Machine learning for chemical-humus correlation in soil

Ivan Lebedev[1*]

[1] Moscow Aviation Institute (National Research University), 125993, Volokolamskoe shosse, 4, Moscow, Russia

Abstract. This article investigates the dependency of the quantitative content of humus in soil on phosphate (P_2O_5), potassium oxide (K_2O), hydrolytic acid, as well as the pH value in aqueous and saline environments through machine learning. Linear regression was chosen as the primary model. The mean absolute error (MAE) was found to be 0.517, mean squared error (MSE) – 0.460, and the coefficient of determination after cross-validation reached 0.685. The search for the most significant covariate among the listed ones identified hydrolytic acid as the most impactful due to its influence on microbial activity in the soil and metabolism.

1 Introduction

In recent years, there has been a significant growth in interest in applying machine learning methods across various scientific and practical fields, opening new perspectives for analyzing and interpreting complex data. The integration of these technologies in agrochemistry, where they can play a crucial role in optimizing agricultural production and sustainable soil resource management, is particularly noteworthy. One of the main aspects in this area is predicting the humus content in the soil, which is crucial for assessing its fertility as humus directly affects the physical, chemical, and biological properties of the soil, determining its water retention capacity, structure, nutrient composition, and microbial activity.

The analysis of correlations between soil chemical indicators, such as phosphorus and potassium content, acidity levels, and others, and humus levels using machine learning methods forms the basis for developing effective prediction models. These models can adapt to changing conditions and the specifics of different soil types. Such models become an indispensable tool for agronomists and farmers, providing them with information for making decisions on soil treatment, fertilization, and crop rotation plans aimed at increasing yield and maintaining the ecological balance in agricultural ecosystems [1-3]. This approach can complement field analysis using satellite imagery [4] and unmanned aerial vehicles [5].

2 Materials and Methods

In the conducted research, 70 soil samples were selected for comprehensive chemical analysis to determine key parameters, including concentrations of phosphate (P_2O_5),

* Corresponding author: lebedev.ivan.ig@yandex.ru

potassium oxide (K_2O), hydrolytic acid, as well as pH values in aqueous and saline environments and humus content. Given the relatively small data sample size, the application of deep learning models was deemed impractical due to their tendency to overfit in limited dataset conditions. Instead, linear regression was chosen as a method demonstrating high efficiency with small and medium data volumes and capable of providing good result interpretability. Linear regression establishes linear dependencies between independent variables (chemical indicators) and the dependent variable (humus content).

After training the model, the following performance indicators were obtained: the mean absolute error (MAE) was 0.517, the mean squared error (MSE) was 0.460, and the coefficient of determination (R^2 score) reached 0.668. The mean absolute error reflects the average deviation of predicted values from actual ones, providing an understanding of the overall accuracy of the model in absolute terms. The mean squared error, being the square of the difference between predicted and real values, penalizes large errors more severely than small ones, serving as an indicator of the model's consistency. The coefficient of determination, ranging from 0 to 1, shows the proportion of the variance in the dependent variable that is predictable from the model, and is a key indicator of its predictive power. The higher the value, the better.

To strengthen confidence in the results and assess their robustness, cross-validation [6-8] was applied, allowing for a stricter test of the model's predictive power. During cross-validation, data is divided into several parts, the model is trained on one part of the data and tested on another, which is repeated several times for different partitions. This provides a more objective assessment of the model's efficiency, minimizing the impact of the randomness of the selection of training and test samples. The average coefficient of determination obtained through cross-validation was 0.685, not only confirming the reliability of the preliminary results but also indicating the high stability and reliability of the linear regression model [6] in the given task.

In addition to linear regression, after obtaining predictions, the random forest algorithm [7] is applied to assess the importance of individual features. The random forest, being an ensemble of decision trees, reduces the risk of overfitting by aggregating the results of multiple trees.

Effective use of this method allowed for identifying key factors that have the most significant impact on humus content in the soil, enabling the exclusion of less significant variables from the models. This optimizes their structure and improves result interpretability.

3 Results

The following graphs show the dependency of humus content in the soil on each of the covariates.

A blue trend line on the graph represents the trend line calculated using the least squares method, showing the main direction of dependency between two variables. In the context of linear regression, this line represents the best linear approximation of the data, meaning it's the line through which the sum of the squares of the vertical distances from each observed point to the line is minimized. The area around the trend line, often referred to as the confidence band or interval, reflects uncertainty in trend line estimates. This band indicates the degree of confidence one can have that the true trend line falls within this area. The width of this band depends on the data variability: the greater the variability, the wider the confidence band. The marked points on the graph represent the actual observed data. Each point corresponds to a pair of values — one for the independent variable (on the x-axis) and one for the dependent variable (on the y-axis). The distribution of these points and their proximity to the trend line give an indication of the strength of the linear relationship between the variables. If the points are widely scattered around the trend line and the confidence band

is wide, this indicates a weak linear dependency. Conversely, if the points are closely aligned with the trend line and the confidence band is narrow, this indicates a strong linear dependency between the variables.

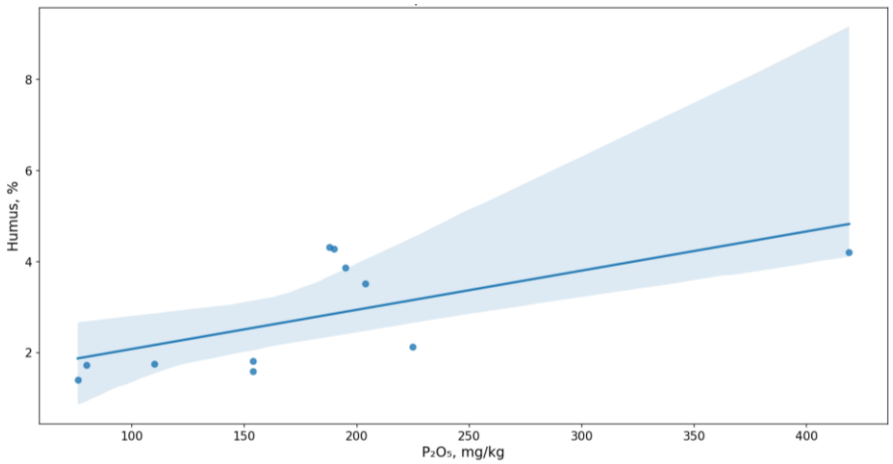

Fig 1. The relationship between humus and P_2O_5

The trend line (figure 1) through the data points shows a positive correlation, suggesting that an increase in P_2O_5 levels is associated with an increase in soil humus content. This relationship can be explained by the interaction of phosphorus with organic substances in the soil. The confidence band surrounding the trend line is relatively narrow, indicating that the model's predictions are reliable and can be generalized with a given level of confidence. However, there are a few points outside the confidence band, which may indicate potential outliers or the presence of unaccounted variables that could influence humus levels.

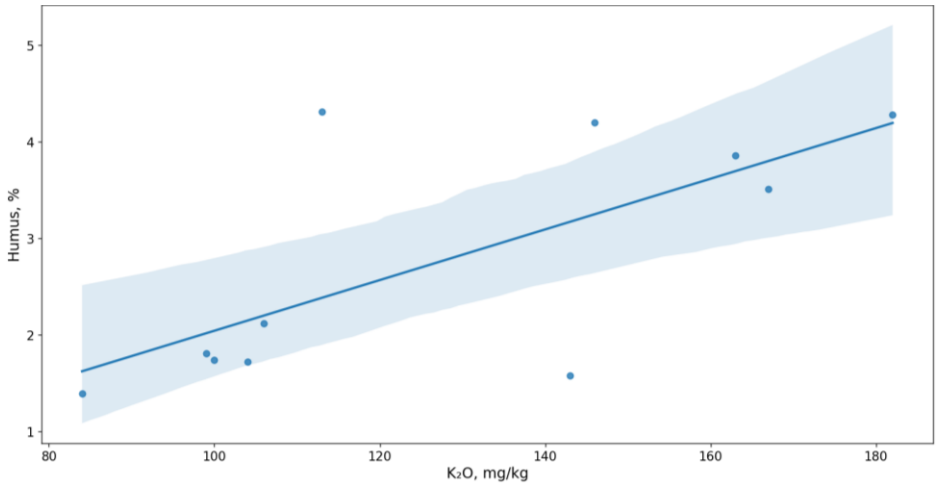

Fig 2. The relationship between humus and K_2O

The trend line (figure 2) suggests a moderate positive correlation between the increase in K_2O levels and an increase in humus content. The confidence interval covering the trend line spans a broad range, indicating greater uncertainty in the data compared to previous variables. This result may be an indicator that potassium oxide plays a role in improving soil quality, though the data variability highlights the need for further research to clarify the relationship.

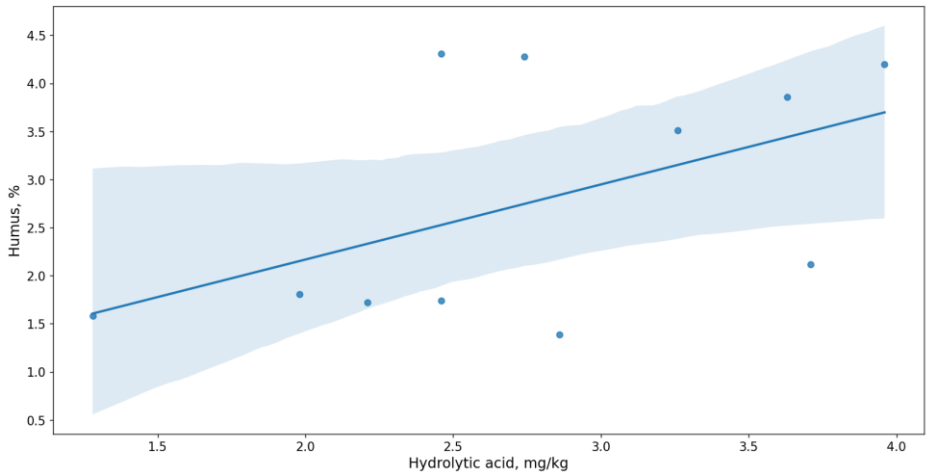

Fig 3. The relationship between humus and hydrolytic acid

The graph (figure 3) depicting the relationship between soil hydrolytic acid level and humus content shows a positive correlation, where an increase in hydrolytic acid content is accompanied by an increase in humus level. This link points to the role of hydrolytic acid in processes affecting the enrichment of soil with organic matter. The confidence interval band is relatively narrow, indicating less uncertainty in predictions compared to other variables.

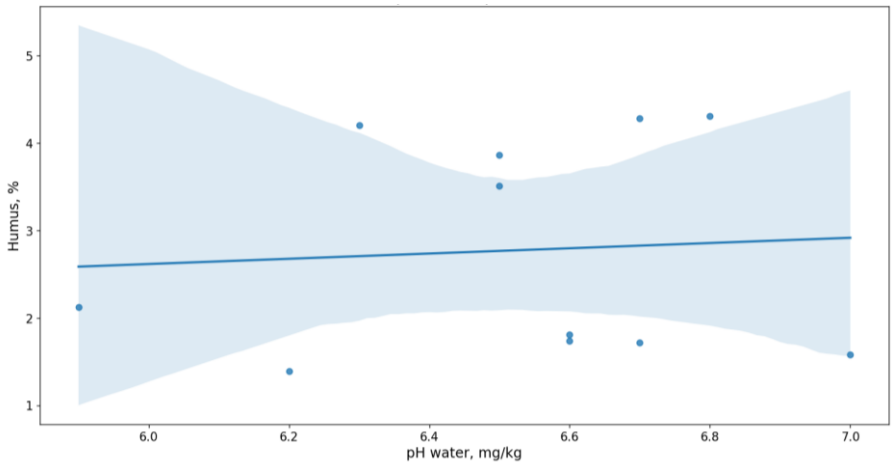

Fig 4. The relationship between humus and water pH

The trend line (figure 4) shows a relatively stable humus level as water pH changes, indicating a weak or nonexistent correlation between these two variables within the studied pH range. The wide confidence interval around the trend line signifies significant data uncertainty and the potential influence of other factors on humus content. These results suggest that water pH is not a primary indicator of humus content in the studied soils or that the influence of pH may only manifest over a broader range of values or in combination with other soil properties.

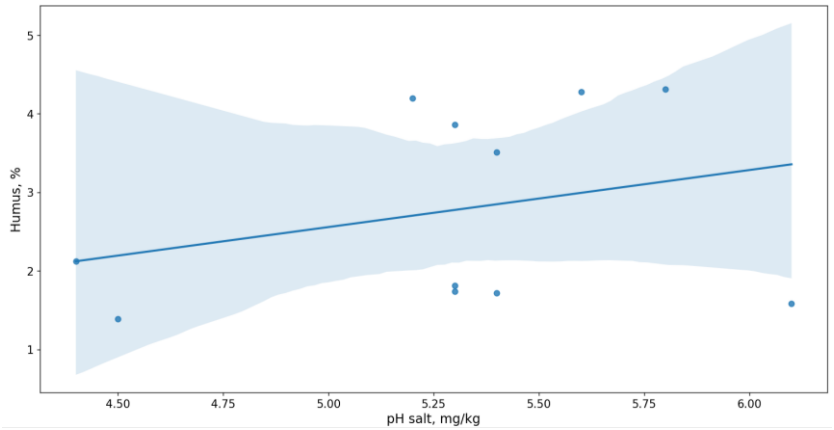

Fig 5. The relationship between humus and salt pH

The trend line (figure 5) indicates a moderate positive correlation, suggesting that an increase in the pH of the saline extract leads to an increase in humus content. The wide confidence band speaks to significant data variability, and while a tendency for increased humus with higher pH is observed, the possible influence of other soil characteristics should be considered. This relationship may be important for agronomy, as the pH of the saline extract affects nutrient availability and microbial activity, which, in turn, can influence soil fertility and agricultural productivity.

4 Conclusions

Based on the presented graphs, conclusions can be drawn about the degrees of correlation between the measured soil chemical parameters and humus content. Notably, the presence of phosphorus as P_2O_5 and hydrolytic acid correlates with humus levels, suggesting their significant role in ensuring soil fertility. It's worth noting that the values of water pH and saline extract pH were less significant, which may indicate their secondary role or the need for further study of their interaction with other soil components.

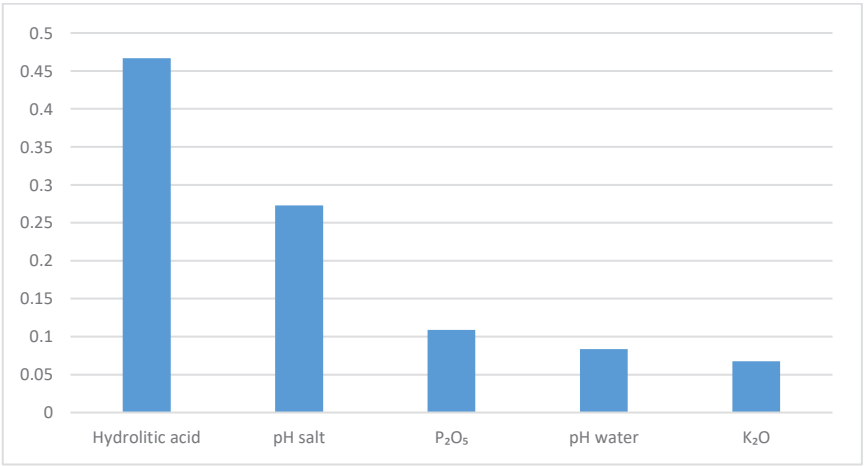

Fig 6. The degree of influence of each of the covariates

Moving to the feature importance analysis presented in the bar chart (figure 6), hydrolytic acid has the highest relative importance, related to its impact on microbial activity in the soil and metabolism, which, in turn, directly influences humus formation. The influence of saline extract pH also proved significant, highlighting the importance of the soil's saline balance for retaining organic matter and nutrient availability. On the other hand, P_2O_5 concentration, as well as water pH and K_2O levels, have less relative importance, which may reflect the more complex interactions of these elements with the soil environment or their indirect impact on humus content.

Acknowledgements

The research was carried out at the expense of the grant of the Russian Science Foundation No. 23-74-01050, https://rscf.ru/en/project/23-74-01050

References

1. Ogorodnikov, S.S. IOP Conference Series: Earth and Environmental Science **723(4)**, 042053(2021)
2. Ogorodnikov, S.S. IOP Conference Series: Earth and Environmental Science **1010(1)**, 012040 (2022)
3. Rozanov, V. BIO Web of Conferences **84**, 01018 (2024)
4. Ogorodnikov, S.S., Sorokin, A.E. Russian Engineering Research **42(6)**, 639-641 (2022)
5. Sorokin, A.E., Zhelonkin, M.V., Ogorodnikov, S.S. Russian Engineering Research, **42(12)**, 1315-1317 (2022)
6. Wang, Y., Huang, Q., Yao, Z., Zhang, Y. Journal of Complexity **(82)**, 101826 (2024)
7. He, Z., Wang, J., Jiang, M., Hu, L., Zou, Q. Information Sciences **(667)**, 120478 (2024)
8. Cross-valid Deng, A. Economics Letters **(233)**, 111369 (2023)