DOI 10.52209/1609-1825_2025_3_134 UDC 004.93'1

Analysis of Ust-Kamenogorsk Soil Radioactive Contamination Data Using Machine Learning Algorithms

[1]**UVALIYEVA Indira,** *PhD, Professor, iuvalieva@mail.ru,*
[1]***IDRISHEVA Zhanat,** *PhD, Senior Researcher, zhanat.idr@mail.ru,*
[1]**BURUNOVA Almira,** *Cand. of Med. Sc., Associate Professor, abukunova@edu.ektu.kz,*
[1]**SHAYAKHMETKYZY Karina,** *Master's Student, karinasaahmetova88@gmail.com,*
[2]**SKWAREK Ewa,** *dr.hab. PhD, Professor, ewa.skwarek@mail.umcs.pl,*
[1]NPJSC «D. Serikbayev East Kazakhstan Technical University», D. Serikbayev Street, 19, Oskemen, Kazakhstan,
[2]Maria Curie-Sklodowska University, M. Curie-Sklodowska Square, 3, Lublin, Poland,
**corresponding author.*

Abstract. *Radioactive contamination of soil is one of the most important environmental problems affecting natural and anthropogenic ecosystems. The study of radiation background and radionuclide content in soil is necessary to assess their impact on the environment and human health and to develop mitigation strategies. Within the framework of the article the results of analysis of data of radioactive contamination of soils of Ust-Kamenogorsk city with the help of machine learning algorithms are obtained. The aim of the study is to analyze alpha, beta, Ra226 and Th232 indicators of the database of radioactive contamination of the soil of Ust-Kamenogorsk city with the help of statistical methods and machine learning algorithms. Models of comparison of experimental and calculated data of laboratory tests of samples taken for chemical, radionuclide and mineralogical composition were created, general correspondence of observations was studied.*

Keywords: *radioactive contamination, machine learning algorithms, statistical analysis; correlation matrix, alpha activity, beta activity.*

Introduction

The article describes the analytical processing of land in the campus area around the building of the East Kazakhstan Technical University named after D. Serikbayev based on the following data:

- Gamma-ray shooting on the ground at a scale of 1:2000;
- sampling of shurfs up to 1 meter deep at specified points of individual sections of the contaminated area;
- laboratory examination of samples selected for chemical, radionuclide and mineralogical composition.

Radioactive pollution began to take shape in the 40-50s of the last century as a tailings depot for waste from processing ores of rare earth and radioactive metals. Currently, it belongs to the historical pollution of the residential area of the city with radioactive materials. The radionuclide composition of contaminated areas contains radionuclides of the uranium-238 and Thorium-232 series in quantities significantly exceeding the average content of them in this area and in Ust-Kamenogorsk as a whole [1].

The aim of the project is to analyze the indicators of alpha, beta, Ra226 and Th232 databases of radioactive soil pollution in Ust - Kamenogorsk using statistical methods and machine learning algorithms.

The purpose of the subject area of the analysis of radioactive contamination of soils of the city of Ust-Kamenogorsk is to determine the scale, nature and sources of pollution in order to ensure environmental safety and develop effective measures for the rehabilitation of polluted territories [2]. To do this, it is necessary to create a system for collecting, recording and analyzing data, which includes the results of laboratory tests, field measurements and calculations [3].

In addition, an important aspect is to take into account the spatial distribution of radionuclides in the soil and their ability to migrate in underground horizons. This makes it possible

to identify the main areas of pollution, predict the further spread of radionuclides and assess potential threats to population health and ecosystems [4].

It is also necessary to develop a system for analyzing and interpreting the data obtained. It should provide the ability to assess the level of radiation hazard, create reports and recommendations to minimize the consequences of pollution. Such a system contributes to the effective adoption of managerial decisions aimed at protecting the environment and the population.

Research methods

To simplify the work, a database structure was created with the corresponding tables and relationships between them. Each table covers individual aspects of the study: gamma-ray imaging, laboratory data, quantitative interpretation and assessment of qualitative characteristics. Particular attention is paid to the Coordination of data on Spurs, depths and types of measurements.

As part of the analysis, calculations were made on the main indicators of radiation activity, including the values of Alpha and beta activity, as well as the concentration of radium and thorium in different layers of soil. The results were supplemented by statistical analysis to identify patterns of changes in activity depending on the depth.

Within the framework of the project, a machine learning method was used, in particular the random Forest algorithm (Random Forest), to classify radiation levels in the soil based on measurements of gamma activity carried out at different depths [5]. The use of machine learning allows you to effectively analyze large amounts of data, make predictions, and identify hidden patterns that are not visible in traditional analysis methods [6].

Machine learning (ML) is a field of artificial intelligence that allows computer systems to learn from data, find patterns and make predictions, or make decisions without specific programming [7]. Unlike traditional algorithms that follow clearly defined steps, machine learning algorithms analyze data and create experience-based models, that is, they are taught using examples.

This project used the Random Forest algorithm to classify the level of soil gamma activity depending on measurements at different depths. The task was to classify activity levels as low, medium and high based on radiation activity data at different depths (from 0 to 100 cm).

Results

To apply statistical analysis and machine learning algorithms, a database of the following data was designed:

- Data for conducting spur gamma surveys with points for sampling layered samples from shurfs at intervals of 0-20 cm, 20-40 cm, 40-60 cm, 60-80 cm and 80-100 cm;
- Data on the power of the exposure dose of gamma radiation (mcr/h) in the spur at a depth of 0 to 100 cm;
- results of laboratory studies of radionuclide and chemical composition samples, assessment of the ability of radionuclides to migrate to underground horizons;
- results of comparison of experimental and estimated values for the indicator of the total specific beta activity of samples;
- results of comparison of experimental and estimated values for the indicator of specific activity of radium-226 (Ra 226), Thorium-232 (Th 232) in samples.

To analyze the data of each table in the project, the main statistical characteristics were calculated: number of records (count), mean (mean), standard deviation (std), maximum value (max) and minimum value (min). These indicators allow you to get a general idea of the distribution of data and identify possible deviations. The results of statistical analysis of indicators of laboratory tests are shown in Figure 1.

As a result of the analysis, data on the average activity of radionuclides are obtained for each site and the level of contamination is determined: low, medium or high. This makes it possible to classify zones by the degree of radiation pollution and highlight the most problematic areas that need priority when developing measures to reduce the radioactive background.

The graph allows you to see how activity changes with depth and allocate abnormal values, such as sharp jumps or consistently high levels of activity. This information helps to assess the vertical distribution of pollution in the soil and isolate Spurs that need to be studied in detail. The graph of the distribution of gamma activity (μr/H) by the depth of each Spur is shown in Figure 2.

The resulting graph depicts the depth distribution of alpha activity for each spur. The graph shows how the values of alpha activity at different soil levels change, which makes it possible to identify trends, fluctuations or local areas with high activity. This information is useful for assessing the vertical distribution of Alpha-active radionuclides contamination and helps identify areas that require additional analysis or intervention. The depth distribution graph for each spur of alpha activity is shown in Figure 3.

The construction of color correlation matrices allows to analyze the relationships between experimental and computational data on dif-

	ID_shurf_l	0-20cm	20-40cm	40-60cm	60-80cm	80-100cm
count	1148.000000	1148.000000	1148.000000	1148.000000	1148.000000	1148.000000
mean	14.500000	21796.478223	21638.053136	21749.425958	21758.089721	21799.371080
std	8.081268	98213.973729	98495.475951	98854.773914	99571.080956	99816.105981
min	1.000000	2.000000	2.000000	2.000000	2.000000	2.000000
25%	7.750000	24.750000	20.000000	17.000000	12.000000	12.000000
50%	14.500000	120.000000	104.000000	94.000000	82.000000	81.000000
75%	21.250000	941.000000	963.000000	858.250000	611.250000	569.750000
max	28.000000	654294.000000	656511.000000	657697.000000	649961.000000	655761.000000

Figure 1 - Results of statistical analysis of indicators of laboratory tests

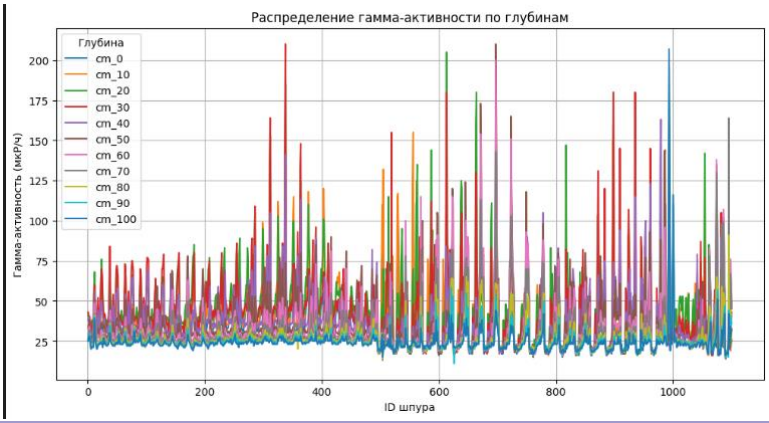


Figure 2 - Graph of the distribution of gamma activity (μr/H) by the depth of each spur

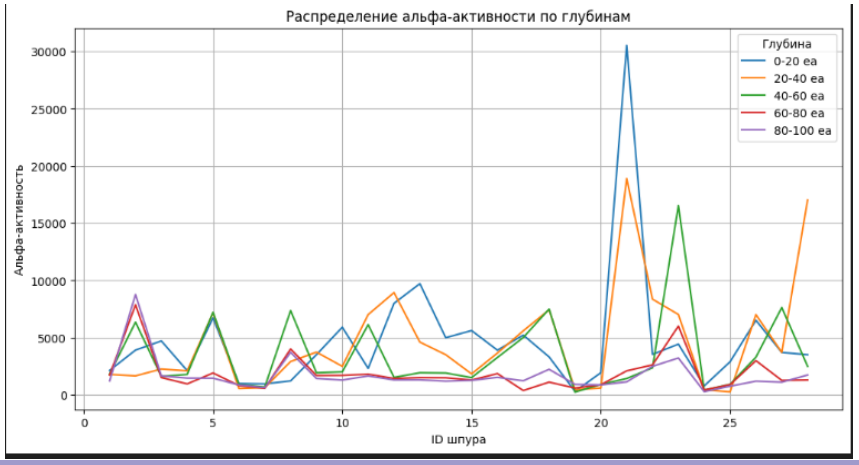


Figure 3 - Graph of the depth distribution of alpha activity for each spur

ferent types of activity (Alpha, Beta, Ra226, Th232) at a depth of 0-100 CM [8]:

• Ra226: the correlation matrix shows the degree of correspondence between experimental (er) and calculated activity values at all depths. High correlation indicates the accuracy of the calculation model.

• Th232: a similar analysis determines the nature of the relationship between experimental and calculated data on thorium activity, which is especially important for the study of deep distribution.

• Beta activity: the correlation matrix helps to assess how close the experimental and cal-

culated data are, which makes it possible to identify possible deviations in the calculations.

• Alpha activity: a similar approach is used to analyze alpha activity across different layers, giving a complete idea of the quality and accuracy of the data.

The creation of these color correlation matrices helps to assess the quality and accuracy of calculations compared to experimental measurements, identify possible deviations or discrepancies between layers, and decide on the need for further improvement or additional analysis of computational models. To analyze the relationship of gamma activity at different depths, a correlation matrix is developed that reflects the degree of relationship between soil layers. In increments of 10 cm, activity indicators from 0 to 100 cm are considered and the correlation coefficient is calculated for each depth pair, which shows how much changes in one layer depend on changes in another. Correlation values-from 1 (full negative connection) to 1 (full positive connection), where 0 indicates the absence of a connection. The results are presented as a heat map, where the color scale shows the strength of the correlation: the brighter the color, the stronger the connection. The correlation matrix of gamma activity is shown in Figure 4.

Correlation matrices have helped determine the relationship between activity at different depths, which is important for understanding the migration patterns of radionuclides in soils.

Before applying machine learning algorithms, the data is imported from a file containing information about the results of gamma radiation at different soil depths. These data include radiation values at levels from 0 to 100 cm at intervals of 10 centimeters. Each of these levels is assigned names, which makes the table more readable.

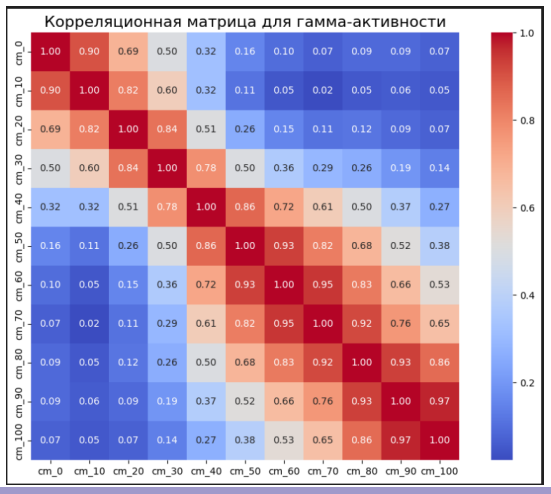


Figure 4 - Gamma activity correlation matrix

A machine learning model is trained to classify radiation levels of shurfs using a random forest algorithm (Random Forest).For this, a set of signs (X) differs from the initial data, which are radiation values at different depths (from 0 CM to 100 cm) for each spur, and the target variable (y) is a classification of radiation levels (for example, "low", "medium", "high"). The data is then divided into training and test samples using the train_test_split function. 80% of the data is used to train the model and the remaining 20% is used to test it. The model is trained using a random forest algorithm.

After training the model, it is used to predict classes for test data. Then the accuracy of the model is calculated, which shows how correctly it classified the test data. The result is produced in the form of classification accuracy, expressed as a percentage. This process allows us to assess how effectively the model can classify shurfs by radiation levels using the data provided. At the final stage, a process is carried out that allows the user to enter data into a new Spur, and this data is then used to predict the radiation activity class (e.g. "Low", "Medium" or "high") using a trained machine learning model.

Steps to take to make a forecast:

Step 1. Get input data from the user: In the first function (get_input_data), the user is asked to enter values for each layer of soil at different depths (e.g. 0 CM, 10 cm deep, etc.)

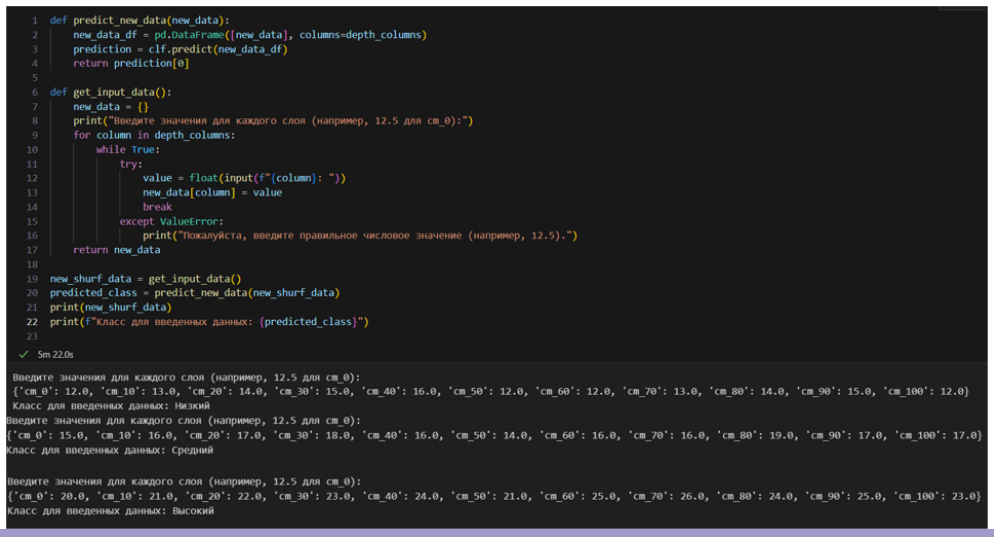
Step 2. Edit The entered data: The entered data is collected in a dictionary, where the keys are layer names (for example, 'cm_0', 'cm_10', etc.) and the values are numeric values entered by the user After that, the dictionary becomes a DataFrame, which is transferred to the machine learning model.

Step 3. Radiation activity class prediction: In the second function (predict_new_data), the data from the new shurf is passed to the classification model (Random Forest). According to the existing data, the pre-prepared model predicts the radiation class for the entered values (for example, "low", "medium" or "high").

Step 4. Output results: After the forecast is made, the program will display on the screen the data entered by the user (for each layer) and the classification result — the predicted radiation class for the spur data.

This process makes it easy to integrate machine learning to assess the radiation background at new sites, obtaining results immediately after entering the necessary data. Making a forecast and the result of forecasting is shown in Figure 5.

The use of a random forest algorithm in this study made it possible to build a reliable model



```
def predict_new_data(new_data):
    new_data_df = pd.DataFrame([new_data], columns=depth_columns)
    prediction = clf.predict(new_data_df)
    return prediction[0]

def get_input_data():
    new_data = {}
    print("Введите значения для каждого слоя (например, 12.5 для cm_0):")
    for column in depth_columns:
        while True:
            try:
                value = float(input(f"{column}: "))
                new_data[column] = value
                break
            except ValueError:
                print("Пожалуйста, введите правильное числовое значение (например, 12.5).")
    return new_data

new_shurf_data = get_input_data()
predicted_class = predict_new_data(new_shurf_data)
print(new_shurf_data)
print(f"Класс для введенных данных: {predicted_class}")
```

✓ 5m 22.0s

```
Введите значения для каждого слоя (например, 12.5 для cm_0):
{'cm_0': 12.0, 'cm_10': 13.0, 'cm_20': 14.0, 'cm_30': 15.0, 'cm_40': 16.0, 'cm_50': 12.0, 'cm_60': 12.0, 'cm_70': 13.0, 'cm_80': 14.0, 'cm_90': 15.0, 'cm_100': 12.0}
Класс для введенных данных: Низкий
Введите значения для каждого слоя (например, 12.5 для cm_0):
{'cm_0': 15.0, 'cm_10': 16.0, 'cm_20': 17.0, 'cm_30': 18.0, 'cm_40': 16.0, 'cm_50': 14.0, 'cm_60': 16.0, 'cm_70': 16.0, 'cm_80': 19.0, 'cm_90': 17.0, 'cm_100': 17.0}
Класс для введенных данных: Средний

Введите значения для каждого слоя (например, 12.5 для cm_0):
{'cm_0': 20.0, 'cm_10': 21.0, 'cm_20': 22.0, 'cm_30': 23.0, 'cm_40': 24.0, 'cm_50': 21.0, 'cm_60': 25.0, 'cm_70': 26.0, 'cm_80': 24.0, 'cm_90': 25.0, 'cm_100': 23.0}
Класс для введенных данных: Высокий
```

Figure 5 - Making predictions and the result of predictions

for classifying levels of gamma activity in soils. The model not only showed high accuracy, but also showed good resistance to retraining. This approach can be used to monitor radiation pollution, predict radiation activity in different ecosystems, and develop measures to reduce the impact of radiation on the environment.

Conclusion

As a result of the analysis of data on radioactive activity in soil layers, important results were obtained that made it possible to assess the level of pollution and its distribution to depth. Comparison of experimental and computational data for the activity of alpha, beta, Ra226 and Th232 showed a general correspondence of models and observations, which confirms the accuracy of the calculation methods used. Pollution categories were identified based on the pollution index and the type of distribution, which made it possible to distinguish between high, deep and mixed types of pollution.

All practical work is carried out at the D. Serikbayev school. Conducted at the East Kazakhstan Technical University (Oskemen city, Kazakhstan).

REFERENCES

1. Кумарбекулы С. и др. Analysis of radiation city of Ust-Kamenogorsk // Актуальные научные исследования в современном мире. – 2020. – № 3-2. – С. 107-110.
2. Манапов Д. Д. и др. Radiation condition of river channel Komendantka // Там же. – № 3-7. – С. 152-157.
3. Алибаева Л. Ж., Жолтабарова Ш. М., Иминова Д. Е. Современная радиометрическая обстановка населенных пунктов повышенного радиационного риска бывшего СИЯП // Перспективы развития науки в современном мире. – 2019. – С. 132-137.
4. Маркин М. Ю. Оценка объемной активности радона тектонических разломов в границах городской агломерации Усть-Каменогорск (Республика Казахстан) // Геология и геофизика Юга России. – 2023. – Т. 13. – № 2. – С. 29-39.
5. Попова А. А. Методы машинного обучения в экологии // Современные технологии в российской и зарубежных системах образования. – 2020. – С. 151-154.
6. Костромин Н. С., Сивова А. Н. Применение методов машинного обучения для решения экологических задач // Modern science. – 2019. – № 5-3. – С. 144-148.
7. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – Litres, 2022.

8. Лаврентьева, Г. В., Удалова, А. А., Непогодина, Я. В., Мельникова, Т. В. (2024). Экологический мониторинг почвенно-растительной системы в зоне влияния радиационно опасного объекта неэнергетического профиля // Экология урбанизированных территорий. 2024. (3), 42-47.
9. Mahur, A. K., Sharma, R. L., Mehra, R., Chand, S., Singh, H., & Sharma, S. (2024). Natural radioactivity, radon exhalation rates and radiation doses in soil and cement samples. Journal of Radioanalytical and Nuclear Chemistry, 1-8.
10. Баврина А. П., Борисов И. Б. Современные правила применения корреляционного анализа // Медицинский альманах. – 2021. – № 3 (68). – С. 70-79.

Өскемен қаласының топырағының радиоактивті ластану деректерін машиналық оқыту алгоритмдері арқылы талдау

[1]УВАЛИЕВА Индира Махмутовна, *PhD, профессор, iuvalieva@mail.ru,*
[1]*ИДРИШЕВА Жанат Кабылбековна, *т.ғ.к., аға ғылыми қызметкер, zhanat.idr@mail.ru,*
[1]БУКУНОВА Альмира Шошановна, *м.ғ.к., қауымдастырылған профессор, abukunova@edu.ektu.kz,*
[1]ШАЯХМЕТҚЫЗЫ Карина, *магистрант, karinasaahmetova88@gmail.com,*
[2]SKWAREK Ewa, *dr. hab. PhD, профессор, ewa.skwarek@mail.umcs.pl,*
[1]«Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті» КеАҚ, Д. Серікбаев көшесі, 19, Өскемен, Қазақстан,
[2]Мария Кюри-Склодовская университеті, М. Кюри-Склодовская алаңы, 3, Люблин, Польша,
**автор-корреспондент.*

Аңдатпа. *Топырақтың радиоактивті ластануы табиғи және антропогендік экожүйелерге әсер ететін ең маңызды экологиялық проблемалардың бірі болып табылады. Радиациялық фонды және топырақтағы радионуклидтердің құрамын зерттеу олардың қоршаған ортаға және адам денсаулығына әсерін бағалау және әсерді азайту стратегияларын әзірлеу үшін қажет. Мақала аясында Өскемен қаласының топырағының радиоактивті ластану деректерін машиналық оқыту алгоритмдері арқылы талдау нәтижелері алынды. Жобаның мақсаты – Өскемен қаласының топырағының радиоактивті ластанудың альфа, бета, Ra226 және Th232 деректер базасының көрсеткіштерін статистикалық әдістер арқылы және машиналық оқыту алгоритмдері көмегімен талдау. Химиялық, радионуклидтік және минералогиялық құрамға іріктелген сынамаларды зертханалық зерттеудің эксперименттік және есептік деректерін салыстыру модельдері құрылды мен бақылаулардың жалпы сәйкестігі зерттелді.*

Кілт сөздер: *радиоактивті ластану, машиналық оқытудың алгоритмдері, статистикалық талдау, корреляциялық матрица, альфа белсенділігі, бета белсенділігі.*

Анализ данных радиоактивного загрязнения почвы г. Усть-Каменогорска с применением алгоритмов машинного обучения

[1]УВАЛИЕВА Индира Махмутовна, *PhD, профессор, iuvalieva@mail.ru,*
[1]*ИДРИШЕВА Жанат Кабылбековна, *к.т.н., старший научный сотрудник, zhanat.idr@mail.ru,*
[1]БУКУНОВА Альмира Шошановна, *к.м.н., ассоциированный профессор, abukunova@edu.ektu.kz,*
[1]ШАЯХМЕТҚЫЗЫ Карина, *магистрант, karinasaahmetova88@gmail.com,*
[2]SKWAREK Ewa, *dr. hab. PhD, профессор, ewa.skwarek@mail.umcs.pl,*
[1]НАО «Восточно-Казахстанский технический университет имени Д. Серикбаева», ул. Д. Серикбаева, 19, Усть-Каменогорск, Казахстан,

[2]Университет Марии Кюри-Склодовской, площадь М. Кюри-Склодовской, 3, Люблин, Польша,

**автор-корреспондент.*

Аннотация. *Радиоактивное загрязнение почвы – одна из важнейших экологических проблем, затрагивающих природные и антропогенные экосистемы. Изучение радиационного фона и содержания радионуклидов в почве необходимо для оценки их воздействия на окружающую среду и здоровье человека и разработки стратегий смягчения последствий. В рамках статьи получены результаты анализа данных радиоактивного загрязнения почв г. Усть-Каменогорска с помощью алгоритмов машинного обучения. Цель исследования - анализ альфа, бета, Ra226 и Th232 показателей базы данных радиоактивного загрязнения почвы г. Усть-Каменогорска с помощью статистических методов и алгоритмов машинного обучения. Созданы модели сравнения экспериментальных и расчетных данных лабораторных исследований проб, отобранных на химический, радионуклидный и минералогический состав, изучено общее соответствие наблюдений.*

Ключевые слова: *радиоактивное загрязнение, алгоритмы машинного обучения, статистический анализ, корреляционная матрица, альфа-активность, бета-активность.*

REFERENCES

1. Kumarbekuly S. i dr. Analysis of radiation city of Ust-Kamenogorsk // Aktual'nye nauchnye issledovanija v sovremennom mire. – 2020. – № 3-2. – Pp. 107-110.
2. Manapov D. D. i dr. Radiation condition of river channel Komendantka // Tam zhe. – № 3-7. – Pp. 152-157.
3. Alibaeva L. Zh., Zholtabarova Sh. M., Iminova D. E. Sovremennaja radiometricheskaja obstanovka naselennyh punktov povyshennogo radiacionnogo riska byvshego SIJaP // Perspektivy razvitija nauki v sovremennom mire. – 2019. – Pp. 132-137.
4. Markin M. Ju. Ocenka ob#emnoj aktivnosti radona tektonicheskih razlomov v granicah gorodskoj aglomeracii Ust'-Kamenogorsk (Respublika Kazahstan) // Geologija i geofizika Juga Rossii. – 2023. – T. 13. – № 2. – Pp. 29-39.
5. Popova A. A. Metody mashinnogo obuchenija v jekologii // Sovremennye tehnologii v rossijskoj i zarubezhnyh sistemah obrazovanija. – 2020. – Pp. 151-154.
6. Kostromin N. S., Sivova A. N. Primenenie metodov mashinnogo obuchenija dlja reshenija jekologicheskih zadach // Modern science. – 2019. – № 5-3. – Pp. 144-148.
7. Flah P. Mashinnoe obuchenie. Nauka i iskusstvo postroenija algoritmov, kotorye izvlekajut znanija iz dannyh. – Litres, 2022.
8. Lavrent'eva, G. V., Udalova, A. A., Nepogodina, Ja. V., Mel'nikova, T. V. (2024). Jekologicheskij monitoring pochvenno-rastitel'noj sistemy v zone vlijanija radiacionno opasnogo ob#ekta nejenergeticheskogo profilja // Jekologija urbanizirovannyh territorij. 2024. (3), 42-47.
9. Mahur, A. K., Sharma, R. L., Mehra, R., Chand, S., Singh, H., & Sharma, S. (2024). Natural radioactivity, radon exhalation rates and radiation doses in soil and cement samples. Journal of Radioanalytical and Nuclear Chemistry, 1-8.
10. Bavrina A. P., Borisov I. B. Sovremennye pravila primenenija korreljacionnogo analiza // Medicinskij al'manah. – 2021. – № 3 (68). – Pp. 70-79.