(5), 553-565.

[9] Mani, V.R.S., Saravanaselvan, A., Arumugam, N., 2022. Performance comparison of CNN, QNN and BNN deep neuralnetworks for real-time object detection using ZYNQ FPGA node. Microelectronics Journal. 119(1), 319-331.

[10] Zhang, Z.H., Huang, X.H., Zhang, T.H., 2022. Analytical redundancy of variable cycle engine based on variable-weights neural network. Chinese Journal of Aeronautics. 28(1), 28-40.

[11] Zhao, Y.P., Chen, Y.B., 2022. Extreme learning machine based transfer learning for aero engine fault diagnosis. Aerospace Science and Technology. 121(2), 311-326.

[12] Han, E.F., Ghadimi, N., 2022. Model identification of proton-exchange membrane fuel cells based on a hybrid convolutional neural network and extreme learning machine optimized by improved honey badger algorithm. Sustainable Energy Technologies and Assessments. 52(8), 5-19.

[13] Chen, Z.J., Duan, F.B., Blondeau, F.C., 2022. Training threshold neural networks by extreme learning machine and adaptive stochastic resonance. Physics Letters A. 432, 8-21.

[14] Xiao, L., Zhang, L.Y., Yan, Z., 2022. Diagnosis and distinguishment of open-switch and current sensor faults in PMSM drives using improved regularized extreme learning machine. Mechanical Systems and Signal Processing. 171(3), 866-879.

[15] Yang, J., Yuan, Y.L., Yu, H.L., 2016. Selective ensemble learning algorithm for extreme learning machine based on ant colony optimization. Computer Science. 43(10), 266-271.

[16] Wang, J.H., Hu, J.W., Cao, J., et al., 2022. Multi-fault diagnosis of rolling bearings based on adaptive variational mode decomposition and integrated extreme learning machine. Journal of Jilin University. 52(2), 318-328.

[17] Dhibi, K., Mansouri, M., Bouzrara, K., et al., 2022. Reduced neural network based ensemble approach for fault detection and diagnosis of wind energy converter systems. Renewable Energy. 194, 778-787.

[18] Klein, L., Seidl, D., Fulneček, J., et al., 2023. Antenna contactless partial discharges detection in covered conductors using ensemble stacking neural networks. Expert Systems with Applications. 213.

[19] Panja, M., Chakraborty, T., Nadim, S., et al., 2023. An ensemble neural network approach to forecast Dengue outbreak based on climatic condition. Chaos, Solitons & Fractals. 167.

[20] Hu, X.Y., Zeng, Y., Qin, C., et al., 2022. Bagging-based neural network ensemble for load identification with parameter sensitivity considered. Energy Reports. 8, 199-205.

[21] Huang, G.B., Zhu, Q.Y., Siew, C.K., 2004. Extreme learning machine: A new learning scheme of feedforward neural networks. IEEE Int. Joint Conf. Neural Netw. 2, 985-990.

[22] Liu, Y.J., 2017. Research on mixed forecast model of summer precipitation in Jilin Province [Master's thesis]. Changchun: Northeast Normal University.

[23] Wei, F.Y., Cao, H.X., 1990. Mathematical models of long term forecasting and their applications. Meteorological Press: Beijing.

[24] Zhang, D.P., 2021. Research on customer credit management of mobile companies based on principal component analysis [Master's thesis]. Beijing: North China Electric Power University, School of economics and management.

[25] Ma, M.J., Yang, J.H., Liu, R.B., 2022. A novel structure automatic-determined Fourier extreme learning machine for generalized Black-Scholes partial differential equation. Knowledge-Based Systems. 238(2), 904-912.

[26] Tummalapalli, S., Kumar, L., Krishna, A., 2022. Detection of web service anti-patterns using weighted extreme learning machine. Computer Standards & Interfaces. 82(8), 621-632.

BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Outdoor Air Quality Monitoring with Enhanced Lifetime-enhancing Cooperative Data Gathering and Relaying Algorithm (E-LCDGRA) Based Sensor Network

G. Pius Agbulu**, G. Joselin Retnar Kumar

Department of Electronics and Instrumentation Engineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulatur, Kancheepuram, Chennai, TN, 600083, India

ABSTRACT

The air continues to be an extremely substantial part of survival on earth. Air pollution poses a critical risk to humans and the environment. Using sensor-based structures, we can get air pollutant data in real-time. However, the sensors rely upon limited-battery sources that are immaterial to be alternated repeatedly amid extensive broadcast costs associated with real-time applications like air quality monitoring. Consequently, air quality sensor-based monitoring structures are lifetime-constrained and prone to the untimely loss of connectivity. Effective energy administration measures must therefore be implemented to handle the outlay of power dissipation. In this study, the authors propose outdoor air quality monitoring using a sensor network with an enhanced lifetime-enhancing cooperative data gathering and relaying algorithm (E-LCDGRA). LCDGRA is a cluster-based cooperative event-driven routing scheme with dedicated relay allocation mechanisms that tackle the problems of event-driven clustered WSNs with immobile gateways. The adapted variant, named E-LCDGRA, enhances the LCDGRA algorithm by incorporating a non-beacon-aided CSMA layer-2 un-slotted protocol with a back-off mechanism. The performance of the proposed E-LCDGRA is examined with other classical gathering schemes, including IEESEP and CERP, in terms of average lifetime, energy consumption, and delay.

Keywords: Air quality; Cluster; Delay; Energy; Lifetime; WSN

*CORRESPONDING AUTHOR:
G. Pius Agbulu, Department of Electronics and Instrumentation Engineering, SRM Institute of Science and Technology, SRM Nagar, Kattankulatur, Kancheepuram, Chennai, TN, India; Email: gpagbulu@gmail.com

ARTICLE INFO
Received: 31 December 2022 | Revised: 14 January 2023 | Accepted: 25 January 2023 | Published Online: 11 February 2023
DOI: https://doi.org/10.30564/jcsr.v5i1.5383

CITATION
Agbulu, G.P., Kumar, G.J.R., 2023. Outdoor Air Quality Monitoring with Enhanced Lifetime-enhancing Cooperative Data Gathering and Relaying Algorithm (E-LCDGRA) Based Sensor Network. Journal of Computer Science Research. 5(1): 13-20. DOI: https://doi.org/10.30564/jcsr.v5i1.5383

1. Introduction

In recent years, air pollution has intensified in practically all societies on the planet [1]. The amount of particles in the atmospheric air is ascending, and the repercussion on the ecosystem cannot be overlooked [2]. Air contamination is a major root of mortality worldwide [3-5]. Air pollutant status is broadly tracked by adopting traditional fixed-monitoring mechanisms [5-7]. These typical monitors are extremely expensive, massive, and bulky [6]. Besides, air contamination zones may deviate in seconds, and typical monitoring apparatuses cannot recognize such swift divergences. Using sensor-based structures, we can get air pollutant data in real-time [8-11].

WSN (wireless sensor network) is an autonomous wireless configuration, set up by mini-sized sensor node devices [12]. The sensors are supposed to observe numerous environmental parameters and transfer the observation to a gateway [13]. The sensors rely upon limited-battery sources that are unreal to be swapped always amid considerable broadcast costs. Generally, the sensors disperse their total-energy prematurely from constant tracking and broadcast tasks [14,15]. Thus, effective energy administration measures must be enforced to handle the expense of power-dissipation.In this situation, routing policies play a principal role, regulate the QoS and energy diffusion at the sensors. Clustered aggregation is an outstanding technology widely used to lessen redundancy, energy allocation and lengthen WSN longevity [16]. In a clustered structure, the sensors are split into cells while a leader is allocated to accumulate readings of their cell and convey them to the remote gateway [17]. Various clustered aggregation schemes can be found in the literature [18-24].

To enhance the lifetime of diverse clustered WSNs, the lifetime-enhancing cooperative data gathering and relaying algorithm (LCDGRA) was proposed lately [25]. LCDGRA is a cluster-based cooperative event-driven routing scheme with dedicated relay allocation mechanisms that tackle the problems of event-driven clustered WSNs with immobile gateways. It uses a centralized hybrid clustering strategy based on Huffman coding and K-means clustering to section sensors into the K-number of clusters. In LCDGRA, relay nodes are committed in the diverse cluster fields to support the CH's transporting their assembled sensory data to the central gateway. Random linear coding is realized at each hop from the event cluster to the central gateway/BS to assure minimum energy consumption. Hence, the relays exploit decode and forward techniques to cooperatively relay the observations to the central gateway. In this study, we propose outdoor air quality monitoring using WSN with an enhanced lifetime-enhancing cooperative data gathering and relaying algorithm (E-LCDGRA). The adapted E-LCDGRA improves the original LCDGRA algorithm by incorporating a non-beacon-aided CSMA layer-2 un-slotted protocol with a back-off mechanism. The performance of E-LCDGRA is examined with other typical clustered event-driven gathering schemes, including IEESEP and CERP, in terms of average lifetime, energy consumption, and delay.

2. System model and communication protocol

Figure 1 reveals the WSN model examined in this work for air quality monitoring. $G= (V, S)$ denotes the network's directed graph. V denotes the vertexes, which comprise arrays of nodes spread arbitrarily in the outdoor air quality monitoring zone and a central base station (BS) located at the monitoring zone-end. S signifies the links or edges. According to their functions, each node fits into relay node (RN), normal node (NN), and cluster head (CH) categories. The proposed solution named E-LCDGRA is developed in cognizance of existing clustered event-driven routing design. In the examined sensor network, the sensor devices run over Zigbee/IEEE 802.15.4 protocol whilst being cognizant of the in-network aggregation methodology. Every node device possesses the same ability to run as a full-function device (FFD) and reduced function-devices (RFD). Hence, the node devices can run in either sensor monitoring or communication modes to transfer recorded air quality data to other sensor neighbours in their reach.

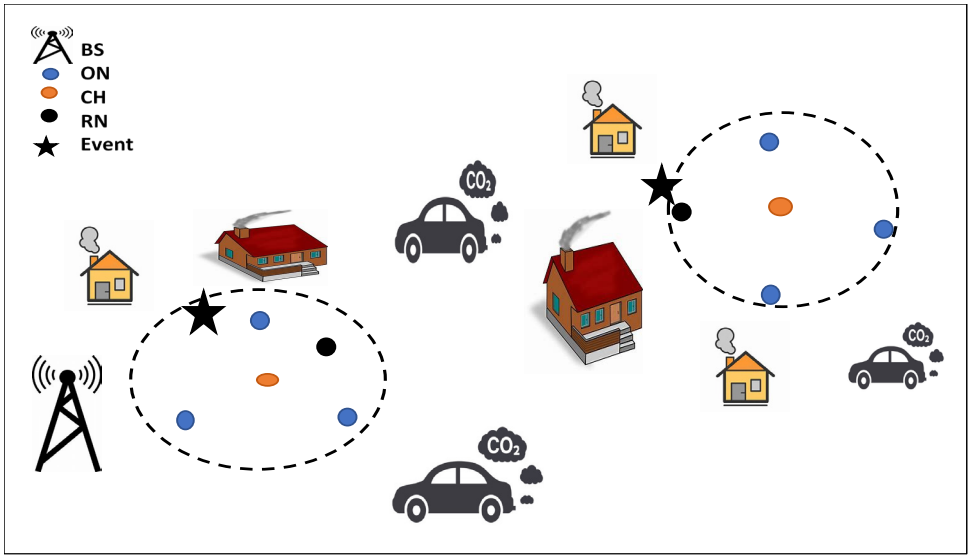


Figure 1. Proposed network model.

The protocol operation of the RN and CH is network coding-cognizant, and based on a non-beacon-aided CSMA layer-2 un-slotted protocol with a back-off mechanism. To boost more coding chances at the coding-layer, the CH's and RN's use CSMA-based listening and a back-off method to realize coding actions. It permits them to briefly hold their transmissions and listen to broadcasts from MAC-layers of their upper layers before disseminating their packets. This is exclusively meant to boost coding gains at the intermediate nodes, as opposed to the archetypal collision-evasion mechanisms. All coded packets ordered by NC-layer header and a notification message are routed to the MAC-layer. The receiver's hash for the MAC-address is included with the NC-header, to guarantee ease in the decoding operation. The operation of decode and forward (DF) on the packets at the relays, from the event zone is realized per-hop until they are obtained at the BS. **Figure 2** shows the proposed IEEE 802.15.4-based asynchronous communication protocol for two CH and RN.

3. Design methods and phases of LCDGRA algorithm

This section elaborates on the design methods of the adapted scheme named E-LCDGRA, which consists of three phases as follows:

- Initialization and clustering;
- Relay nodes allocation;
- Data aggregation and broadcasting.

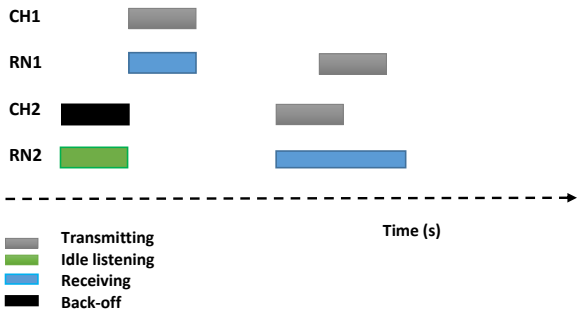


Figure 2. Proposed IEEE 802.15.4 based asynchronous communication protocol of 2-CH's and 2- RN's.

3.1 Initialization and clustering

In this phase, the sensor network is initialized, then the sensors are clustered into equal K cells. The central gateway starts this phase with messages of initialization (*I-REQ*) forwarded to every node in the network space; whilst each node replies to the request by sending responses for initialization (*I-REP*). The responses (*I-REP's*) from sensors have information about their locations and energy.

By the procedures defined above, the network is initialized. A centralized hybrid clustering strategy based on Huffman coding and K-means clustering is employed to section sensors into K-number of clus-

ters. It is intended to augment the node's coverage distances with their energy usage at the K formation stage. The allocation of the CHs is based on gauged competing value for each node regarding the distance from a competing node to its K-unit members, the contending node's residual energy with reference to the energy desired for the transceiving of the member's K-bit packets, and the energy for RLNC-based aggregation. The clustering and CH allocation scheme are defined in the following subsections.

Cluster development stage

At this epoch, the nodes are shared into K-divisions of clusters. Firstly, the optimal overall K-points are computed using Equation (1).

$$k_O = \frac{A}{D(N,BS)^2} \times \sqrt{{d_o}^2} \times \sqrt{\frac{N}{2\pi}} \quad (1)$$

Here, D (N, BS) denotes the node's Euclidean distances to BS, d_o signifies the threshold for communication and A represents the monitoring area. Once the overall quantity of K-points is worked out, the sensors are allocated to their nearby cluster centroids. The distance between the cluster center-point (centroid) and the nodes is defined by Equation (2).

$$d(j) = \sum_{i=1}^{N} \sqrt{\left(X_{Ni} - X_j\right)^2 + \left(X_{Yi} - Y_j\right)^2} \quad (2)$$

where, $i=1,2,....,N$, X_{Ni} and X_j signify the node and cluster-centroid's X coordinates, while Y_{Ni} and Y_j represent their respective Y coordinates, meanwhile; N stands for total nodes. Lastly, fresh center-points are calculated for all clusters till the points become unchaining.

Nomination of CH's

In this phase, CH's are nominated. Firstly, a competing-value N_{compi} is premeditated for every node in entire clusters employing Equation (3).

$$N_{compi} = \frac{E_{resi}}{\sum_{i=1}^{N} d_{Node}^2 + k\,(E_{rx} + E_{tx}) + E_{DAN}} \quad (3)$$

where, $i=1,2\ldots,N$, and N signifies the cluster overall members, d_{node} connotes distance from the members to the competing node,E_{resi}signifies the competing node's residual energy, E_{rx} signifies the energy desirable for the receipt of the member's k-bit packets, E_{tx} signifies the energy required to transfer the packets to the adjoining relays, while E_{DAN} represents the requisite energy for realizing RLNC and In-network aggregation.

Next, succeeding evaluation of contending value N_{compi} for every node, each sensor node's cost is multiplied by a random value in the line of 0 and 1, in order to ascertain their various probabilities. The obtained probabilities for all node is subsequently summed to one and set up in a descending set. Later, a code is found for all the nodes with Huffman coding method to figure out their weights. Eventually, the sensor node that possesses the lightest weight in the distinct clusters is adopted as the head node and introduced to the K members. In each round-cycle, other CH's are appointed in all K-clusters to promote load balancing until every node drains its battery power within the sensing area.

3.2 Relay node allocation

Here, the relay node appointment takes place. Relay nodes are committed in the diverse cluster fields to support the CH's transporting their assembled sensory data to the central gateway. Analogous to the CH appointment, the relay allocation scheme is computed by the central gateway employing a gradient-descent (GD) based relay allocation algorithm. Algorithm 1 gives the GD relay assignment scheme.

3.3 Data aggregation and broadcasting

After the CH's and relays are determined, the nodes switch into idle states forecasting events. An event means a mutation in the perceived sensory value of air quality (AQ) above defined-thresholds. Accompanied by incidents of an event within the outdoor air quality monitoring region are phases of aggregation and broadcasting. It is well acquainted that in WSN, the number of data transmissions substantially affects the network's energy usage. Accordingly, it becomes vital to mitigate the estimate of communications to maintain remarkably less energy usage. In the designed scheme, random linear coding is realized at each hop from the event cluster

Algorithm 1. Relay Node Election Algorithm

Step1: Input: Initial Guess (x_0), Step length (a),
Step 2**: Initialization**
Repeat
Step 3: **For** i=0, i< Max i;
i++ **do**
Step 4: **For** ***N*** **do**
Step 2: Calculate model parameters
Step 3: Set the parameters and assess $f_{(GD)}$ gradient with reference to the weights
Step 4: Modify the weights taking steps proportionate to the $f_{(GD)}$ –Gradient, to find the optimal values that minimizes $f_{(GD)}$
Step 5: Update relays
Step 6: **End for**
Step 7: **End for**
Step 8: Until the iteration Maximum
Step 9: Assign the best node as "RN"

to the central gateway/BS to assure minimum energy consumption. Hence, the relays exploit decode and forward techniques to cooperatively relay the observations to the central gateway.

Upon event incidence in the monitoring space, the event region K-members convey their recorded data to the head node. Then, the cluster head, gathers received outdoor air quality (AQ) variation beyond a set event threshold and arranges into n-blocks of packets P_i=[P_1, P_2..., P_N] in accordance with the node's IDs. The CH allocate 2^8 coding vectors a_i= [1, 2..., a_N] and codes them mutually by linear mixture as represented in Equation (4).

$$P_{RLNC} = \sum_{i=0}^{N} P_i * a_i \quad (4)$$

Here, i=1, 2…N, P_{RLNC} represents the coded-packet, P_i signifies the source-packet, a_i is the coding-vectors. Following the coding operation at the CH, the coded-packet is transferred to the nearby relay hop. Recovery of the source-packet from the coded-packet at the relay destination depends on the acquired amount of packet. Firstly, this involves Gaussian extinction. The header-message is then set-up to n*n matrix and eventually to (reff) reduced-row-echelon. Eventually, the source-packets are reconstructed upon evaluating a few series of underlying linear equations. The operation of decode and forward (DF) on the packets at the relays, from the event zone is realized per-hop until they are obtained at the BS. **Figure 3** illustrates the full-flowchart of the suggested E-LCDGRA.

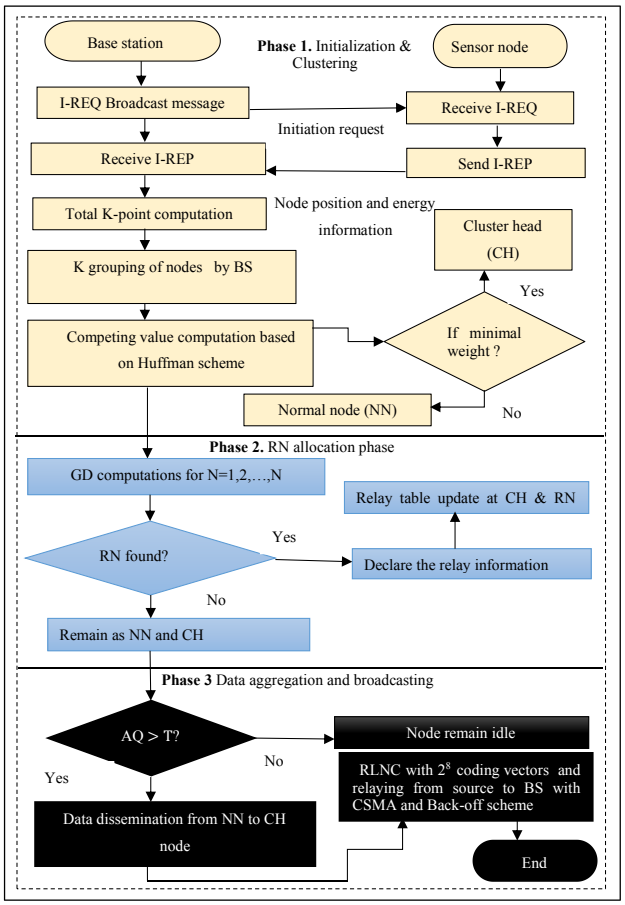


Figure 3. E-LCDGRA full-flowchart.

4. Comparative experimental simulation results

We assess E-LCDGRA performance employing MATLAB 2018b simulations. The experimental study was performed using 100 sensors spread arbitrarily across (X = 100 m, Y = 100 m) 2D interest zone with one a gateway positioned at (X = 100 m, Y = 50 m) remotely from the 2D sensing zone. Further basic parameters employed in the experimental study are offered in **Table 1**. We chose IEESEP and CERP data-gathering schemes to authenticate the soundness

of our adapted algorithm, named E-LCDGRA.

Table 1. Simulation parameters.

Parameter	Value
Sensed-traffic	Event driven
Round-time	4000
No of nodes	100
Dimension of field	X=100 m, Y=100 m
No. of BS	1
BS placement	X=100 m, Y = 50 m
Initial-Energy	5J
Data packet size	100(bytes)
E_{elec}	50 nJ/bit
Aggregation-energy	5 nJ/bit/signal
e_{mp}	0.0013pJ/bit/m
e_{fs}	100 pJ/bit/m^2
Size of broadcast packet	25 (bit)

Figure 4 measures the average network latency with each of the schemes. The latency signifies the interval between data dissemination from the source to the receivers' reception time. It comprises delays in propagation, data queuing, and data processing. It is noticeable from the experimental results obtainable in **Figure 4** that the network's average latency when using E-LCDGRA is rationally less compared to IEESEP. Again, when paying attention to the system's latency in **Figure 4**, it is indisputable that the average latency of IEESEP is relatively lesser than that of CERP.

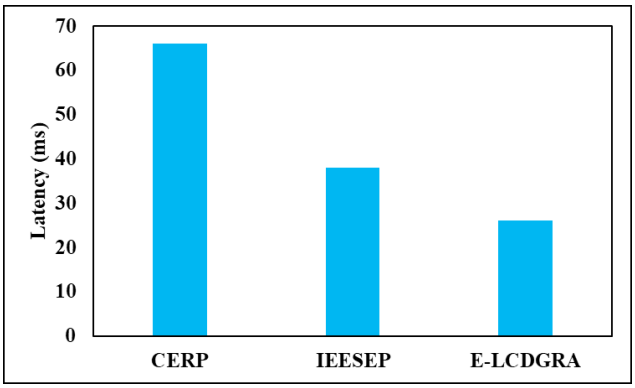


Figure 4. Network latency.

Figure 5 examines the network's average network energy consumption in data gathering. The energy consumption per round is defined by Equation (5).

$$E = \sum_{i=1}^{N} \frac{E_{ri}}{R} \quad (5)$$

where, i=1,2,…N, N denotes the over-all sensors, E represents the round energy consumption, and R signifies the total rounds, E_{ri} signifies the node's residual energy upon a round conclusion. As can be inferred from **Figure 5**, it is clear that network energy consumption for E-LCDGRA is considerably lower than IEESEP and CERP, respectively.

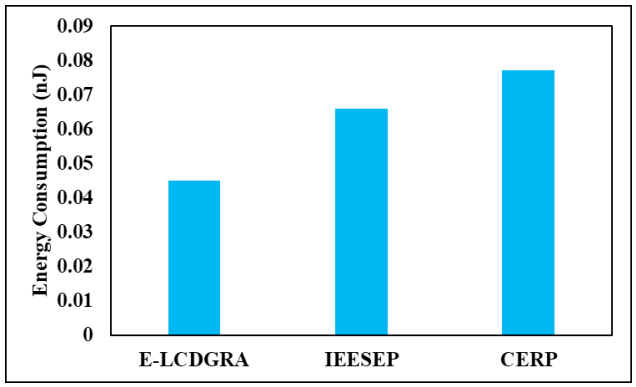


Figure 5. Energy consumption.

Figure 6 evaluates the sensor network's lifetime for each protocol. The lifetime is viewed as the epoch from initialization to the time halves of sensors drain their energy over the sensor network. It is explicit from **Figure 6** that in E-LCDGRA, halves of the sensor deplete their energy at around 3500. On the other hand, it can be seen that in IEESEP and TEEN, halves of the sensors drained their energy around 3000 and 2600.

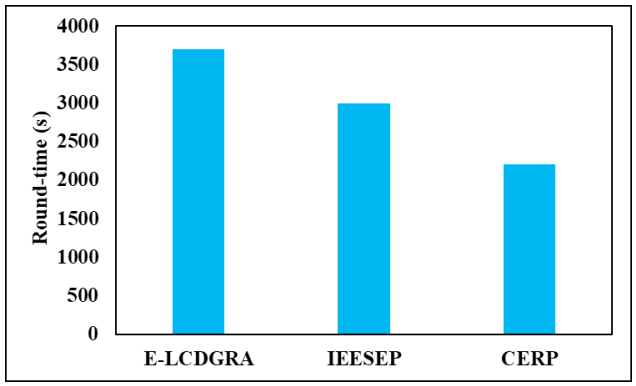


Figure 6. Network lifetime.

The above experimental results affirm that the adapted solution named E-LCDGRA incorporates the aids of a non-beacon-aided CSMA layer-2 un-slotted protocol with a back-off mechanism, co-

operative multi-hop communication, and random linear coding technology with the most suitable relays in a clustered topology. These results verify that the designed strategy does not solely lessen system latency and energy expenditure, but again augments the longevity and energy efficiency above comparable schemes. Thus, our solution, named E-LCDGRA, proves its supremacy in sufficing the need for further energy-efficient, reliable, and well-timed event dissemination in energy-constrained WSN set-ups for air quality monitoring.

5. Conclusions

Air pollution is a severe problem that has raised the concerns of communities, the public, and scientists globally. By employing sensor-based structures, we can get pollutant data in real time and implement preventive measures. However, associated with sensor structures deployed for air pollution monitoring are the problems of the constrained lifetime of sensors and poor broadcast reliability. Thus, to tackle these issues, this work proposed outdoor air quality monitoring using a sensor network with an enhanced lifetime-enhancing cooperative data gathering and relaying algorithm (E-LCDGRA). LCDGRA is a cluster-based cooperative event-driven routing scheme with dedicated relay allocation mechanisms that tackle the problems of event-driven clustered WSNs with immobile gateways. The adapted variant, named E-LCDGRA, enhances the LCDGRA algorithm by incorporating a non-beacon-aided CSMA layer-2 un-slotted protocol with a back-off mechanism. The performance of E-LCDGRA was examined with other typical clustered event-driven gathering schemes, including IEESEP and CERP, in terms of average lifetime, energy consumption, and delay.

Conflict of Interest

There is no conflict of interest.

Funding

This research received no external funding.

References

[1] Concas, F., Mineraud, J., Lagerspetz, E., et al., 2021. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. ACM Transactions on Sensor Networks (TOSN). 17(2), 1-44.

[2] Pramanik, J., Samal, A.K., Pani, S.K., et al., 2022. Elementary framework for an IoT based diverse ambient air quality monitoring system. Multimedia Tools and Applications. 81(26), 36983-37005.

[3] Alsaber, A.R., Pan, J., Al-Hurban, A., 2021. Handling complex missing data using random forest approach for an air quality monitoring dataset: A case study of Kuwait environmental data (2012 to 2018). International Journal of Environmental Research and Public Health. 18(3), 1333.

[4] Kumar, G., Agbulu, G.P., Rahul, T.V., et al., 2022. A cloud-assisted mesh sensor network solution for public zone air pollution real-time data acquisition. Journal of Ambient Intelligence and Humanized Computing. (IF 3. 662). 1-15.

[5] Lu, K.F., Wang, H.W., Li, X.B., et al., 2022. Assessing the effects of non-local traffic restriction policy on urban air quality. Transport Policy. 115, 62-74.

[6] Kim, J., Jeong, U., Ahn, M.H., et al., 2020. New era of air quality monitoring from space: Geostationary Environment Monitoring Spectrometer (GEMS). Bulletin of the American Meteorological Society. 101(1), E1-E22.

[7] Zhang, D., Woo, S.S., 2020. Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network. IEEE Access. 8, 89584-89594.

[8] Semlali, B.E.B., El Amrani, C., Ortiz, G., et al., 2021. SAT-CEP-monitor: An air quality monitoring software architecture combining complex event processing with satellite remote sensing. Computers & Electrical Engineering. 93, 107257.

[9] Karagulian, F., Barbiere, M., Kotsev, A., et al., 2019. Review of the performance of low-cost

sensors for air quality monitoring. Atmosphere. 10(9), 506.

[10] Saini, J., Dutta, M., Marques, G., 2020. A comprehensive review on indoor air quality monitoring systems for enhanced public health. Sustainable Environment Research. 30(1), 1-12.

[11] Sumathi, J., Velusamy, R.L., 2021. A review on distributed cluster based routing approaches in mobile wireless sensor networks. Journal of Ambient Intelligence and Humanized Computing. 12(1), 835-849.

[12] Lakshmanna, K., Subramani, N., Alotaibi, Y., et al., 2022. Improved metaheuristic-driven energy-aware cluster-bajsed routing scheme for IoT-assisted wireless sensor networks. Sustainability. 14(13), 7712.

[13] Maheshwari, P., Sharma, A.K., Verma, K., 2021. Energy efficient cluster based routing protocol for WSN using butterfly optimization algorithm and ant colony optimization. Ad Hoc Networks. 110, 102317.

[14] Roy, S., Das, A.K., 2014. Cluster based event driven routing protocol (CERP) for wireless sensor network. International Journal of Computer Applications. 88, 6-11.

[15] Vaiyapuri, T., Parvathy, V.S., Manikandan, V., et al., 2021. A novel hybrid optimization for cluster-based routing protocol in information-centric wireless sensor networks for IoT based mobile edge computing. Wireless Personal Communications. 127, 39-62.

[16] Moussa, N., El Belrhiti El Alaoui, A., 2021. An energy-efficient cluster-based routing protocol using unequal clustering and improved ACO techniques for WSNs. Peer-to-Peer Networking and Applications. 14(3), 1334-1347.

[17] Mazinani, A., Mazinani, S.M., Mirzaie, M., 2019. FMCR-CT: An energy-efficient fuzzy multi cluster-based routing with a constant threshold in wireless sensor network. Alexandria Engineering Journal. 58(1), 127-141.

[18] Pavani, M., Trinatha Rao, P., 2019. Adaptive PSO with optimised firefly algorithms for secure cluster-based routing in wireless sensor networks. IET Wireless Sensor Systems. 9(5), 274-283.

[19] Barzin, A., Sadegheih, A., Zare, H.K., et al., 2020. A hybrid swarm intelligence algorithm for clustering-based routing in wireless sensor networks. Journal of Circuits, Systems and Computers. 29(10), 2050163.

[20] Senthil, G.A., Raaza, A., Kumar, N., 2022. Internet of Things energy efficient cluster-based routing using hybrid particle swarm optimization for wireless sensor network. Wireless Personal Communications. 122(3), 2603-2619.

[21] Chanak, P., Banerjee, I., Sherratt, R.S., 2020. A green cluster-based routing scheme for large-scale wireless sensor networks. International Journal of Communication Systems. 33(9), e4375.

[22] Dayal, K., Bassoo, V., 2022. Fast-converging chain-cluster-based routing protocols using the Red-Deer Algorithm in Wireless Sensor Networks. Applied Computing and Informatics. ahead-of-print. ahead-of-print.
DOI: https://doi.org/10.1108/ACI-10-2021-0289

[23] Yadav, R.N., Misra, R., Saini, D., 2018. Energy aware cluster based routing protocol over distributed cognitive radio sensor network. Computer Communications. 129, 54-66.

[24] Dixit, E., Jindal, V., 2022. IEESEP: An intelligent energy efficient stable election routing protocol in air pollution monitoring WSNs. Neural Computing and Applications. 34(13), 10989-11013.

[25] Agbulu, G.P., Kumar, G.J.R., Juliet, A.V., 2020. A lifetime-enhancing cooperative data gathering and relaying algorithm for cluster-based wireless sensor networks. International Journal of Distributed Sensor Networks. 16(2), 1550147719900111.

BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

Data Analytics of an Information System Based on a Markov Decision Process and a Partially Observable Markov Decision Process

Lidong Wang*, Reed L. Mosher, Terril C. Falls, Patti Duett

Institute for Systems Engineering Research, Mississippi State University, Vicksburg, MS 39180, USA

ABSTRACT

Data analytics of an information system is conducted based on a Markov decision process (MDP) and a partially observable Markov decision process (POMDP) in this paper. Data analytics over a finite planning horizon and an infinite planning horizon for a discounted MDP is performed, respectively. Value iteration (VI), policy iteration (PI), and Q-learning are utilized in the data analytics for a discounted MDP over an infinite planning horizon to evaluate the validity of the MDP model. The optimal policy to minimize the total expected cost of states of the information system is obtained based on the MDP. In the analytics for a discounted POMDP over an infinite planning horizon of the information system, the effects of various parameters on the total expected cost of the information system are studied.

Keywords: Predictive modelling; Information system; MDP; POMDP; Cybersecurity; Q-learning

1. Introduction

Cyberattacks against federal information systems in the USA are more and more sophisticated. The probability of grave damages keeps increasing in spite of efforts and the use of substantial resources. There are challenges in completely aggregating heterogeneous data from various security tools, analyzing the collected data, prioritizing remediation activities, and reporting in an approach to directing a suitable response [1]. Cyberspace is a dynamic environment. Targets are not always static. No offensive or defensive capability keeps being indefinitely effective. There is no permanent advantage [2].

Cyber attackers generally have advantages over the defender of an information system. The advantages lie in: 1) Attackers can choose the place and time of an attack; 2) Attackers can only exploit a sin-

*CORRESPONDING AUTHOR:
Lidong Wang, Institute for Systems Engineering Research, Mississippi State University, Vicksburg, MS 39180, USA; Email: lidong@iser.msstate.edu

ARTICLE INFO
Received: 26 January 2023 | Revised: 17 February 2023 | Accepted: 20 February 2023 | Published Online: 28 February 2023
DOI: https://doi.org/10.30564/jcsr.v5i1.5434

CITATION
Wang, L.D., Mosher, R.L., Falls, T.C., et al., 2023. Data Analytics of an Information System Based on a Markov Decision Process and a Partially Observable Markov Decision Process. Journal of Computer Science Research. 5(1): 21-30. DOI: https://doi.org/10.30564/jcsr.v5i1.5434

gle vulnerability while the defender has a much more costly task of mitigating all kinds of vulnerabilities. Human-centered cyber-defense practices have not kept pace with threats of targeting and attacking organizations. An integrated approach is needed to speed up detection or responses and slow down attacks. Security automation and intelligence sharing can reduce the defender's costs and save time. Information sharing helps improve the efficiency in detecting and responding to cyberattacks [3].

There are four major categories of attacks [4-6]: 1) Denial of service—trying to stop legitimate users from utilizing services; 2) Probe—trying to get the information of a target host; 3) User to Root (U2R)—unauthorized access to privileges of a local super-user (root); and 4) Remote to Local (R2L)—unauthorized access from a remote machine. Signature-based detection and anomaly-based detection are the two main methods of detecting attacks. Signature-based detection uses predefined attack specifications that are clear and distinct signatures. The database of signatures needs to be updated when there are new signatures. Human security experts are generally required to analyze data related to attacks manually and formulate specifications regarding attacks [7]. Anomaly-based detection is also called behavior-based detection. It models behaviors of the network, computer systems, and users; and raises an alarm when there is a deviation from normal behaviors [8].

Many cyberattacks are characterized by a high level of sophistication. Typically, an advanced persistent threat (APT) is a kind of attack targeting an asset or a physical system with high values. APT attackers frequently leverage stolen credentials of users or zero-day exploits to avoid triggering alerts. This kind of attacks could continue over an extended period of time [9]. Artificial intelligence (AI) or intelligent agents are needed to fight attack, especially an APT. Therefore, the mechanisms of cyber defense should be 1) increasingly intelligent, 2) very flexible, and 3) robust enough to detect various threats and mitigate them. Much research has been done on intrusion detection and prevention systems. Various methods and algorithms of artificial intelligence have been used for cybersecurity. The algorithms include support vector machines (SVM), convolution neural networks, recursive neural networks, general artificial neural networks (ANN), Q-learning (QL), decision trees (DT), *k*-means, *k*-nearest neighbors (*k*-NN), etc. [10]. MDP and POMDP are used in this paper because they deal with the optimal policy or actions based on computed benefits or costs.

During an attack, both the attacker and the defender are in the process of learning about each other. The knowledge evolution of the attacker and the defender indicates the process of learning. A defender's knowledge includes, for example, attackers' objectives, methods utilized, possible technical levels, etc. An attacker's knowledge can be the topology of a defender's network or information system, the operating system version and applications running on servers, etc. When an attack is detected, the defender can expel the attacker or keep it in the information system in order to observe or learn about it. The policy of always expelling the attacker is not optimal in many situations. There is a trade-off between the opportunity of learning about the attacker and the risk of the attacker's damage during the defender's learning process [11]. MDP and POMDP can handle the trade-off and decide on optimal policies or actions.

This paper aims to conduct analytics of an information system based on an MDP and a POMDP. Various methods and algorithms were used, including value iteration (VI), policy iteration (PI), and Q-learning in the analytics of a discounted MDP over an infinite planning horizon to evaluate the MDP model validity and parameters in the model. In the modelling of a discounted POMDP over an infinite planning horizon, the effects of several important parameters on the total expected reward of the system were studied. The data analytics of the MDP and POMDP in this paper was conducted using the *R* language and its functions. The organization of this paper is as follows: the next section introduces the methods of MDP; Section 3 introduces the methods of POMDP; Section 4 presents an MDP model of an information system; Section 5 shows the analytics

of the information system based on MDP; Section 6 presents the analytics of the information system based on POMDP; and the final section is the conclusions.

2. Markov decision process

An MDP can be defined by a tuple <*S*, *A*, *P*, *R*, γ> [12-14]: *S* refers to a set of states; *A* is a set of actions; *P* represents a transition probability matrix that describes the transition from state *s* to state *s'* ($s \in S, s' \in S$) after action *a* ($a \in A$); *R* refers to the immediate reward after action *a*; and γ ($0 < \gamma < 1$) is a discounted reward factor. Solving an MDP is often a process of finding an optimal policy to maximize the total expected reward or minimize the total expected cost.

Policy iteration, value iteration, and Q-learning are often used to obtain an optimal policy for an MDP. Data analytics results based on the algorithms of the three methods may be noticeably different, or there can be convergence problems during iterations if the MDP model is not reasonable due to unsuitable model parameters or an incorrect model structure. Therefore, the three methods are employed in this paper, and results are compared to evaluate the model's validity.

PI tries to find a better policy (compared to the previous policy). An iterative process of policy evaluation and policy improvement is stopped when two successive policy iterations result in the same policy, indicating the optimal policy is achieved. The policy iteration is described in Algorithm 1 [15,16]. $P(s, a, s')$ is the probability of the transition. *R*(*s, a, s'*) is the immediate transition reward from the state *s* to the state *s'* after the action *a*. *V*(*s*) and *V*(*s'*) are the expected total reward of state *s* and state *s'*, respectively. $\pi(s)$ is an optimal policy of state *s*.

An optimal policy of the MDP can also be achieved by utilizing VI [15,17]. The stopping criterion is that the value difference of two successive iterative steps is less than the tolerance τ (a very small positive number). Algorithm 2 shows the value iteration process.

Algorithm 1. Policy Iteration.

1	Initial policy Choose an initial policy arbitrarily for all $s \in S$ $V(s) \in R$ and $\pi(s) \in A$
2	Policy evaluation Repeat $\Delta \leftarrow 0$ For each $s \in S$ $v \leftarrow V(s)$ $V(s) \leftarrow max_a \sum_{s'} P(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s'))$ $\Delta \leftarrow \max(\Delta, \lvert V(s) - v \rvert)$ until $\Delta < \tau$ (a very small positive number)
3	Policy improvement routine For each state *s* $\pi(s) \leftarrow argmax_a(\sum_{s'} P(s, a, s')(R(s, a, s') + \gamma V(s')))$
4	Stopping rule If policy is stable, then stop; else go to step 2

Algorithm 2. Value Iteration.

1	Initialization Select *V*(*s*) arbitrarily (e.g., *V*(*s*) = 0 for all $s \in S$)
2	Value iteration process Repeat $\Delta \leftarrow 0$ For each $s \in S$ $v \leftarrow V(s)$ $V(s) \leftarrow max_a \sum_{s'} P(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s'))$ $\Delta \leftarrow \max(\Delta, \lvert V(s) - v \rvert)$ until $\Delta < \tau$
3	Output the optimal policy and the maximal values of *V*(*s*)

Q-learning [17,18] enables an agent to learn the Q-value function which is an optimal action-value function. It can be employed to solve a discounted MDP. Specifically, it is used to compute the expected total reward (or cost) and find the optimal policy in this paper. It can be used to perform data analytics and simulation of a discounted MDP over an infinite planning horizon if the number of iterations to perform is large enough. A Q-learning algorithm is shown in Algorithm 3. *Q*(*s,a*) is the action-value function. $\beta \in (0, 1)$ is the learning rate and it is often chosen to be decreased appropriately, e.g., $\beta = 1/\sqrt{(n+2)}$ (*n* is the iteration step number or the epoch number). The iterative process and the Q-learning update continue until the final step of an episode. The best action *a* at state *s* is chosen according to the optimal policy $\pi(s)$.

Algorithm 3. Q-learning.

1	Initialization Initialize $Q(s,a)$ arbitrarily (e.g., $Q(s,a) = 0$, $\forall s \in S, \forall a \in A$)
2	Iterative process and Q-learning update Repeat For each $s \in S$ $Q(s,a) \leftarrow \sum_{s'} P(s,a,s')(R(s,a,s') + \gamma V(s'))$ Q-learning update is as follows: $Q(s,a) \leftarrow (1-\beta)Q(s,a) + \beta[R(s,a,s') + \gamma \max_a Q(s',a)]$ until the final step of episode
3	Output the optimal policy and maximal values of states

3. Partially observable Markov decision process

In many applications, a POMDP is a more realistic model than the classic MDP [19]. The transition model $P(s'|s, a)$, actions $A(s)$, and the reward function $R(s, a, s')$ in a POMDP are the same elements as those in an MDP. The optimal action of the POMDP depends only on the agent's current belief state. The agent does not know its real state; all it knows is the belief state [20]. Besides the three elements, there are a set of observations $O = \{o_1, o_2, ..., o_k\}$ and a set of conditional observation probabilities $B(o|s', a)$ in a POMDP [21].

If b was the previous belief state, and the agent takes action a and then perceives evidence o, then the new belief state [20] is obtained using the following formula:

$$b'(s') = \alpha P(o|s') \Sigma_s P(s'|s, a)\, b(s) \quad (1)$$

where a is a normalizing constant, making the belief state sum to 1.

The optimal value of any belief state b is the infinite expected sum of discounted rewards starting in state b, and executing the optimal policy. The value function, $V^*(b)$, is expressed as follows [22]:

$$V^*(b) = max_{a\in A}\left[b(s)R(s,a) + \gamma \sum_{o\in O} P(o|b,a)V^*(b')\right] \quad (2)$$

4. A Markov decision process model of the information system

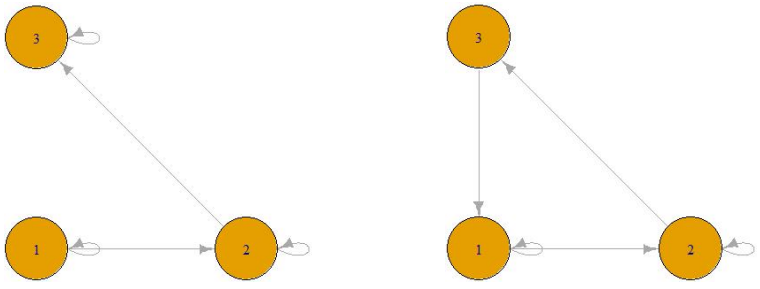
4.1 The structure of the MDP model

The information system has the following states: State 1—no attacker is connected to the information system; state 2—an attacker is connected to the information system, but it has not been detected; and state 3—the attacker is detected. The defender needs to make a decision: wait (no action) or expel only when an attack is detected (state 3). After an expelling action, the system will return to state 1.

The MDP model of the information system is established. State transitions among three states (states 1-3) of two decisions are shown in **Figure 1**.

4.2 State transitions and rewards

Transitions among states in the created MDP model of the information system rely on decisions and there are two main probabilities P_1 and P_2. P_1 is the probability of the transition from state 1 (no attacker's connection) to state 2 (connected). P_2 is the probability of the transition from state 2 to state 3 (detected). There are no transitions from state 1 to state 3 directly and no transitions from the state 3 to the state 2. The probability of a transition from state



(a) (b)

Figure 1. State transitions of two decisions: (a) decision 1 (wait) and (b) decision 2 (expel).

3 to state 1 is 0 for decision 1 and 1 for decision 2. The probability matrix of state transitions P_d and the reward matrix R_d for the two decisions are expressed as follows:

1) P_d and R_d for decision 1 are:

$$P_d = \begin{bmatrix} 1-P_1 & P_1 & 0 \\ 0 & 1-P_2 & P_2 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} = \begin{bmatrix} 0 & -C_a & 0 \\ 0 & -C_a & B_i - C_a \\ 0 & 0 & B_i - C_a \end{bmatrix} \quad (4)$$

where C_a is the cost due to attacking and B_i is the defender's benefit due to collecting information during the learning process of knowing about the attack.

2) P_d and R_d for decision 2 are:

$$P_d = \begin{bmatrix} 1-P_1 & P_1 & 0 \\ 0 & 1-P_2 & P_2 \\ 1 & 0 & 0 \end{bmatrix} \quad (5)$$

$$R_d = \begin{bmatrix} 0 & r_{12} & 0 \\ 0 & r_{22} & r_{23} \\ r_{31} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -C_a & 0 \\ 0 & -C_a & B_i - C_a \\ -C_e & 0 & 0 \end{bmatrix} \quad (6)$$

where C_e is the cost due to expelling.

5. Data analytics of the information system based on the MDP

5.1 Analytics based on MDP over an infinite planning horizon

Let $P_1 = 0.15$, $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$. The analytics of the information system with a discount $\gamma = 0.85$ over an infinite planning horizon is conducted. Policy iteration and value iteration are used in the data analytics and the obtained optimal policies in both the two methods are d (1, 1, 2), indicating that decision 1, decision 1, and decision 2 are made on the state 1, the state 2, and the state 3, respectively. The total expected costs of the two methods and Q-learning are listed in **Table 1** to evaluate the model validity in this paper. Gauss-Seidel's algorithm is employed in VI for an improved convergence speed. The accuracy is also improved compared with the result of Jacob's algorithm. In Q-learning, the learning rate β is set to $1/\sqrt{n+2}$ in this paper and N is the number of iterations to perform. The results of policy iteration and the Gauss-Seidel method are the same and are close to that of Q-learning, which indicates the parameters in the MDP model are reasonable, and the created model is valid.

Table 1. Total expected costs of three states in the information system based on various algorithms over an infinite planning horizon ($\gamma = 0.85$).

Algorithms	C_1	C_2	C_3
VI (Jacob' algorithm)	12.68322	21.77953	11.77344
VI (Gauss-Seidel's algorithm)	12.73186	21.82816	11.82208
PI	12.73186	21.82816	11.82208
Q-learning ($N = 120{,}000$)	12.67515	21.63394	11.90482

5.2 Analytics over a finite planning horizon

The total expected costs of three states (states 1-3) are calculated utilizing the VI algorithm over a 40-step planning horizon with and without a discount, respectively. The rewards (the negative values of the costs in this paper) at the end of the planning horizon are set to 0 for three states for the beginning of the backward recursion of the VI. **Table 2** and **Table 3** show the computation results. $C_1(n)$, $C_2(n)$, and $C_3(n)$ represent the total expected cost at step n for the state 1, the state 2, and the state 3, respectively. It is shown that the total expected costs $C_1(n)$, $C_2(n)$, and $C_3(n)$ in **Table 2** are very close to C_1, C_2, and C_3 for infinite planning horizon in **Table 1**, respectively when Epoch $n \leq 10$ for a 40-step planning horizon ($\gamma = 0.85$).

5.3 Analytics of the information system with various parameters of the transition probability

Analytics of the information system with various state transition probability parameters P_1 and P_2 is performed based on the PI over an infinite planning horizon. The following data are utilized: $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$, and $\gamma = 0.85$. The total expected cost C_i ($i = 1, 2, 3$) for states 1-3 at various P_1 is analyzed and the result is shown in **Figure 2**. All the values of C_1, C_2, and C_3 are increased with the increase of P_1.

Table 2. Total expected costs of three states computed using the VI algorithm over a 40-step planning horizon ($\gamma = 0.85$).

Epoch n	$C_1(n)$	$C_2(n)$	$C_3(n)$
0	12.7065	21.8028	11.7967
5	12.6746	21.7710	11.7649
10	12.6029	21.6992	11.6931
15	12.4412	21.5375	11.5315
20	12.0769	21.1731	11.1672
25	11.2565	20.3509	10.3471
30	9.4191	18.4836	8.5165
33	7.3930	16.3143	6.5226
35	5.4787	14.0296	4.6918
36	4.3432	12.4763	3.6503
37	3.1180	10.5134	2.5912
38	1.8720	7.9649	1.6375
39	0.75	4.55	1.00
40	0	0	0

Table 3. Total expected costs of three states computed using the VI algorithm over a 40-step planning horizon ($\gamma = 1.0$).

Epoch n	$C_1(n)$	$C_2(n)$	$C_3(n)$
0	85.3155	93.7710	76.7897
5	75.1760	83.7475	66.7897
10	64.9085	73.6947	56.7897
15	54.4104	63.5756	46.7897
20	43.5248	53.3072	36.7897
25	32.0626	42.7023	26.7897
30	19.9701	31.3391	16.7897
33	12.6354	23.7993	10.7897
35	7.9649	18.2667	6.7897
36	5.7897	15.2911	4.7946
37	3.7946	12.0945	3.0700
38	2.0700	8.5675	1.7500
39	0.75	4.55	1.00
40	0	0	0

Let $P_1 = 0.15$, $C_e = 1$, $B_i = 3$, $C_a = 5$, and $\gamma = 0.85$. The PI over an infinite planning horizon is utilized. The total expected cost C_i ($i = 1, 2, 3$) at various P_2 is shown in **Figure 3**. All the values of C_1, C_2, and C_3 are decreased with the increase of P_2.

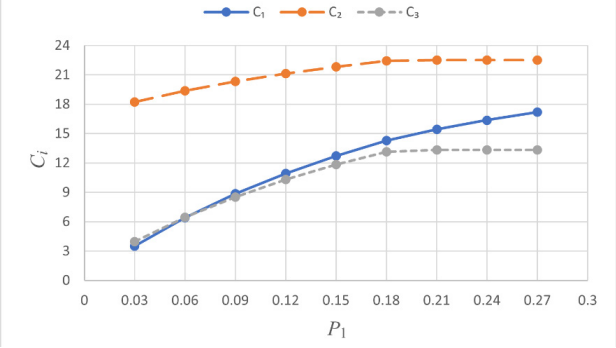


Figure 2. Total expected cost C_i ($i = 1, 2, 3$) at various P_1.

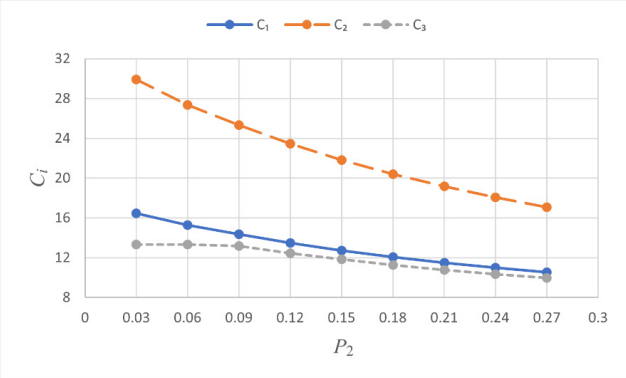


Figure 3. Total expected cost C_i ($i = 1, 2, 3$) at various P_2.

5.4 Analytics of the information system with various transition cost parameter C_a

Analytics of the information system with various transition cost parameters C_a is performed based on the PI over an infinite planning horizon. The following data are used: $P_1 = 0.15$, $P_2 = 0.15$, $C_e = 1$, $B_i = 3$, and $\gamma = 0.85$. **Figure 4** illustrates the total expected cost C_i ($i = 1, 2, 3$) at various C_a. The greater the value of C_a, the larger the value of the expected total cost C_i.

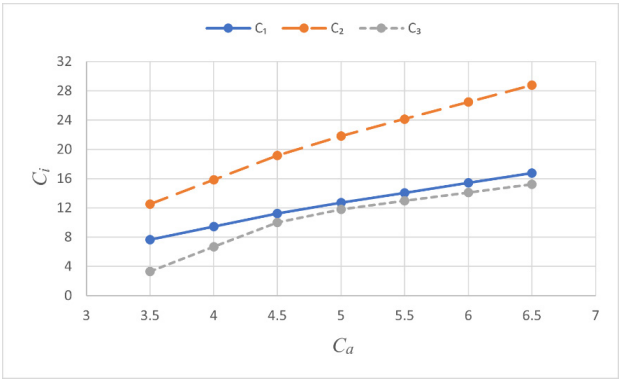


Figure 4. Total expected cost C_i ($i = 1, 2, 3$) at various C_a.

6. Data analytics of the information system based on POMDP

6.1 Analytics based on the POMDP over an infinite planning horizon

Analytics of the information system is performed based on a discounted POMDP over an infinite planning horizon. The following data are utilized: P_1 = 0.15, P_2 = 0.15, C_e = 1, B_i =3, C_a = 5, and γ = 0.85. The following solution methods or algorithms are used in solving the POMDP problem: "grid", "enum", "twopass", "witness", "incprune", and "SARSOP" [23]. The total expected cost C_t is shown in **Table 4**, indicating that the result of SARSOP is very close to the results of the other five methods (with the same results).

6.2 The effects of various parameters on POMDP solutions

The following data are used to study the effects of various parameters on the total expected cost C_t: C_e = 1, B_i =3, and γ = 0.85. **Figure 5** shows the effect of the connecting probability P_1 on C_t at various P_2 (0.03, 0.15, and 0.27) when C_a = 5. **Figure 6** shows the effect of P_1 on C_t at various C_a (3.5, 5.0, and 6.5) when P_2 = 0.15. It is shown that C_t is increased with an increase of P_1. Similarly, the effects of the detecting probability P_2 on the total expected cost C_t are studied. The results are shown in **Figure 7 and Figure 8**. It is shown that $C_t$is decreased with an increase of P_2. **Figure 9** shows the effect of C_a on C_t at various P_1 (0.03, 0.15, and 0.27) when P_2 = 0.15. **Figure 10** shows the effect of C_a on C_t at various P_2 (0.03, 0.15, and 0.27) when P_1 = 0.15. It is shown that C_t is increased with the increase of C_a.

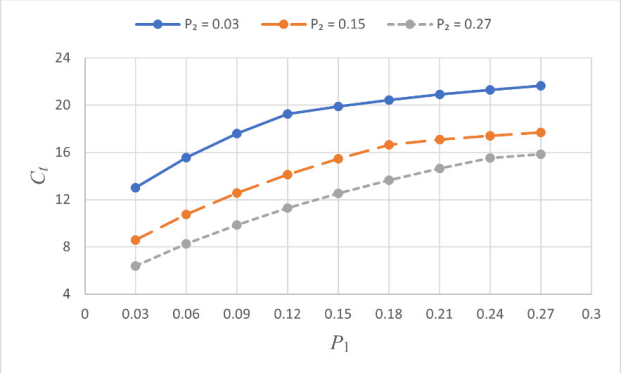


Figure 5. The effect of P_1 on C_t at various P_2 when C_a = 5.

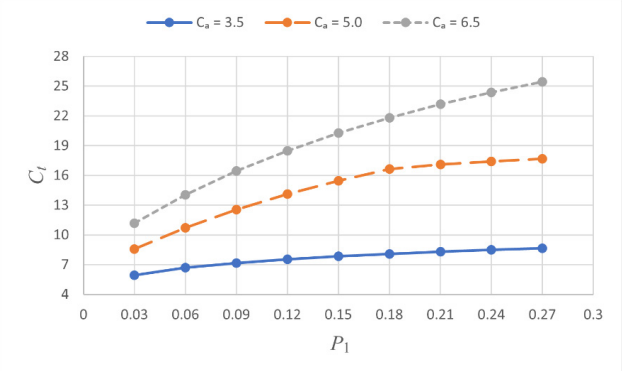


Figure 6. The effect of P_1 on C_t at various C_a when P_2 = 0.15.

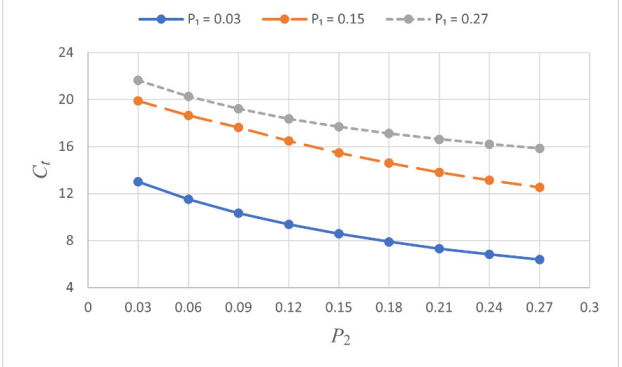


Figure 7. The effect of P_2 on C_t at various P_1 when C_a = 5.0.

Table 4. The total expected cost C_t based on six various methods.

Methods	grid	enum	twopass	witness	incprune	SARSOP
C_t	15.46070	15.46070	15.46070	15.46070	15.45570	15.46073

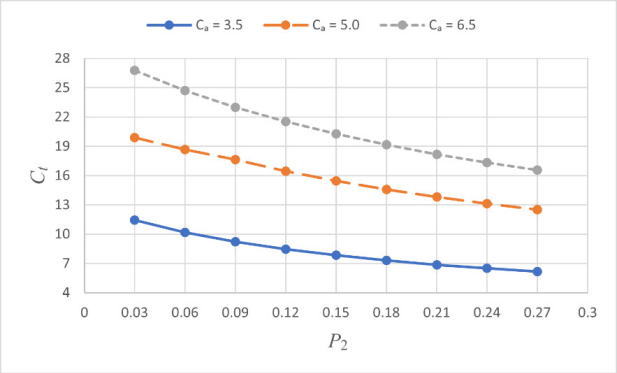


Figure 8. The effect of P_2 on C_t at various C_a when $P_1 = 0.15$.

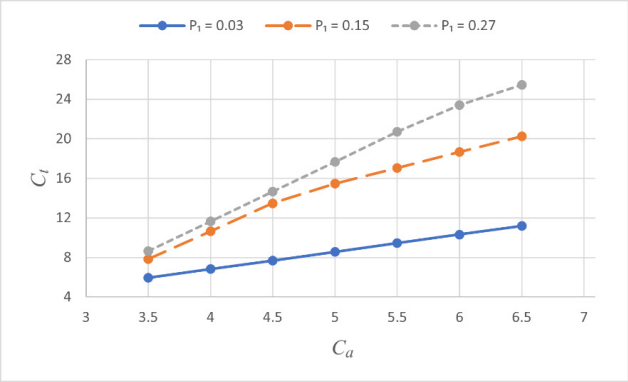


Figure 9. The effect of C_a on C_t at various P_1 when $P_2 = 0.15$.

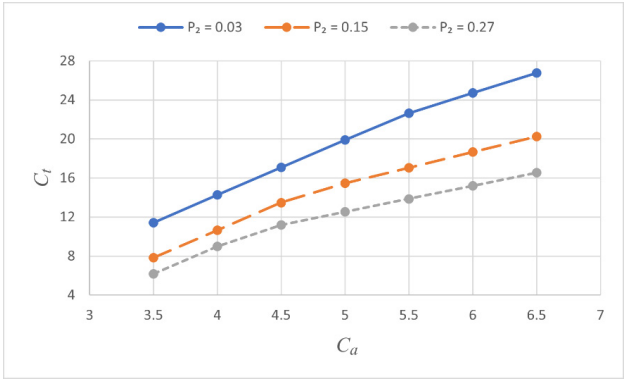


Figure 10. The effect of C_a on C_t at various P_2 when $P_1 = 0.15$.

7. Conclusions

Data analytics of an information system based on the MDP demonstrates that the algorithms in this paper are effective in achieving optimal policies to minimize the total expected costs of states of the information system. These algorithms are effective in analytics over a finite planning horizon and an infinite planning horizon (for a discounted MDP). The VI (Gauss-Seidel's algorithm) and the PI achieve the same results, and the result of Q-learning is very close to the results of the VI and the PI, indicating the MDP model is valid. The pros of data analytics of the information system based on the MDP lie in: 1) Multiple methods can be used to check the validity of the created MDP model; 2) It is convenient to perform predictive modelling and study the effects of various parameters on the total expected cost of the information system.

One of the main cons of the MDP-based method is that the state uncertainty is not considered while this problem is fixed in the POMDP method. In the analytics of a discounted POMDP (over an infinite planning horizon) of the information system, the total expected cost of the information system is increased with an increase in the connecting probability and is decreased with an increase in the detecting probability. The cost caused by the attacker is a primary factor in increasing the total expected cost of the information system.

Conflict of Interest

There is no conflict of interest.

Funding

This research received no external funding.

Acknowledgement

This paper is based upon work supported by Mississippi State University, USA.

References

[1] AlSadhan, T., Park, J.S. (editors), 2021. Leveraging information security continuous monitoring to enhance cybersecurity. 2021 International Conference on Computational Science and Computational Intelligence (CSCI); 2021 Dec 15-17; Las Vegas, NV, USA. USA: IEEE. p. 753-759.

[2] United States Cyber Command, 2018. Achieve and Maintain Cyberspace Superiority, Command Vision for U.S. Cyber Command [Internet]. Available from: https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20 Vision%20 April%202018.pdf?ver=2018-06-14-152556-010

[3] Wendt, D., 2019. Addressing both sides of the cybersecurity equation. Journal of the Cyber Security & Information Systems Information Analysis Center. 7(2).

[4] Anuar, N.B., Sallehudin, H., Gani, A., et al., 2008. Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree. Malaysian Journal of Computer Science. 21(2), 101-115.

[5] Kukielka, P., Kotulski, Z. (editors), 2008. Analysis of different architectures of neural networks for application in intrusion detection systems. 2008 International Multiconference on Computer Science and Information Technology; 2008 Oct 20-22; Wisla, Poland. USA: IEEE. p. 807-811.

[6] Faisal, M.A., Aung, Z., Williams, J.R., et al., 2012. Securing advanced metering infrastructure using intrusion detection system with data stream mining. In: Chau, M., Wang, G.A., Yue, W.T., et al. (editors), intelligence and security informatics. PAISI 2012. Lecture Notes in Computer Science. Springer, Berlin: Heidelberg. pp. 96-111. DOI: https://doi.org/10.1007/978-3-642-30428-6_8

[7] Raiyn, J., 2014. A survey of cyber attack detection strategies. International Journal of Security and Its Applications. 8(1), 247-256.

[8] Singh, J., Nene, M.J., 2013. A survey on machine learning techniques for intrusion detection systems. International Journal of Advanced Research in Computer and Communication Engineering. 2(11), 4349-4355.

[9] Cardenas, A.A., Manadhata, P.K., Rajan, S.P., 2013. Big data analytics for security. IEEE Security & Privacy. 11(6), 74-76.

[10] Wiafe, I., Koranteng, F.N., Obeng, E.N., et al., 2020. Artificial intelligence for cybersecurity: A systematic mapping of literature. IEEE Access. 8, 146598-146612.

[11] Bao, N., Musacchio, J., 2009. Optimizing the decision to expel attackers from an information system. 2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton); 2009 Sep 30-Oct 2; Monticello, IL, USA. USA: IEEE. p. 644-651.

[12] Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of machine learning. Adaptive computation and machine learning. MIT Press: USA.

[13] Alsheikh, M.A., Hoang, D.T., Niyato, D., et al., 2015. Markov decision processes with applications in wireless sensor networks: A survey. IEEE Communications Surveys & Tutorials. 17(3), 1239-1267.

[14] Chen, Y., Hong, J., Liu, C.C., 2018. Modeling of intrusion and defense for assessment of cyber security at power substations. IEEE Transactions on Smart Grid. 9(4), 2541-2552.

[15] van Otterlo, M., Wiering, M., 2012. Reinforcement learning and Markov decision processes. Reinforcement Learning. Springer, Berlin: Heidelberg. pp. 3-42.

[16] Sutton R.S., Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press: USA.

[17] Zanini, E., 2014. Markov Decision Processes [Internet]. Available from: https://www. lancaster. ac. uk/pg/zaninie. MDP. pdf

[18] Liu, D., Khoukhi, L., Hafid, A. (editors), 2017. Data offloading in mobile cloud computing: A Markov decision process approach. 2017 IEEE International Conference on Communications (ICC); 2017 May 21-25; Paris, France. USA: IEEE. p. 1-6.

[19] Xiang, X., Foo, S., 2021. Recent advances in deep reinforcement learning applications for solving partially observable Markov decision processes (POMDP) problems: Part 1—fundamentals and applications in games, robotics and natural language processing. Machine Learning and Knowledge Extraction. 3(3), 554-581.

[20] Russell, S.J., Norvig, P., 2021. Artificial intelligence a modern approach, 4th edition. Pearson Education, Inc: UK.

[21] Kurniawati, H., Hsu, D., Lee, W.S., 2008. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. Robotics: Science and systems. MIT Press: USA. pp. 65-72.

[22] Cassandra, A.R., Kaelbling, L.P., Littman, M.L., 1994. Acting optimally in partially observable stochastic domains. Aaai. 94, 1023-1028.

[23] Kamalzadeh, H., Hahsler, M., 2019. POMDP: Introduction to Partially Observable Markov Decision Processes [Internet]. Available from: https://mran.revolutionanalytics.com/snapshot/2020-04-25/web/packages/pomdp/vignettes/POMDP.pdf

BILINGUAL PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
https://journals.bilpubgroup.com/index.php/jcsr

ARTICLE

On Software Application Database Constraint-driven Design and Development

Christian Mancas**, Cristina Serban, Diana Christina Mancas

Math. & Computer Science Department, Ovidius University, Constanta, 900720, Romania

ABSTRACT

This paper presents a methodology driven by database constraints for designing and developing (database) software applications. Much needed and with excellent results, this paradigm guarantees the highest possible quality of the managed data. The proposed methodology is illustrated with an easy to understand, yet complex medium-sized genealogy software application driven by more than 200 database constraints, which fully meets such expectations.

Keywords: Database constraint-driven design and development; Database constraint; Data plausibility; Software architecture; Design and development; The (elementary) mathematical data model; MatBase

1. Introduction

Software design and development research started as a mathematical branch [1,2]. Since then and to the end of the previous millennia, the emphasis was put on tackling complexity and delivering high quality, derived from rigor, and user-friendliness software [3,4].

Despite good fundamental textbooks on software engineering [5-8], unfortunately, the state of the art in the field became largely dominated by technologies and glossy graphic user interfaces (GUI) [9], instead of principles and sound methodology. However, fortunately, there is still research and results inspired by the roots of this discipline. Among them, we were always interested in *constraint-driven approaches*.

1.1 Literature survey

Almost three decades ago, for example, Hoog et al. [3] put it forward as an alternative to the waterfall model. Then, Lano [10] added constraints to UML class diagrams and state machines in the framework of

*CORRESPONDING AUTHOR:
Christian Mancas, Math. & Computer Science Department, Ovidius University, Constanta, 900720, Romania; Email: christian.mancas@gmail.com

ARTICLE INFO
Received: 14 February 2023 | Revised: 28 February 2023 | Accepted: 1 March 2023 | Published Online: 10 March 2023
DOI: https://doi.org/10.30564/jcsr.v5i1.5476

CITATION
Mancas, C., Serban, C., Mancas, D.C., 2023. On Software Application Database Constraint-driven Design and Development. Journal of Computer Science Research. 5(1): 31-45. DOI: https://doi.org/10.30564/jcsr.v5i1.5476

model-driven development (MDD), advocating for constraint-driven development (CDD). In the aftermath, Demuth et al. [11] went further and explored constraint-driven modeling (CDM), extended by Rebmann et al. [12] who proposed to automate the generation of model constraints instead of generating entire models.

In parallel, constraint-driven approaches were considered as well in narrower subfields of software engineering. For example, Siddiqui [13] proposed his *Pike*, a tool for checking code conformance to specifications; Shrotri et al. [14] use it for Machine Learning; Ciortuz [15] applies it to concurrent parsing of a natural language.

Moreover, such approaches are used outside the software engineering realm as well. For example, in linguistics, Kumaran [16] extends correspondingly Noam Chomsky's *Agree*, while, in PCB hardware design, OrCAD [17] makes heavy use of the constraint-driven paradigm.

Getting back to software engineering, let us first note that most of the applications designed, developed, maintained, and used are database (db) ones: extremely few software applications of today are not managing databases (dbs). However, db constraints are not systematically considered in software engineering design and development approaches anymore. Moreover, what is very intriguing for us is the spreading of the JSON technology, which gives the false impression that there is no need for db design anymore: you just design objects for the applications and JSON is automatically mapping them into db tables, with all needed constraints.

We advocate a dual approach: you should carefully design and implement a db and then use an advanced tool of the 5th generation of programming languages, e.g., *MatBase* [18], to automatically generate accordingly the software application for managing that db. It is true that the Relational Data Model (RDM) [19,20], which is powering most of today's DB Management Systems (DBMS), as well as the NoSQL datastores are not at all suited for such an approach: RDM provides only six constraint types, while NoSQL, practically, only one of them. This is probably why even otherwise excellent recent textbooks on db software application design like, for example, the one by Kleppmann [21], is almost not even mentioning db constraints.

In fact, while software engineering is still craftsmanship, dbs are pure applied math, namely the naïve algebraic theory of sets, relations, and functions, plus the first-order predicate logic (FOPL). In particular, *db constraints* are formalized by closed FOPL expressions, while db queries by the open ones [20] (recall that a FOPL expression is closed whenever all of its variable occurrences are bound to at least one quantifier and open when at least one of them is free, i.e. not bound to any quantifier; for example, all variable occurrences within a SQL SELECT clause are free, while all those in either WHERE or HAVING ones are bounded to a universal quantifier).

MatBase is a prototype intelligent db and knowledge base management system, based mainly on the (Elementary) Mathematical Data Model ((E)MDM) [22], but also on the Entity-Relationship (E-R) one (E-RDM) [20,23], RDM, and Datalog [19,24]. Its (E)MDM GUI accepts mathematical db schemes, translates them into both RDM and E-RDM ones, and automatically generates corresponding db software applications for managing them.

(E)MDM provides 73 constraint types on sets, relations, and functions (that includes, either explicitly, or implicitly, the 6 relational ones provided by the RDM). All these 73 types belong to the Horn clauses class, the largest FOPL one for which the implication problem is decidable.

1.2 Paper outline

MatBase's strategy to enforce constraints (which was manually used by Mancas [9]), based on our proposed DB Constraint-Driven Design and Development (DBCDDD) approach, is presented in the next section of this paper.

The third section presents and discusses the results of applying it to an interesting sub-universe centered around the genealogy trees. The paper ends with conclusions and references.

2. The DB constraint-driven design and development approach in software engineering

2.1 Proposed methodology

The DB Constraint-Driven Design and Development (DBCDDD) approach that we are proposing in this paper is made up of the 6 methodological steps summarized in **Figure 1**.

2.2 Sub-universe analysis

You might want to apply in this step Algorithm *A*0 from Mancas [20], such as to obtain for the sub-universe of interest an E-R data model [20], which is made of the following 3 deliverables:

(i) A comprehensive set of E-R diagrams (E-RDs);

(ii) An associated set of restrictions (business rules);

(iii) An informal description of the corresponding sub-universe.

This E-R data model (the only one that business-oriented people may understand) should be obtained with the help of and, finally, negotiated with, and approved by our customers. The E-R GUI of *MatBase* [25] may be used to draw, store, and maintain E-RDs.

During this step, the domain-driven approach [5,7] is very useful as well.

Obviously, not even Artificial Intelligence (AI) might ever fulfill this task, but only, at most, help software architects!

2.3 Translation of the resulting E-R data model into a(n) (E)MDM scheme

This step, detailed in Algorithm *A*1 from Mancas [26], can be partially done automatically, with the help of an intelligent DBMS like *MatBase*, which is translating E-RDs into (E)MDM schemes, but only software architects may formalize restrictions (business rules) as FOPL constraints.

2.4 (E)MDM scheme validation and enhancement

For validation, you might want to apply in this step the Algorithm *A*2 from Mancas [26], to correct any modeling errors done in the first step (e.g., declaring a set as being of the relationship type when,

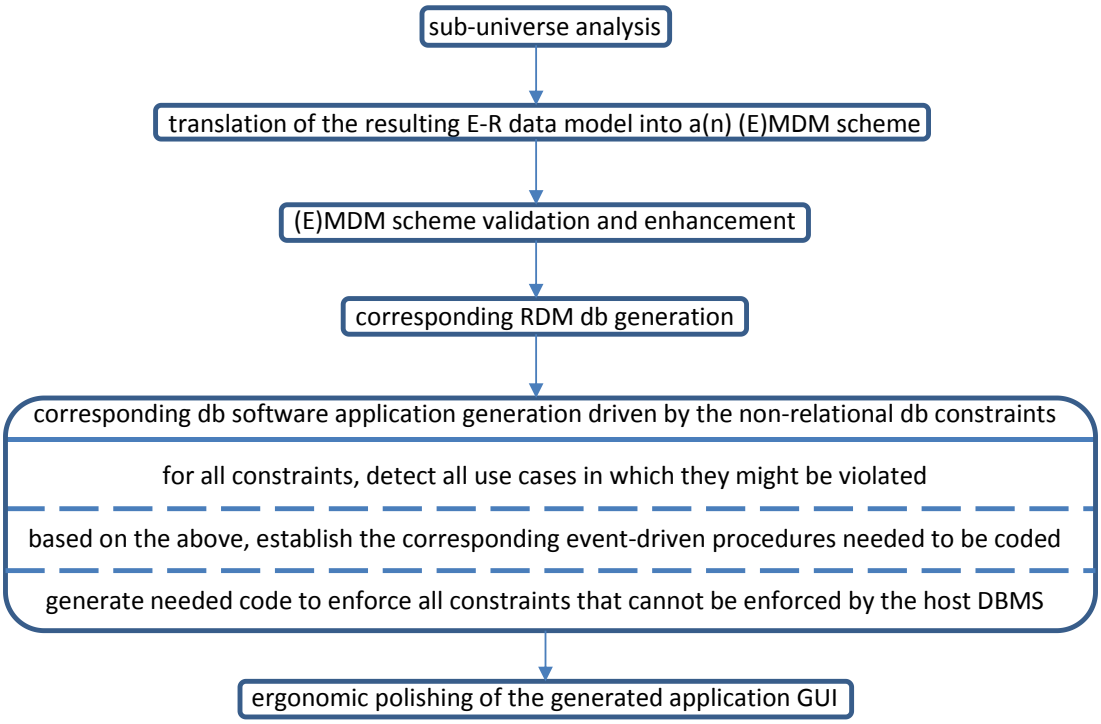


Figure 1. The DBCDDD methodology steps.

in fact, it can only be of the entity one or adding a constraint that does not exist in reality).

In our opinion, data correctness is utopian: for example, very probably, almost nobody knows or will ever know what the height of HM Queen Elizabeth II in Her last days on Earth was (moreover, we bet that most of us do not exactly know our current height, most of the time). Dually, anybody should be sure that the height in centimeters of any world person, any time, is a number between 20 (under which no premature baby managed to survive) and 275 (as the tallest man recorded was 272), so that only the values in this interval are plausible for the *Height* property of persons.

We understand by *plausible data value* (abbreviated as *plausible data*), in any sub-universe of discourse, any value of the data associated with a property (i.e., function codomain) that is satisfying all the business rules of that universe (or, equivalently, is not violating any of them). As such, *data plausibility*, i.e., the fact that a db instance stores only plausible values, is the highest possible form of data quality.

Beware that any existing constraint in the modeled sub-universe which is missing in your (E)MDM (or any other data model) scheme allows for storing unplausible data in your db (e.g., two persons with the same SSN (i.e., US Social Security Number), two countries with a same name, persons living a negative number of days or more than 120 years, etc.); dually, any constraint in your data model scheme that does not exist in the corresponding sub-universe prevents your software application end-users to store valid data in your db (e.g., enforcing for a *MARRIAGES* set/table the constraint *Husband* • *Wife* minimally one-to-one, i.e., declaring this set a relationship, instead of an entity type one, prevents storing data on remarriages, like, for example, the famous ones between Richard Burton and Elizabeth Taylor).

Moreover, enforcing redundant constraints (e.g., that *Mother*: *PEOPLE* $\rightarrow$ *PEOPLE* is not only acyclic, i.e., nobody may be his/her own mother, neither directly, nor indirectly, but also irreflexive and asymmetric, as acyclicity implies both of them), while not tampering with the db instances plausibility, is slowing down your corresponding software application for nothing. Consequently, redundant constraints should never be enforced, but only minimal constraint sets must be [22].

Dually, and much more important, we always need to make sure that our constraint sets are always coherent [22]: For example, if a constraint set contains both the constraint *CurrentCity acyclic*, i.e., no city may be its current one, neither directly, nor indirectly, and the constraint *CurrentCity reflexive*, i.e. the current city of any city is itself, then the corresponding *CurrentCity* column (of a *CITIES* table) would ever remain void (i.e., the corresponding function's image would always be the empty set), because acyclicity implies reflexivity, and any set containing both reflexivity and reflexivity is incoherent. Consequently, we should always remove incoherence from our constraint sets, preferably before coding an incoherent one.

Enhancements involve constraint discovery, as well as guaranteeing the coherence and minimality of the constraint sets. This second sub-step is the crucial one in the process and needs thorough deep thinking. Both (E)MDM and *MatBase* provide assistance algorithms for detecting all missing constraints [26-30], as well as for guaranteeing the coherence and minimality of the constraint sets [22,26].

Obviously, this step too may only be taken by software and db architects: For example, only humans may decide whether, in a given sub-universe, a function is a one-to-one, or a function product is minimally one-to-one or not (e.g. Mormons, some Arabs, some Chinese, etc. may have several simultaneous marriages, orthodox Christians may have at most 4 sequential marriages in a lifetime, catholic ones only one, except for exceptional papal approvals, etc.).

2.5 Corresponding RDM db generation

This step may be fully automated by an intelligent DBMS and *MatBase* is successfully doing it. Alternatively, you might do it manually, by using Algorithm *A*7 from Mancas [26].

This step also produces the sets of the non-rela-

tional constraints and of the relational ones that cannot be enforced by the target DBMS (e.g., MS SQL Server wrongly assumes implicitly that the NULLS set contains only one value, not infinite many ones; as such, it cannot enforce uniqueness constraints on table columns that might contain more than one null value). All constraints from both these sets must be enforced in the next step.

2.6 Corresponding db software application generation

This step is the core DBCDDD one: It takes as input the two above constraint sets that cannot be enforced by the DBMS host and generates the corresponding software application, which must enforce them instead. This step has the 3 sub-steps separated in **Figure 1** by dashed lines.

Especially this step might never be totally entrusted to anything or anybody else than a software and db architect. (E)MDM and *MatBase* are only assisting this process with Algorithm *A*9 from Mancas [26] and are automatically generating corresponding code whenever possible.

2.7 Ergonomic polishing of the generated application GUI

Even when using an intelligent tool like *MatBase*, at the end of the previous step you end up with only a set of MS Windows forms and their classes that are enforcing all the constraints. However, they must be ergonomically architectured in a hierarchy of forms and sub-forms that are called by a menu of the corresponding application.

Moreover, basic ergonomic principles should incite you to replace all context-independent (and, generally, incomprehensible to application's end-users, as they are hard to understand sometimes even by senior db developers) DBMS error messages with context-sensitive ones, to add facilities like pre-programmed queries and reports, navigation shortcuts between related data, to embellish the standard GUI with end-users fancied options, etc.

Obviously, all these may only be accomplished manually, by developers.

3. Results and discussion on applying DBCDDD to a genealogy sub-universe

Mancas [9] considered an extended genogram sub-universe, by adding to the genealogy trees data on countries, cities, monuments, marriages, and reigns of rulers over countries.

The MS SQL Server 2022 Developer edition was chosen as the application db host.

3.1 The sub-universe objects and their main properties

The corresponding E-R data model contains the following 13 object types (with their main properties in parentheses):

1) PERSONS (Name, Sex, Birth and PassedAway Dates and Cities, Mother, Father, Killer, BurialMonument, Family/Dynasty, Title, Nationality, Website, Picture, Notes);

2) DYNASTIES/FAMILIES (Name, Country, Founder, ParentHouse);

3) TITLES (Name);

4) MARRIAGES (Husband, Wife, Marriage, and Divorce Dates);

5) COUNTRIES (Name, Capital city, CurrentCountry, MainNationality);

6) CITIES (Name, Country, CurrentCity);

7) COUNTRIES_CAPITALS (Country, City, EstablishingYear);

8) CITIES_PICTURES (City, Picture, PictDescription);

9) MONUMENTS (Name, Type, City, Website, Notes);

10) MONUMENT_TYPES (Name);

11) MONUMENTS_PICTURES (Monument, Picture, PictDescription);

12) REIGNS (Person, Title, Country, Start and End Dates, Notes);

13) PARAMS (maxLifeYears, minMFertileAge, minFFertileAge, maxMFertileAge, maxFFertileAge, maxSurvivalMDays, maxSurvivalFDays).

The *Sex* property accepts 3 values: 'F' for fe-

males, 'M' for males, and 'N' for anything else (e.g., military occupations, international bodies administrations, etc.).

The corresponding structural E-RD [20] is shown in **Figure 2**.

3.2 The sub-universe constraints

This sub-universe is governed by 210 business rules. Their corresponding constraints are grouped as follows:

(i) 172 relational constraints, out of which:

- 21 domain (range) ones;
- 53 totality (not-null) ones;
- 2 default value ones;
- 12 primary key ones;
- 26 unique ones;
- 28 reference integrity (foreign key) ones;
- 30 tuple (check) ones.

(ii) 38 non-relational constraints.

Out of these 210 constraints, only the following 65 might raise issues (as the domain, totality, except for 2 of them, the ones for pictures, default, primary and foreign keys, as well as most of the tuple/check ones are simple to have them enforced by the MS SQL Server):

- C_1: There may not be two persons of the same dynasty (family) born in the same year and having the same names.
- C_2: No mother gives the same names to two of her children.
- C_3: No father gives the same names to two of his children.
- C_4: No person may live less than 0 days and more than *maxLifeYears* years.
- C_5: Mothers' sex must be 'F'.
- C_6: Wives' sex must be 'F'.
- C_7: Fathers' sex must be 'M'.
- C_8: Husbands' sex must be 'M'.
- C_9: Nobody may be his/her own mother, neither directly, nor indirectly (i.e., no ancestor, other than his/her mother, or descendant of somebody may be that somebody's mother).
- C_{10}: Nobody may be his/her own father, neither directly, nor indirectly (i.e., no ancestor, other than his/her father, or descendant of somebody may be that somebody's father).
- C_{11}: Nobody may be his/her ancestor or descendant.
- C_{12}: No woman may give birth before being *minFertileFAge* or after being *maxFertileFAge* years old, or after her death.
- C_{13}: No man may have a child before being *minMertileFAge* or after being *maxMertileFAge* years old, or more than *maxMSurvivalDays* after his death.
- C_{14}: Nobody may get married before being born or after death.

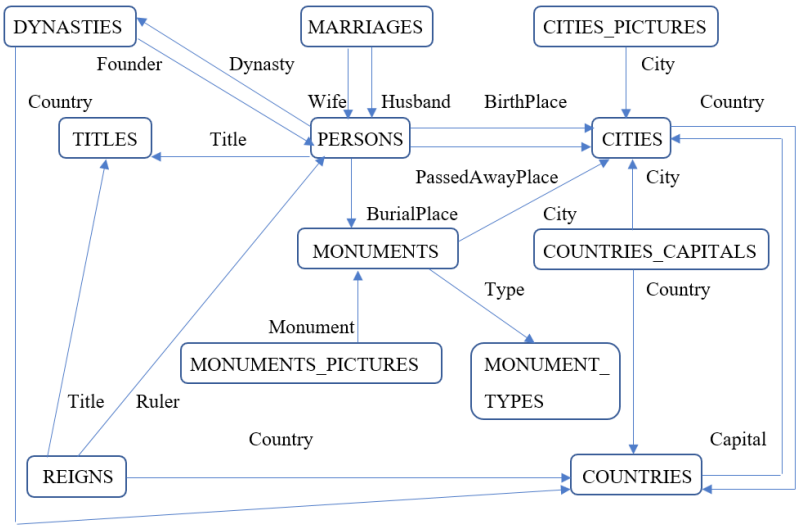


Figure 2. The structural E-RD of the genealogy db from Mancas [9].

- C_{15}: Nobody may divorce before being born or after death.
- C_{16}: Nobody may divorce before getting married.
- C_{17}: Nobody may get married twice on a same date.
- C_{18}: Nobody may get divorced twice on a same date.
- C_{19}: Nobody can get married while still being married.
- C_{20}: For any marriage, both spouses must be simultaneously alive for at least one day.
- C_{21}: No woman may be the wife of one of her ancestors or descendants.
- C_{22}: No man may be the husband of one of his ancestors or descendants.
- C_{23}: Nobody may be killed by somebody who was not alive when the assassination occurred.
- C_{24}: Nobody may belong to a dynasty (family) founded after his/her death.
- C_{25}: The founder of a dynasty (family) must belong to that dynasty or to its parent house.
- C_{26}: Nobody may have found more than one dynasty (family).
- C_{27}: There may not exist two dynasties with the same name.
- C_{28}: Any parent house must be established before any of its child dynasties.
- C_{29}: No dynasty (family) may be its ancestor or descendant, neither directly, nor indirectly.
- C_{30}: It does not make sense to store more than once a title.
- C_{31}: Nobody may reign before birth or after death.
- C_{32}: No country may be simultaneously ruled by two persons, except for spouses and for regencies.
- C_{33}: No reign may end before its start.
- C_{34}: It does not make sense to store more than once the fact that somebody started his/her rule in a country at any given date.
- C_{35}: It does not make sense to store more than once the fact that somebody ended his/her rule in a country at any given date.
- C_{36}: There may not be two countries having the same names.
- C_{37}: No country maybe its current one, neither directly, nor indirectly.
- C_{38}: No former country may be a current one.
- C_{39}: There may not be two cities of the same country having the same names.
- C_{40}: No city may be its current one, neither directly, nor indirectly.
- C_{41}: No former city may be a current one.
- C_{42}: The capital city of any country must either belong to that country, or to the current country of it, or to a former country whose current one is that country.
- C_{43}: No country establishes more than one city as its capital in any given year.
- C_{44}: It does not make sense to store more than once a picture from a city.
- C_{45}: Picture descriptions for the same city must be unique.
- C_{46}: It does not make sense to store more than once a picture of a monument.
- C_{47}: Picture descriptions for the same monument must be unique.
- C_{48}: There may not be two monuments in the same city having the same names.
- C_{49}: The website of a monument may not be shared by another monument.
- C_{50}: It does not make sense to store more than once a monument type.
- C_{51}: Whenever birth month and/or day are known, the birth year must be known too.
- C_{52}: Whenever the death month and/or day are known, the death year must be known too.
- C_{53}: Whenever the reign start month and/or day is known, the reign start year must be known too.
- C_{54}: Whenever the reign end month and/or day are known, the reign end year must be known too.
- C_{55}: Persons of sex 'N' may not have either parents or children, may not marry, and may not belong to dynasties (families).
- C_{56}: There may not be two persons having no

parents, no birth year, but the same name, sex, notes, and dynasty, or no dynasty.

- C_{57}: Nobody may have a brother as his/her father.
- C_{58}: Nobody may have a sister as his/her mother.
- C_{59}: City pictures are mandatory.
- C_{60}: Monument pictures are mandatory.
- C_{61}: There may not be more than one value for any application parameter.
- C_{62}: Parameter values may not be deleted.
- C_{63}: $0 < minMFertileAge < maxMFertileAge < maxLifeYears$
- C_{64}: $0 < minFFertileAge < maxFFertileAge < maxLifeYears$
- C_{65}: *maxSurvivalFDays* and *maxSurvivalMDays* parameter values may not be modified.

Out of these 65 constraints, 28 are relational ones, but only the following 9 out of them may be enforced by the MS SQL Server, namely: C_{27}, C_{30}, C_{36}, C_{39}, C_{43}, C_{48}, C_{50}, C_{63}, and C_{64}.

The 19 remaining ones (13 of type uniqueness, namely C_1, C_2, C_3, C_{17}, C_{18}, C_{26}, C_{34}, C_{35}, C_{44}, C_{45}, C_{46}, C_{47}, and C_{49}, as well as 4 of type tuple/check, namely C_4, C_{16}, C_{33}, C_{55}, and 2 of type totality, namely C_{59} and C_{60}) may not be enforced through the MS SQL Server, because the first 17 ones include at least one table column (which corresponds to a function defined on the set represented by its table, which corresponds in its turn to an object property) that accepts nulls, whereas the last two ones are on columns of type VARBINARY, on which no constraints are allowed. Consequently, all these 19 constraints must be enforced by the software application, just like the 38 non-relational ones.

Unfortunately, in the end, two of these 57 constraints may not be enforced at all, namely C_{44} and C_{46}, as large, good quality pictures (for both cities and monuments, in this case) may not be manipulated in memory either, not even by the Variant type of VBA (although they are linked or embedded as OLEDB objects).

3.3 The use cases that might violate the 55 constraints to be enforced through application code

Please note that, as expected, persons for whom passed away dates are null are considered still alive. Similarly, reigns for which end dates are null are considered still ongoing. Marriages for which divorce dates are nulls are considered still ongoing only while both spouses are alive.

Constraint C_1

(i) Current person's dynasty (family) is replaced by a not-null one;

(ii) Current person's name is modified;

(iii) Current person's birth year is replaced by a not-null one.

Constraint C_2

(i) Current person's mother is replaced by a not-null one;

(ii) Current person's name is modified when his/her mother is known.

Constraint C_3

(i) Current person's father is replaced by a not null one;

(ii) Current person's name is modified when his/her father is known.

Constraint C_4

(i) Current person's birth or/and passed away dates are replaced (for birth by a not null one);

(ii) For persons still alive, simply by the passing time (i.e., not when data is modified).

Constraint C_5

(i) Selecting as the mother of the current person somebody of sex 'M' or 'N';

(ii) Changing the current person's sex to 'M' or 'N' when that person is a mother.

Constraint C_6

(i) Selecting as the wife of current marriage somebody of sex 'M' or 'N';

(ii) Changing the sex of a wife to 'M' or 'N'.

Constraint C_7

(i) Selecting as the father of the current person somebody of sex 'F' or 'N';

(ii) Changing the current person's sex to 'F' or 'N' when that person is a father.

Constraint C_8

(i) Selecting as the husband of current marriage somebody of sex 'F' or 'N';

(ii) Changing the sex of a husband to 'F' or 'N'.

Constraint C_9

Might be violated only when, for the current person, is selected as his/her mother that person or a maternal ancestor or descendant of him/her.

Constraint C_{10}

Might be violated only when, for the current person, is selected as his/her father that person or a paternal ancestor or descendant of him/her.

Constraint C_{11}

(i) Selecting as the father of the current person somebody who is an ancestor or descendant of his/her mother;

(ii) Selecting as the mother of the current person somebody who is an ancestor or descendant of his/her father.

Constraint C_{12}

(i) Selecting as the mother of the current person somebody who does not satisfy this condition;

(ii) Modifying birth and/or death dates of a mother;

(iii) Modifying birth and/or death dates of a child of a known mother.

Constraint C_{13}

(i) Selecting as the father of the current person somebody who does not satisfy this condition;

(ii) Modifying birth and/or death dates of a father;

(iii) Modifying birth and/or death dates of a child of a known father.

Constraint C_{14}

(i) Selecting as a spouse of current marriage somebody who does not satisfy this condition;

(ii) Modifying marriage date;

(iii) Modifying birth and/or death dates of a spouse.

Constraints C_{15} and C_{16}

Let us consider constraint $C_{15'}$: Nobody may divorce before getting married or after death. Together with C_{14}, $C_{15'}$ obviously imply both C_{15} and C_{16}; consequently, we replace them with $C_{15'}$, which might be violated only in the following 3 use cases:

(i) Selecting as a spouse of current marriage somebody who does not satisfy this condition;

(ii) Modifying marriage and/or divorce dates for the current marriage;

(iii) Modifying the death date of a spouse.

Constraint C_{17}

(i) Replacing the marriage date for the current marriage with a not-null one;

(ii) Modifying a spouse of the current marriage.

Constraint C_{18}

(i) Replacing the divorce date for the current marriage with a not-null one;

(ii) Modifying a spouse of the current marriage.

Constraint C_{19}

(i) Selecting as a spouse of current marriage somebody who does not satisfy this condition;

(ii) Modifying marriage and/or divorce dates for the current marriage.

Constraint C_{20}

(i) Selecting as a spouse of current marriage somebody who does not satisfy this condition;

(ii) Modifying marriage and/or divorce dates for the current marriage;

(iii) Modifying birth and/or death dates of a spouse.

Constraint C_{21}

Might be violated only when, for the current marriage, is selected as husband somebody who is

an ancestor or descendant (either maternally or/and paternally) of the corresponding wife.

Constraint C_{22}

Might be violated only when, for the current marriage, is selected as wife somebody who is an ancestor or descendant (either maternally or/and paternally) of the corresponding husband.

Constraint C_{23}

(i) Selecting as a killer of the current person somebody else who does not satisfy this condition;

(ii) Modifying birth and/or death dates of a killer or/and the death date of the current person.

Constraint C_{24}

(i) Selecting as the founder of the current dynasty somebody who does not satisfy this condition;

(ii) Modifying the birth date of a founder of a dynasty;

(iii) Modifying birth and/or death dates of a member of a dynasty;

(iv) Replacing the current person's dynasty with a not-null one.

Constraint C_{25}

(i) Selecting as a founder of the current dynasty a known person;

(ii) Modifying the dynasty of its founder;

(iii) Replacing the current dynasty's parent house when the current dynasty's founder is not null.

Constraint C_{26}

C_{26} is redundant, as implied by C_{25}: Any founder belonging to its dynasty may not belong to another one as well.

Constraint C_{28}

(i) Selecting as founder of the current dynasty a known person;

(ii) Modifying the birth date of the dynasty founder;

(iii) Replacing the current dynasty's parent house with a not-null one.

Constraint C_{29}

Might be violated only when, for the current dynasty, is selected as the parent house either the current dynasty or one of its ancestors or descendants.

Constraint C_{31}

(i) Selecting as ruler of current reign somebody who does not satisfy this condition;

(ii) Modifying birth and/or death dates for a ruler;

(iii) Modifying start and/or end dates of the current reign.

Constraint C_{32}

(i) Selecting as co-ruler of a reign somebody who does not satisfy this condition;

(ii) Modifying birth and/or death dates for a co-ruler;

(iii) Modifying marriage and/or divorce dates for a co-ruler;

(iv) Modifying start and/or end dates of the current reign;

(v) Modifying the country of the current reign;

(vi) Modifying the title of a co-ruler.

Constraint C_{33}

Might be violated only when, for the current reign, start and/or end dates are modified.

Constraint C_{34}

(i) Modifying the start date of the current reign;

(ii) Modifying the country of the current reign;

(iii) Modifying the ruler of the current reign.

Constraint C_{35}

(i) Modifying the end date of the current reign;

(ii) Modifying the country of the current reign;

(iii) Modifying the ruler of the current reign.

Constraint C_{37}

Might be violated only when, for a country, is selected as its current one itself or one of its former ones.

Constraint C_{38}

Might be violated only when, for a country, is selected as its current country or a former one.

Constraint C_{40}

Might be violated only when, for a city, is selected as its current one itself or one of its former ones.

Constraint C_{41}

Might be violated only when, for a city, is selected as its current city or one of its former ones.

Constraint C_{42}

(i) Selecting for the current country in COUNTRIES_CAPITALS a city that does not satisfy this condition;

(ii) Modifying for a country occurring in COUNTRIES_CAPITALS its current one;

(iii) Modifying for a capital occurring in COUNTRIES_CAPITALS its current city;

Constraint C_{45}

(i) Replacing the description of the current city picture with a not-null one;

(ii) Replacing the city of the current city picture with another one.

Constraint C_{47}

(i) Replacing the description of the current monument picture with a not-null one;

(ii) Replacing the monument of the current monument picture with another one.

Constraint C_{49}

Might be violated only when replacing the website URL of a monument with a not null one.

Constraint C_{51}

Might be violated only when modifying the birthday and/or month and/or year of a person.

Constraint C_{52}

Might be violated only when modifying the death day and/or month and/or year of a person.

Constraint C_{53}

Might be violated only when modifying the start day and/or month and/or year of a reign.

Constraint C_{54}

Might be violated only when modifying the end day and/or month and/or year of a reign.

Constraint C_{55}

(i) Selecting a not null dynasty, father, or mother for a person of sex 'N';

(ii) Replacing the sex value of a person with 'N'.

Constraint C_{56}

(i) Attempting to enter corresponding duplicate data for a new person;

(ii) Replacing the mother and/or father and/or birth year of the current person with nulls;

(iii) Replacing name or/and sex or/and notes or/and dynasty of the current person.

Constraint C_{57}

(i) Adding/replacing a brother to the current person;

(ii) Adding/replacing the father of the current person.

Constraint C_{58}

(i) Adding/replacing a sister to the current person;

(ii) Adding/replacing the mother of the current person.

Constraint C_{59}

Might be violated only when adding to the current city a picture description without a picture.

Constraint C_{60}

Might be violated only when adding to the current monument a picture description without a picture.

Constraint C_{61}

Might be violated only when a second line is saved in the PARAMETERS table.

Constraint C_{62}

(i) Replacing a parameter value with a null one;

(ii) Deleting the only line of the PARAMETERS table.

Constraint C_{63}

Might be violated only when modifying the values of at least one of these 3 parameters.

Constraint C_{64}

Might be violated only when modifying the values of at least one of these 3 parameters.

Constraint C_{65}

Might be violated only when modifying the value

of at least one of these 2 parameters.

3.4 Establishing the corresponding event-driven procedures needed to be coded

This sub-step heavily depends on the platform used for coding the application. For example, *MatBase*, which has two versions: -one for students and small dbs and a professional one- that uses VBA and C#, respectively. Mancas [9] opted for VBA, which is both simpler, very robust, and provides an extensive set of data-oriented events and associated event-driven procedures.

It is out of the scope of this paper to enter into details on software application development on any platform, as this would need at least another 20 pages per platform and would not be of any academic, but only of technological interest.

The only general aspect about this sub-step is the fact that there are two possible algorithmic approaches to enforce db constraints in software applications, just like in healthcare, namely:

(i) *preventively* (i.e., providing users to choose from only plausible data in combo-boxes),

(ii) *curatively* (i.e., letting users enter desired data and reject unplausible ones).

The preventive ones are the best and, for example, in VBA they may be coded in the *Form_Current* event-driven procedures, which are automatically called each time the cursor is set on another data line of the current form. For example, in the *DYNASTIES* form, this procedure should dynamically modify the SQL SELECT statements that compute the combo-boxes *Founder* and *ParentHouse* and then re-query them, such as to eliminate from *ParentHouse* the current dynasty and from *Founder* all persons that are not belonging to either the current dynasty or its parent house, as well as those dead before the birth of the current founder (thus preventively enforcing constraints C_{24}, C_{25} , and C_{29}, respectively).

Sometimes, however, this is not possible (not even for all combo-boxes and all types of constraints involving their corresponding columns, hence functions). For example, to enforce constraint C_{49} you can only let the user type any desired text string in the *Website* text-box control of the *MONUMENTS* form's current data line and then reject it within the *Website_BeforeUpdate* VBA event-driven procedure (corresponding to the *Validating* event type of .NET) if that URL is already stored in the db for another monument.

3.5 Comparative analysis

In Mancas [9], state-of-the-art analysis of genealogy software applications available on the market was conducted as well, starting from the No1Reviews.com website post on the top 10 of such applications in 2022 [31]. Only 8 of them have been analyzed (as one is only for Apple hardware and software and the other is a website builder not freely available for evaluation) and only 3 of them provide a rudiment of data quality consideration: For a few unplausible values (e.g. passed away date less than birth one) they warn you and ask a confirmation message to which, unfortunately, you can answer *Yes*, thus saving that unplausible data in their dbs. In all 8 of them we easily manage to save aberrantly unplausible data, like persons living centuries, getting married or/and baptized before birth or after death, mothers of sex 'M', fathers of sex 'F', persons being buried before death, etc.

Unfortunately, this is not an exception: Such software applications abound in all fields, not only in the genealogy one. Some might say that the corresponding software companies lack software and/or db architects or that all fine ones are working only for giants like Microsoft, Google, Apple, Tesla, etc.

We strongly believe, however, that the main reason for this catastrophic reality is that, on one hand, software engineering treats db applications just as the not-db ones and, on the other, it completely lacks consideration of the main asset of any db application, namely its managed data quality. And, as we've explained, data quality may be guaranteed only by plausible data values and data plausibility may be guaranteed only by discovering all business rules governing the considered sub-universes and enforcing all their corresponding constraints.

This is why we consider that our proposed db

constraint-driven design and development methodology described in this paper is a crucial approach to take towards the delivery of high-quality software db applications, not only exhibiting glossy GUIs, but, especially, guaranteeing the highest quality possible of the managed data.

Using the DB Constraint-Driven Design and Development approach, the genogram software application described in Mancas [9] and in the previous section of this paper successfully and elegantly enforced all 208 constraints governing this sub-universe that can be enforced with the currently available technologies. The contrast between this application and the ones considered in reviews [31] as the best ones in this field could not be more spectacular.

4. Conclusions and further work

We introduced a novel database constraint-driven methodology for designing and developing software database applications. We exemplified it with a complex medium-sized software database application for managing genograms. We argued that, using this methodology, this application guarantees the highest possible quality of the data it is managing, whereas most of the similar applications available and considered to be the best ones in this field have almost no concern at all about data quality.

Moreover, although Mancas [9] used this paradigm manually, our previous research and the *MatBase* prototype embedding it provide powerful tools to program while modeling, which is the future of software, as fewer and fewer developers and testers, while more and more architects and designers will soon be needed with the generalization of automatic code generation.

Further work is needed to automate software applications' code generation for the (E)MDM general object constraints [22,26] in *MatBase*.

Author Contributions

Dr. Christian Mancas is the author of the software application DB Constraint-Driven Design and Development approach, which he was teaching for decades to his MSc. students with both the Math. & Computer Science Dept. of the Ovidius University at Constanta, Romania, and the English Stream Computer & Telecomm. Engineering Taught in Foreign Languages Dept. of the Politehnica University at Bucharest, Romania. Dr. Mancas wrote the first 2 sections of this paper, plus its last sentence (the one on further work above). Miss Diana Christina Mancas wrote the rest of it. Associate Professor Cristina Serban is the scientific coordinator of her work, as well as for her entire MSc. Dissertation Thesis [9].

Conflict of Interest

There is no conflict of interest.

Funding

This research received no external funding.

Acknowledgement

We are grateful to Mihaela Virginia Mancas for her thorough revision of the final manuscript, which, hopefully, thus got rid of any typo or syntax errors and is fully intelligible, as well as to the whole JCSR editorial team for their kind and competent assistance.

References

[1] Daylight, E.G., Niklaus, W., Hoare, T., et al., 2012. The dawn of software engineering: From Turing to Dijkstra. Lonely Scholar bvba: Belgium.

[2] Dijkstra, E.W., 1982. Selected writings on computing: A personal perspective. Springer Verlag: NY, Heidelberg, Berlin.

[3] Hoog, R. de, Jong, T. de, Vries, F. de, 1995. Constraint-driven software design: An escape from the waterfall model. 7(3), 48-63. DOI: https://doi.org/10.1111/j.1937-8327.1994.tb00637.x

[4] Hunt, A., Thomas, D., 1999. The pragmatic programmer: From journeyman to master. Addison-Wesley Professional: IL, USA.

[5] Evans, E., 2003. Domain-driven design: Tack-

ling complexity in the heart of software. Addison-Wesley Professional: IL, USA.

[6] Taylor, R.N., Medvidovic, N., Dashofy, E.M., 2010. Software architecture: Foundations, theory, and practice. Wiley: NJ, USA.

[7] Vernon, V., 2016. Domain-driven design distilled. Addison-Wesley: IL, USA.

[8] Ousterhout, J., 2021. A Philosophy of Software Design, 2nd edition. Yaknyam Press: CA, USA.

[9] Mancas, D. C., 2023. Design and development of a DB software application for managing genealogical trees [Master's thesis]. Constanta, Romania: Ovidius University. p. 670.

[10] Lano, K., 2008. Constraint-driven development. Information and Software Technology. 50(5), 406-423.
DOI: https://doi.org/10.1016/j.infsof.2007.04.003

[11] Demuth, A., Lopez-Herrejon, R.E., Egyed, A. (2012). Constraint-Driven Modeling through Transformation. In: Hu, Z., de Lara, J. (editors), Theory and Practice of Model Transformations. ICMT 2012. Lecture Notes in Computer Science. 7307, 248-263.
DOI: https://doi.org/10.1007/978-3-642-30476-7_17

[12] Rebmann, A., Weidlich, M., Aa, H. van der (editors), 2022. GECCO: Constraint-driven abstraction of low-level event logs. 38th IEEE International Conference on Data Engineering; 2022 May 9-12; Kuala Lumpur, Malaysia. USA: IEEE. pp. 150-163.
DOI: https://doi.org/10.1109/ICDE53745.2022.00016

[13] Siddiqui, J.H., 2012. Improving Systematic Constraint-driven Analysis using Incremental and Parallel Techniques [PhD thesis]. USA: University of Texas at Austin. [cited 2023 Feb 14]. Available from: https://repositories.lib.utexas.edu/bitstream/handle/2152/19568/siddiqui_dissertation_201221.pdf?sequence=1&isAllowed=y

[14] Shrotri, A.A., Narodytska, N., Ignatiev, A., et al., 2022. Constraint-driven explanations for black box ML models. Proceedings of the AAAI Conference on Artificial Intelligence. 36(8), 8304-8314.
DOI: https://doi.org/10.1609/aaai.v36i8.20805

[15] Ciortuz, L., 1997. Constraint-Driven Concurrent Parsing Applied to Romanian Transitive VP [Internet]. Proceedings of the International Workshop on Parsing Technologies [cited 2023 Feb 14]. Available from: https://aclanthology.org/1997.iwpt-1.26.pdf

[16] Kumaran, E., 2022. Constraint-driven Agree. Proceeding of Linguist Society America. 7(1), 5282.
DOI: https://doi.org/10.3765/plsa.v7i1.5282

[17] OrCAD, 2023. Integrated Front-To-Back Constraints for Right First Time Designs [Internet]. [cited 2023 Feb 14]. Available from: https://www.orcad.com/tech-solutions/constraint-driven-design

[18] Mancas, C., 2019. *MatBase*—A tool for transparent programming while modeling data at conceptual levels. Computer Science & Information Technology (CSITEC 2019). AIRCC Pub. Corp.: Chennai, India. pp. 15-27.
DOI: https://doi.org/10.5121/csit.2019.91102

[19] Abiteboul, S., Hull, R., Vianu, V., 1995. Foundations of databases. Addison-Wesley: IL, USA.

[20] Mancas, C., 2015. Conceptual data modeling and database design: A completely algorithmic approach. Volume I: The shortest advisable path. Apple Academic Press/CRC Press (Taylor & Francis Group): FL, USA.

[21] Kleppmann, M., 2016. Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. O'Reilly: UK.

[22] Mancas, C., 2018. *MatBase* constraint sets coherence and minimality enforcement algorithms. Advances in databases and information systems. Springer: Switzerland. pp. 263-277.
DOI: https://doi.org/10.1007/978-3-319-98398-1_18

[23] Thalheim, B., 2000. Entity-relationship modeling: Foundations of database technology. Springer Berlin: Heidelberg.

[24] Mancas, C., Dragomir, S., 2004. *Matbase* Datalog Subsystem Metacatalog Conceptual Design [Internet]. Proceedings of the IASTED Conference on Software Engineering and Applications,

November 9-11, 2004, MIT, Cambridge, MA, USA. Acta Press: Canada. pp. 34-41 [cited 2023 Feb 14]. Available from: https://www.actapress.com/PaperInfo.aspx?PaperID=19050&reason=500

[25] Mancas, C., Mancas, S., 2005. *Matbase* Entity-Relationship Diagrams Subsystem Metacatalog Conceptual Design [Internet]. IASTED International Conference on Databases and Applications, Part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria. pp. 83-89. Acta Press: Canada. [cited 2023 Feb 14]. Available from: https://www.actapress.com/PaperInfo.aspx?PaperID=19050&reason=500

[26] Mancas, C., 2023. Conceptual data modeling and database design: A completely algorithmic approach. Volume II: Refinements for an Expert Path. Apple Academic Press/CRC Press (Taylor & Francis Group): FL, USA.

[27] Mancas, C., 2016. Algorithms for key discovery assistance. BIR 2016, lecture notes in business information processing. Springer: Switzerland. pp. 261, 322-338.
DOI: https://doi.org/10.1007/978-3-319-45321-7_23

[28] Mancas, C., 2019. *MatBase* E-RD cycles associated non-relational constraints discovery assistance algorithm. Intelligent computing. Springer: Switzerland. pp. 390-409.
DOI: https://doi.org/10.1007/978-3-030-22871-2_27

[29] Mancas, C., 2019. *MatBase* autofunction non-relational constraints enforcement algorithms. IJCSIT. 11(5), 63-76.
DOI: https://doi.org/10.5121/ijcsit.2019.11505

[30] Mancas, C., 2020. On detecting and enforcing the non-relational constraints associated to dyadic relations in *MatBase.* Journal of Electronic & Information Systems. 2(2), 1-8.
DOI: https://doi.org/ 10.30564/jeisr.v2i2.2090

[31] No1Reviews.com [Internet]. Reviews of the Top 10 Genealogy Software of 2023 [cited 2023 Feb 14]. Available from: https://genealogy-software.no1reviews.com

BILINGUAL
PUBLISHING GROUP
Pioneer of Global Academics Since 1984

Tel:+65 65881289
E-mail: contact@bilpublishing.com
Website:https://journals.bilpubgroup.com

2630-5151
01
9 772630 515230