Scholar Publishing Group
Academic Journal of Environmental Biology
https://doi.org/10.38007/AJEB.2020.010403
ISSN 2789-8407 Vol. 1, Issue 4: 19-28

Environmental Pollution Control by Biotechnology based on Data Mining Algorithm

Zamanian Erfan*

Coventry Univ, W Midlands, England

**corresponding author*

Keywords: Data Mining Algorithm, Biotechnology, Environmental Pollution, Pollution Control

Abstract: With the rapid development of China's social economy, agricultural production efficiency has also been greatly improved. However, with the development of agriculture, the water environment pollution (WEP) continues to increase. As biotechnology comes from the biochemical reaction process of organic organisms in the natural environment, it is a renewable organic technology. Biotechnology occupies a very large advantage in cost and environmental protection. Therefore, this paper takes water pollution control as the research object, and analyzes the role of biotechnology in water pollution control based on Data Mining Algorithm(DMA). This paper briefly introduces the role of biotechnology and data mining in the treatment of WEP; This paper discusses the data analysis of DMA in water pollution control, including chemical analysis and water quality analysis; Finally, improve the data import and analysis of data mining technology, and analyze the implementation in sewage treatment system. The experimental results also verify the feasibility of data mining in biological sewage treatment.

1. Introduction

WEP control has become one of the main measures of national environmental infrastructure construction. Urban water pollution has become the most serious part of water pollution in China, and sewage treatment and recycling have become the focus of attention. The main way of sewage treatment is biological sewage treatment. Due to the long time period of biological treatment and the obvious environmental impact, the amount of chemicals put into the treatment process cannot be accurately estimated, and it is easy to put in excessive chemicals, resulting in excessive drug consumption and secondary pollution of the treated water. If an estimated predicted value can be given as an auxiliary reference according to the measured water quality parameters before putting into use, and the water quality parameters after treatment can be given according to the given drug

dosage, the sewage treatment process can be effectively controlled. Combined with DMA and biotechnology sewage treatment, the key point of this paper is to predict the release of chemicals and the results of water purification.

Many scholars at home and abroad have studied and analyzed the role of DMA in environmental pollution control. Although there are many and relatively mature urban water pollution treatment technologies and schemes in China, for agricultural water pollution with large area, wide dispersion and many factors, the cost is not only very high, but also how to reasonably determine the implementation scheme is a long-term difficult problem [1]. As biotechnology comes from the biochemical reaction process of organic organisms in the natural environment, it is a renewable organic technology. Therefore, compared with physical and chemical technologies, biotechnology has a great advantage in cost and environmental protection. At the same time, with the development of computer technology and information technology, people's way of data management and data storage begins to change, and more and more data information is accumulated and stored. In order to obtain hidden information from huge data and create greater profit value, researchers applied data mining technology to the research on the role of biotechnology in environmental pollution control and obtained good feedback [2].

This paper studies and analyzes the role of data mining technology in the treatment of environmental pollution, and puts forward the data mining technology to help the sewage treatment. Through the study of data mining technology, we understand the application scenarios of data mining and its related patterns. According to the analysis and understanding of existing data, we select the data fitting and regression analysis in the predictive pattern as the application tools. In the process of sewage treatment, due to the use of biotechnology, the role of microorganisms fluctuates greatly. By means of data fitting and regression analysis, the existing laws are analyzed and found according to the existing data, and the law models of various water quality parameters at the time of inflow and outflow are obtained, so that the effluent results can be predicted according to the subsequent new water quality inflow parameters [3-4].

2. The Role of Biotechnology based on DMA in Environmental Pollution Control

2.1. Biotechnology

Biotechnology refers to the technology that people use the principle of material and energy movement in organic organisms to practice and produce the products they need. In a broad sense, biotechnology refers to the technology that people use microorganisms, animals and plants to process raw materials through transformation, purification or modification to provide a product for human beings. The core of the definition of biotechnology molecular biology is the gene engineering centered on DNA recombination technology. In a broad sense, biological technology tends to cell biotechnology, that is, people use isolated cells, microorganisms, animal bodies and plants to obtain a specific product through transformation, purification or modification. The core process is fermentation engineering [5-6].

2.2. Analysis of the Role of Data Mining in the Treatment of WEP

After the introduction of computer technology from industry as a tool for data storage and data measurement, processing data will be left in the computer at every step of industrial production [7]. In the process of sewage treatment, the water quality of the incoming sewage will be measured before each sewage treatment, and these water quality parameters will be measured again after the

treatment is completed. At the same time, some chemicals in the treatment process will be saved. In this paper, the data mining technology is introduced to summarize and discover the change rules of these data, to accurately determine the situation before and after the treatment of water quality, and to study the effect of agent delivery, find the relationship between agent and water quality, and better position the role of agent, which is conducive to optimizing the process of wastewater treatment [8-9].

2.2.1. Treatment of WEP by Biotechnology

The microbial technology is used to decompose the organic matter and nitrogen phosphide in the sewage to eliminate the harmful substances in the sewage, and then the microbial attachments are used to remove the excess substances, eliminate the heavy metals and purify the water quality. At present, in China, biological sewage treatment technology is widely used in urban sewage treatment and is the mainstream treatment method in China [10]. Due to the different microbial treatment methods and biological species of biological sewage technology, there are differences in the treated water quality, and the tendencies of different technologies are also different. Generally speaking, it can be divided into three categories according to the utilization of oxygen by organisms: anaerobic, aerobic and natural [11].

To put it simply, it is to plant aquatic plants on the sewage, and use the plants to absorb organic matter and supply oxygen at the same time. In the water, it is the remaining part of the microbial treatment. The overall treatment process is shown in Fig. 1.

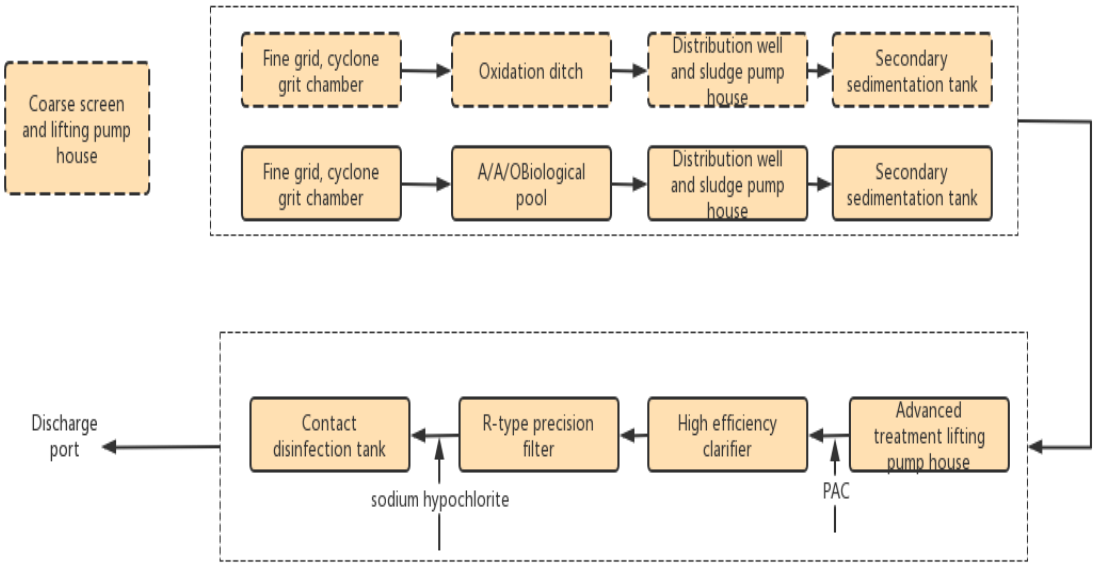

Figure 1. Biological sewage treatment process

2.2.2. Technical Advantages

The microorganisms used can be reused with low cost; The whole process mainly uses electricity for oxygen supply and mixing, which is environmentally friendly and low energy consumption; The excess sludge can be dehydrated and used as compost or industrial use [12-13].

2.3. Analysis of WEP Control based on DMA

2.3.1. Sewage Data Source

The data of sewage treatment mainly include the water quality parameters of sewage, the water quality parameters of purified effluent, and the amount of chemicals added in the treatment process [14]. In fact, the data stored in the database can be divided into three parts. One part is the water quality parameters of the inlet and outlet water, the amount of sedimentation agent input and the parameters of the treatment of sedimentation sludge.

Sludge amount: microbial activated sludge will be precipitated after purification. Since suspended solids in water will be added, the rest will be dehydrated except for partial recycling; Dehydrating agent: the water content of the precipitated sludge is about 99%, which needs dehydration treatment; PAC: sediment of suspended solids in sewage; SS: particulate suspended matter, mainly measuring suspended matter in water; BOD: oxygen consumption of biological decomposition organic matter, calculated in 5 days; Ammonia nitrogen: nitrogen hydride content in sewage [15-16].

2.3.2. Data Mining Mode Selection

The biological sewage treatment data is basically as shown above. According to the analysis of demand and existing data, it is necessary to realize the results of simulating and predicting the effluent quality according to the inflow water quality and the dosage of chemicals used, as well as the required dosage of chemicals according to the inflow water quality and the effluent water quality. The basic model is in the form of $y = f(x)$, and the results are calculated according to the input sample values to be predicted in the future to obtain the prediction results. In the selection of data mining patterns, all descriptive patterns can be excluded first. While the prediction mode is mainly divided into classification mode and regression mode. The classification mode mainly classifies and summarizes the existing data according to the training results of the data set, forms a divided interval, judges the interval for the new sample data, and then gives the prediction result according to the result set corresponding to the interval [17-18]. The regression model is to train the calculation model through data fitting according to the existing data, and calculate the new prediction results according to the new sample data input model.

3. Data Analysis of DMA in Water Pollution Control

3.1. Pharmaceutical Analysis

The agents used are mainly dehydrating agent and PAC. PAC is the precipitant used when the suspended particles are removed by sedimentation in sewage treatment, while dehydrating agent is the agent added when the obtained sludge is dehydrated after sedimentation. Due to the different agents, the independent variable selection for the two dependent variable data is different during data analysis and selection.

The function of the dehydrating agent is to reduce the water content in the sludge. Therefore, when calculating the dehydrating agent model, the independent variable is mainly the amount of sludge left after the sewage sedimentation and the amount of water content reduction. Here, it should be noted that the independent variable is not the state quantity before and after the water content, but the fluctuation quantity. Because the effect of the agent is mainly reflected in the

treatment process and belongs to the process parameter, the independent variable for the calculation of PAC is the inflow of sewage first.

This part of the problem is considered in the design, so the flow of pharmaceutical analysis is roughly as shown in Fig. 2.

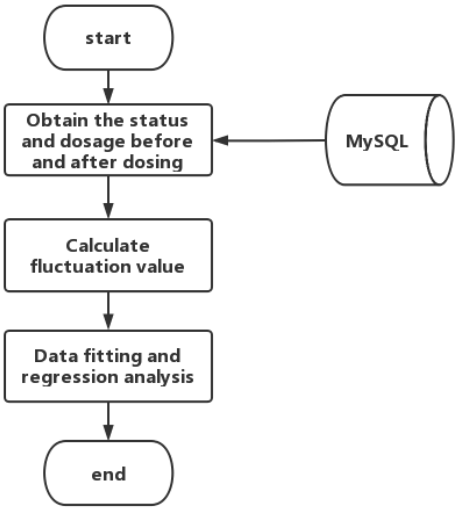

Figure 2. Pharmaceutical analysis flow chart

In the system design, according to the above process, the prediction model of the agent can be obtained for prediction and correlation analysis. In the process of data fitting, this paper considers that the concentration of the agent is also a part that needs to be considered, so the solvent amount of the agent will be selected as a parameter.

3.2. Water Quality Analysis

The water quality analysis is designed for the biological sewage treatment process, which is greatly affected by environmental factors and cannot predict and estimate the water quality parameters of the effluent according to the water quality parameters of the influent sewage and the addition of chemicals. At present, the experiment is mainly aimed at predicting the parameters of SS, BOD, COD, pH and ammonia nitrogen in the sewage water quality. The reason for selecting these parameters is that the updating period of these parameters is fixed, the data volume is large, and the sample representation is strong. In the process of water quality analysis, because different parameters are affected by different factors, they need to be distinguished in the process of treatment, such as SS, COD and BOD. These three parameters are affected by microorganisms and PAC agents in the process of treatment, while the water content of sludge is affected by dehydrating agents, and the independent variable of agent is more than other parts in the analysis. However, the pH value and the acting agent are unknown, and sodium hypochlorite is temporarily used as a substitute. At present, the treatment of ammonia nitrogen is only in the form of microbial decomposition.

In the actual system design, since the data of the agent is the amount of use, and the water quality parameters are basically concentration parameters, they are not a level in the data dimension and need to be processed to convert the agent into concentration. The rest is basically similar to the

analysis of the agent. The flow chart is shown in Fig. 3.

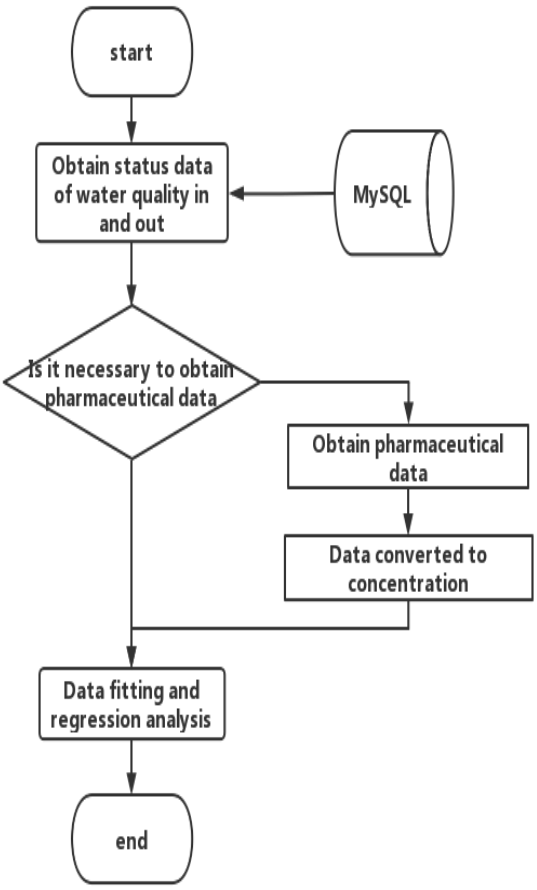

Figure 3. Water quality analysis flow chart

What needs to be paid attention to is the solvent of the agent. The dehydration agent is used for sludge dehydration, so the solvent of the agent is untreated sludge, and the other solvent is sewage.

4. Implementation of DMA in Sewage Treatment System

The water environment treatment system needs to be divided into two parts: data import and data analysis. In terms of data analysis, the drugs and water quality are distinguished and treated separately. In the design of the computing environment, we need to build a matlab computing environment, use data mining, and need a script for data import. Therefore, we choose Python as the experimental language and MySQL as the database for data storage.

4.1. Data Import

In the design of this topic, data import not only serves as the function of data carrier conversion, but also involves simple data filtering function. Due to the problems of missing data and incorrect record format, simple data filtering is performed during import.

Regression mode model evaluation method: since the regression model uses the least square method to fit the possible model functions and selects the closest function as the fitting function model, a quantitative evaluation standard is required as the model evaluation reference. The evaluation coefficient in Python linear fitting is used as the reference in this topic. In the display of the experimental results of this topic, the possibility of the experiment will be displayed, and the model with the highest evaluation coefficient will be selected as the prediction model. The evaluation coefficient mainly calculates the comparison between the predicted results of test samples and the average value of the actual results of test samples, which is similar to the correlation coefficient in the least square method. The calculation formula is shown in formula (1).

$$s^2(b,\hat{b}) = 1 - \frac{\sum_{i=1}^{n}(b_i - \hat{b}_i)^2}{\sum_{i=1}^{n}(b_i - \bar{b}_i)^2} \quad (1)$$

Where $\bar{b}_i$ represents the average value of the calculation results, that is $\sum_{i=1}^{n} b_i$, the average value of, and the number of samples is n, which $\hat{b}_i$represents the prediction result. In addition, during the experiment, two eigenvalues of the function model are mainly recorded, as follows: regression coefficient: the influence degree of the independent variable on the stress variable, that is, the coefficient of the independent variable polynomial in the function. In the multi polynomial function model, this part will be a matrix. Intercept: the intercept between the curve fitted by the model and the coordinate axis.

4.2. Realization of Pharmaceutical Analysis

The pharmaceutical analysis part is mainly aimed at the analysis and prediction of dehydrating drugs and PAC. In the process of pharmaceutical analysis, since the pharmaceutical is a process quantity data, but the corresponding data are all state quantities, the state quantity data will be processed before data processing, and then the data will be fitted to obtain a model. The correlation coefficient is analyzed by regression analysis, and finally the dosage of the pharmaceutical is predicted by the model.

Analysis of dehydrating agent: the dehydrating agent is aimed at the sedimentation part of sewage treatment, and dehydrates the sludge deposited at the bottom of the water. Therefore

The relevant data is the change amount of water in sludge before and after. Since this part is not model processing, data preprocessing is required. The basic formula is:

$$water = mud * \frac{1 - now}{1 - pre} \quad (2)$$

Water represents the water wave momentum, mud represents the amount of sludge, now represents the current water content, and pre represents the previous water content. There is also the amount of sludge. Since the concentration of dehydrating agent will affect the effect, and the sludge will exist as a solvent during the treatment process, the amount of sludge needs to be added. However, it is before the treatment, and the last is the amount of agent. Here, the total amount of agent is used. However, there is a question in the experiment. If the dependent variable here is whether the unit concentration of agent is better, it is temporarily retained due to incomplete data.

In terms of model calculation, it is necessary to manually select a suitable model. In this part, the correlation coefficient of each model is calculated according to the regression analysis. Since the problem of model selection is mainly the degree of polynomial, which belongs to the input

parameter in the code, the test results are required for comparison.

Table 1. Fitting results of dehydrating agent

Polynomial degree	Regression coefficient	Intercept	Evaluation coefficient
1	[[0.52746685,-0.53041068]]	[7.42377634]	0.696461450116
2	[[-3.44729388,3.63202932, 0.00531644, -0.00988678, 0.0045185]]	[33.77130014]	-5.85266000037

From table 1 above, it can be found that the higher the degree, the greater the negative value of the evaluation coefficient, and the greater the proof deviation. Therefore, I chose the polynomial degree of 1.

4.3. Realization of Water Quality Analysis

In this study, because not every water quality index is recorded every day in the sewage treatment process, such as the measurement cycle of E. coli is more than 7 days, or even once a month, which is unreasonable as the data source for data fitting and regression analysis, several parameters with the update cycle of one day are selected: pH, SS, BOD, COD and ammonia nitrogen; In addition, the prediction of sludge water content is also a state prediction, so it is also classified into water quality analysis.

Water content analysis of sludge: the water content analysis of sludge is mainly aimed at the water content of sludge after dehydration. Since the existing data is the water content after treatment, and the water content before treatment cannot be obtained, the default value of this paper is 99%. Since the recording form of water content is proportion, i.e. concentration value, in the actual calculation process, another independent variable - dehydration agent amount is converted into concentration in consideration of reducing the calculation complexity. The calculation formula is as follows:

$$K = med/(mud*(1-now)/(1-pre)) \quad (3)$$

K represents the concentration of dehydrating agent, Med represents the dehydrating agent, mud represents the amount of sludge, now represents the water content after dehydration, and pre represents the water content before dehydration. The results are shown in Table 2.

According to the analysis results in Table 2, when the number of times is 1, one of the regression coefficients is 0. This is because the water content before treatment in the independent variable is 99%, which is a fixed value. Therefore, it is considered that this variable can be attributed to the constant part, that is, the intercept part, during the fitting, so it is hidden and eliminated during the calculation. The evaluation coefficient of the table data shows that both the number of times is 1 and the number of times is 2 are negative. In this part of data fitting processing, data fitting and regression analysis are also affected by the data itself. The fitting model obtained is not as accurate as the average value of historical data.

Table 2. Fitting results of sludge water content

Polynomial degree	Regression coefficient	Intercept	Evaluation coefficient
1	[[0.0, -0.37621007]]	[79.31580389]	-0.0567114275625
2	[[8.87145401e-13, -3.62138555e-04, -1.30137958e-43, -3.58517170e-02, 2.06373823e+01]]	[79.38770554]	-0.0928108570607

5. Conclusion

In this paper, the role of biotechnology in environmental pollution treatment based on DMA is studied and analyzed. The experimental results also verify the feasibility of the application of data mining in biological sewage treatment. In the design of this paper, the sewage treatment data is processed by data fitting and regression analysis, In the verification of data results, it is found that the current scheme has some shortcomings: the data source is the data of sewage treatment all year round. However, since the microbial treatment is very affected by the temperature, and at the same time, human misoperation will have different effects. However, this part of data has not been distinguished or excluded. Although the accuracy of prediction is low due to the lack of data and the lack of close data correlation at present, it is undeniable that the role of this part exists. Therefore, the defects of data and system can be solved in the subsequent research, which can be applied to the production practice of sewage treatment to improve high efficiency.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Sundaravadivel K. A Review On Bioremediation Of Azodyes Using Microbial Consortium From Different Sources. Asian Journal of Microbiology, Biotechnology and Environmental Sciences, 2020, 22(4):614-630.

[2] Anass K, Reda B M, Imad K, et al. Effect of Mycorrhization on Growth and Enzymes involved in Carbon/Nitrogen interaction in Sorghum Plants. Research Journal of Biotechnology, 2020, 16(1):121-126.

[3] Raju M. Performance Assessment Of Soil Biotechnology Treatment Process - A Case Study. Indian Journal of Environmental Protection, 2019, 39(11):1069-1072.

[4] Singh N, Das M K, Gautam R, et al. Assessment of intermittent exposure of zinc oxide

nanoparticle (ZNP)–mediated toxicity and biochemical alterations in the splenocytes of male Wistar rat. Environmental Science and Pollution Research, 2019, 26(32):33642-33653. https://doi.org/10.1007/s11356-019-06225-4

[5] Kolesnikov S I, Kazeev K S, Akimenko Y V. Development of regional standards for pollutants in the soil using biological parameters. Environmental Monitoring and Assessment, 2019, 191(9):1-10. https://doi.org/10.1007/s10661-019-7718-3

[6] Kumar D, Malik D S, Kumar N, et al. Spatial changes in water and heavy metal contamination in water and sediment of river Ganga in the river belt Haridwar to Kanpur. Environmental Geochemistry and Health, 2020, 42(7):2059-2079. https://doi.org/10.1007/s10653-019-00471-8

[7] Saharan B, Sharma D, Ranga P. Bioremediation Of Azo Dye And Textile Effluents Using Pseudomonas Putida Mtcc 2445. Asian Journal of Microbiology, Biotechnology and Environmental Sciences, 2020, 22(2):88-94.

[8] Erhenhi A H, Lemy E E, Vwioko D E, et al. The Roles of Mercury Nitrate in Soil: Effects and Impacts on the Growth of Okra. International Journal of Applied Environmental Sciences, 2019, 14(3):249-258.

[9] Reddy A C, Naresh P, Lakshmana R. Genetics and molecular markers for resistance to major soil borne pathogens in chilli (Capsicum annuum L.). Research Journal of Biotechnology, 2019, 14(1):101-105.

[10] Singh J. Effective technologies and lifestyle changes to reuse and control waste generation from natural and anthropogenic activities for sustainable future. International Research Journal of Environmental Sciences, 2020, 9(2):51-61.

[11] Pattanayak S, Das S, Navyasri K. Bioindicator Emerged as a Potential Environmental Marker. International Journal of Agriculture Environment and Biotechnology, 2020, 13(3):339-344.

[12] Ez-Zriouli R, Yacoubi H E, Benziane Z, et al. Chemical Composition Of Essential Oil Of Eucalyptus Camaldulensis Collected From Forest Moroccan And Determination Their Antifungal Activity On Two Phytopathogenic Fungi. Plant Cell Biotechnology and Molecular Biology, 2019, 20(17-18):770-777.

[13] Kadiri M O, Unusiotame-Owolagba T E. Modelling Toxin-Producing Algae in the Coastal Waters of Nigeria. Journal of water resource and protection, 2020, 12(1):74-92. https://doi.org/10.4236/jwarp.2020.121005

[14] Kushwaha R, Kumar J, Kumar P, et al. Recycling Of Chicken Feather Protein Into Compost By Chrysosporium indicum JK14 And Their Effect On The Growth Promotion Of Zea mays. Plant Cell Biotechnology and Molecular Biology, 2020, 21(37&38):75-80.

[15] Nafissa S, Mohammed B, Ugya A Y, et al. Environmental risk assessment of pesticide use in Algerian agriculture. Journal of Applied Biology & Biotechnology, 2020, 8(5):36-47. https://doi.org/10.7324/JABB.2020.80505

[16] Mindubaev A Z, Fedosimova S V, Grigoryeva T V, et al. Effects of white phosphorus on the cellular morphology and protein profile of Aspergillus niger. Proceedings of universities Applied chemistry and biotechnology, 2020, 11(1):69-79.

[17] Bhat K M, Sharma A, Rao N N, et al. Carrageenan-based edible biodegradable food packaging: A review. Journal of Food Science and Nutrition, 2020, 5(4):69-75.

[18] Tawate S, Gupta R, Jain K. Development of a Technology Commercialization Model for Indian Biotechnology Firms. IEEE Transactions on Engineering Management, 2019, PP(99):1-13.