ANNOTATION
of the dissertation thesis for a degree of Doctor of Philosophy (Ph.D) in specialty "6D060200 – Computer science"

NURZHANOV CHINGIZ ASKAROVICH

Designing an information system for forecasting and decision-making in the remediation of soil containing toxic elements

Research novelty. The advent of computer technology and information technology has presented vast opportunities for studying natural processes. By employing mathematical modelling and cutting-edge computer technology, new methods, models, algorithms, and solutions are being developed to address global environmental challenges concerning human-nature interactions. Currently, machine learning (ML) plays a crucial role in various scientific fields, including information science, ecology, and agriculture, utilizing statistical methods like regression and classification algorithms. However, the widespread use of chemicals in agriculture has led to pollution and reduced crop yields. To enhance agricultural profitability, environmentally friendly practices are now essential. Digitalization in agriculture, enabling real-time tracking of crop yields and production of eco-friendly products, holds a prominent position in the industry. To ensure environmental safety, countries are developing environmental monitoring information systems, gathering Big Data on the state of the environment. The effectiveness of these monitoring systems largely depends on the use of information technologies, typically involving database management systems (DB) for continuous data collection, processing, and storage.

In Kazakhstan, there is currently no centralized computerized database for monitoring the pollution levels in anthropogenically disturbed ecosystems and their waste. Establishing a unified database would empower the government to strategically address environmental concerns, such as the construction of advanced facilities for industrial waste disposal. Moreover, it would enable an accurate assessment of man-made pollution, potential environmental risks, and their impact on public health.

The interest in mathematical modelling for the reclamation of polluted ecosystems, particularly with toxic elements (TEs), is growing each year due to the escalating scale of environmental pollution caused by human activities. Efforts are underway to develop algorithms that can rapidly neutralize pollution sources and create mathematical models to find optimal solutions, providing a comprehensive understanding of soil pollution processes, predicting their consequences on flora, fauna, and the environment, and determining the most effective reclamation strategies. The primary focus of system analysis, coupled with mathematical modelling, is to simulate the restoration process of contaminated sites, with a particular emphasis on hydrocarbons. However, there is currently a lack of theory and models that describe the behavior of other key environmental pollutants in soil and subsoil layers, including their migration within the "soil-plant" system. Developing such models could serve as a foundation for creating technologies to remediate areas contaminated with TEs.

To address this gap, the development of an information system for restoring areas contaminated with xenobiotics, the establishment of a centralized DB on contaminated territories and a DB on plants capable of restoration will play a pivotal role. This integrated approach will enable significant progress in tackling essential environmental projects and tasks related to safeguarding the country's environment and enhancing the well-being of the population living in ecologically hazardous regions.

The purpose of the research is to create an intelligent information system that can effectively process data related to soils contaminated with toxic elements to provide accurate predictions and data-driven decisions for the remediation of contaminated soils in the Republic of Kazakhstan.

Research objectives

1. Modeling the productivity of a bioenergy crop on soils contaminated with toxic elements.

2. Develop a database containing information on woody and herbaceous plant species capable of soil remediation. Creating a database on soils contaminated with toxic elements, taking into account the geographical location of the territory, the amount of obsolete pesticides, and their concentration in the soil.

3. Develop an information system for forecasting and decision-making in cleaning the soil from toxic elements.

4. Using machine-learning to predict of plant productivity that absorb toxic elements from the soil.

Research methods: machine learning methods, multi-row heuristic self-organization method for constructing regression equations, modelling methods, methods During the development of an integrated approach for creating the information system, the research utilized principles from the theory of information systems design, database design methods, and a process-oriented approach. of regression and dispersion multivariate analysis. The study object was data on soils and plants contaminated with toxic elements, as well as climatic conditions (in the example of the Almaty region). The subject of the study was the climate data of AlmatyWeatherDataSet.csv and plant productivity data for 2015–2022 while growing on toxic elements-contaminated soils.

Main research results

1. A model of plant biomass considering environmental factors was adapted using the "Multi-row heuristic method of self-organization." Three factors have the greatest impact on this process: soil moisture evaporation, photosynthetic active radiation and precipitation.

1.1 The Miscancalc model based on Miscanmod was improved to predict plant biomass in contaminated soil taking into account climate data by calculating the difference coefficient between contaminated and clean soil.

2. Two organized data repositories were created: on woody and herbaceous plant species that contribute to soil restoration; the amount of obsolete pesticides and their concentration in the soil.

3. An information system was developed for forecasting and decision-making in soil remediation from TEs. The duration of the cleaning period spent for each toxic element was determined.

4. Intelligent machine learning methods were applied and the best one was determined to predict soil pollution concentrations. According to the results of the study, XGB Regressor has the lowest metrics: R2 = 0.998; MSE = 421.19; MAE = 15.03; MAPE = 0.065.

4.1 Ensemble machine learning methods were explored to predict plant productivity from climate data. The analysis process was carried out through the JupyterLab instrumental software in the Anaconda environment.

4.2 An assessment of regression models for productivity depending on climatic conditions was carried out. The SHAP model identified 13 informative features responsible for productivity, with significant influence from features datetime_1 (months) and datetime_2 (days).

4.3 The duration of the soil-cleaning period for 1 hectare of soil from a depth of 0-20 cm using the Miscantus plant was established for different elements. For particularly toxic elements: lead – 12 years, zinc – 7 years.

Novelty and importance justification of the results obtained:

The scientific novelty of the research lies in the following conclusion:

An improved Miscancalc model based on Miscanmod was proposed for predicting plant biomass on the TE- contaminated soil taking into account climate data.

The method of multi-row self-organization was adapted; its predictive properties are superior in accuracy to regression models, providing automatic selection of informative input variables and selection of the structure of a regression model of optimal complexity.

The soil remediation model based on machine learning methods were expanded using the integrated approach of the XGBoost library, which has high performance and resistance to overfitting.

The significance of the obtained results can be summarized as follows:

1. The algorithms for predicting plant biomass on TEs-contaminated soil, with consideration of climatic data, exhibit high mathematical accuracy.

2. The proposed multi-row self-organization algorithm outperforms traditional regression models in predictive capabilities and offers an automatic selection of informative input variables and optimal complexity for the regression model's structure.

The theoretical and practical significance of the research encompasses the fundamental application of information technology and mathematical modelling in environmental management. The practical value is evident in the swift information dissemination to government agencies responsible for managing land resources and in formulating effective reclamation measures. Additionally, research findings can be practically employed by experimental farms adopting information technologies and organizations engaged in agroecological monitoring.

Compliance with the directions of science development or government programs (projects). The dissertation work aligns with the program and grants No.

AP19678926 “Development of an intelligent system for researching and solving environmental problems of soil and air pollution using data science methods” (2023-2025) of the Ministry of Science and Higher Education of the Republic of Kazakhstan, conducted at the Institute of Information and Computing Technology of the CS MSHE RK. All the results of the doctoral thesis submitted for defense were completed and collected by author. The doctoral candidate personally conducted all aspects of the research, including data collection, analysis, model development, and creation of information systems.

The main results achieved include the creation of the MiscanCalc software, an modification of the MiscanMod model, which predicts plant yield on TEs-contaminated and uncontaminated soil based on climatic conditions. Additionally, 13 machine-learning regression models for crop yield prediction depending on climatic conditions were evaluated. Two databases were created, one for territories contaminated with TEs and another for plants of the Kazakh flora capable of restoring TEs-contaminated soils. Furthermore, an information system for forecasting and decision-making during soil remediation from TEs was developed.

The development of the multi-row self-organization algorithm, analyzing the relationship between plant biomass dynamics and climatic environmental conditions, was conducted under the guidance of a Doctor of Physical and Mathematical Sciences, Professor T. Zh. Mazakov.

The main provisions and research results were reported and discussed at various scientific seminars Department of Computer Science, Faculty of Information Technology of the Al-Farabi Kazakh National University; academic council of the Institute of Information and Computing Technology CS MSHE RK; international conferences: International Scientific Conference IIVT "Modern problems of informatics and computing technologies" June 28-29, 2016; II International Conference on Modern Problems of Computer Science and Computer Technology, National Academy of Sciences of the CS MSHE RK, September 27-30, 2017; 15th International Phytotechnology conference, October 1-5, 2018, University of Novi Sad, Serbia; International scientific conference in the field of information technology, dedicated to the 75th anniversary of Professor U.A. Tukeyev, October 8 2021 and other international conferences.

PhD student contribution to the preparation of each publication.

Published articles and scientific papers describe research results on the topic of the doctoral thesis. Throughout the scientific work, a total of 17 research papers were written: 4 articles published in journals recommended by the Committee for Control in the Sphere of Education and Science of the MSHE of the RK; 8 papers presented in the proceedings of international conferences; 5 articles published in journals included in the Thomson Reuters and Scopus databases.

The structure and scope of the dissertation. The dissertation comprises various components, including designation, abbreviation, introduction, five chapters, a conclusion, a list of references, and an appendix. It spans 191 pages and incorporates 62 figures, 18 tables, and 3 appendices. The reference list encompasses 268 titles.

In the introduction, the dissertation justifies its relevance and outlines the purpose of the research, along with defining the object and subject of study. It also highlights the scientific novelty and practical significance of the work and describes the research findings. Additionally, information is provided regarding the validation and publication of the study results.

Chapter 1 presents a literature review focusing on the information system for environmental monitoring, including data storage technologies and modelling methods applied in ecology, agriculture, and environmental biotechnology.

Chapter 2 delves into three approaches to assess plant productivity, specifically using bioenergy, agricultural species Miscanthus, grown on TEs-contaminated soil depending on climatic conditions (in the example of the Almaty region). These approaches encompass (1) the utilization of 13 regression models of machine learning; (2) the development of the MiscanCalc application derived from the MiscanMod model modification; (3) the creation of a mathematical model called "Multi-row heuristic self-organization method for constructing regression equations".

Chapter 3 details an integrated approach for creating an information system utilizing databases of territories contaminated with toxic elements and plants capable of restoring technogenic landscapes. The chapter includes functional requirements, program creation stages, and system construction. The information system was designed to generate reports on waste generation, disposal, storage, and soil reclamation.

Chapter 4 presents data concerning the development of a mathematical model describing the accumulation of heavy metals in plants' vegetative organs and their migration within the "soil-plant" system, contingent on soil type and environmental conditions.

Chapter 5 provides the data on the creation of an information system for predicting the soil remediation process for 1 ha from a depth of 0–20 cm, using plants suitable for TEs removal. The system's concept is based on an integrated approach involving data collection, transmission, accumulation, and processing of measurement data, database information, plant productivity models, soil phytotoxicity models, and models of TEs absorption by plants and their content in the soil.

The conclusion formulates the main results achieved in the dissertation research.