JEE Journal of Ecological Engineering

*Journal of Ecological Engineering*, 2026, 27(5), 27–45
https://doi.org/10.12911/22998993/215563
ISSN 2299–8993, License CC-BY 4.0

Received: 2025.12.10
Accepted: 2026.03.16
Published: 2026.04.01

# Soil quality assessment based on agrochemical indicators and optimized multiple linear regression

Dmytro Tymoshchuk[1*], Halyna Humeniuk[2], Andrii Sverstiuk[3], Mariana Prokopiak[4], Oksana Matsiuk[4]

[1] Department of Artificial Intelligence Systems and Data Analysis, Ternopil Ivan Puluj National Technical University, Ruska Street, 56, 46001 Ternopil, Ukraine
[2] Department of General Biology and Methods of Teaching Natural Sciences, Ternopil Volodymyr Hnatiuk National Pedagogical University, 2 Maxyma Kryvonosa Street, 46027 Ternopil, Ukraine
[3] Department of Medical Informatics, I. Horbachevsky Ternopil National Medical University, Maidan Voli, 1, 46001, Ternopil, Ukraine
[4] Department of Botany and Zoology, Ternopil Volodymyr Hnatiuk National Pedagogical University, 2 Maxyma Kryvonosa Street, 46027, Ternopil, Ukraine
* Corresponding author's e-mail: dmytro.tymoshchuk@gmail.com

## ABSTRACT

Soil quality assessment is a key component of sustainable land use, agroecosystem monitoring, and the optimization of agrochemical management. The aim of this study is to evaluate the soil quality within the Rozsoshansk community of the Khmelnytskyi region using the soil quality index (SQI) and to develop a statistically sound predictive model based on agrochemical indicators. The analysis was conducted using a set of soil parameters, including *pH*, $C_{org}$, $NH_4^+$, $NO_3^-$, $P_2O_5$, $Ca^{2+}$ i $K^+$. Spearman correlation analysis was performed, followed by the construction of a multiple linear regression model ($R^2 = 0.904$, Adjusted $R^2 = 0.900$), an assessment of multicollinearity (VIF), and ANOVA. Based on the results of the regression model with seven predictors, together with ANOVA and VIF diagnostics, an optimized model with four predictors (*pH*, $C_{org}$, $NO_3^-$, $Ca^{2+}$) was developed. This refined model demonstrated strong explanatory power ($R^2 = 0.856$, Adjusted $R^2 = 0.853$) and low prediction error. Residual diagnostics indicated deviations from normality and heteroscedasticity; however, robust estimation methods (OLS-HC3 and RLM using HuberT) confirmed the stability of the coefficient estimates. The findings suggest that soil organic carbon, exchangeable calcium, acidity, and nitrate nitrogen are the key indicators of soil quality in the study area. Future work will focus on applying machine learning methods for soil classification, integrating Explainable AI techniques to enhance the interpretability of predictive models.

**Keywords:** SQI, agrochemical indicators, MLR; ANOVA, VIF, robust methods, OLS-HC3, RLM.

## INTRODUCTION

Sustainable agricultural development in Ukraine requires continuous monitoring of soil conditions, particularly their chemical composition and fertility, which enables the maintenance of high crop yields without compromising environmental integrity or resource potential (Shchesniak et al., 2025). The Khmelnytskyi region is one of the key agricultural areas of Ukraine, where farming plays a central role in local economic development. The agroecological characteristics of this region deserve special attention (Pichura et al., 2023), the Khmelnytskyi region demonstrates substantial potential for sustainable land management, while simultaneously requiring the implementation of innovative soil quality assessment technologies, including digital tools and modeling approaches (Pichura et al., 2023).

In the context of integrated soil quality (SQ) assessment and the prediction of land productivity, regional soil characteristics are of particular importance. These features must be considered when developing predictive models, especially

in the Khmelnytskyi region, where soils vary in acidity, organic carbon content, and nutrient availability (Maksymenko et al., 2022).

Agriculture is a key sector that ensures food security and substantially contributes to the country's economic development. Despite the rapid advancement of modern technologies, the agricultural sector remains a priority area where technical innovations are actively implemented and highly valued. Improving the efficiency of crop cultivation and optimizing fertilizer application requires a thorough understanding of soil chemical properties, particularly the concentrations of essential nutrients (agrochemical indicators) (Madhumathi et al., 2020).

Soil nutrition plays a crucial role in maintaining soil fertility and supporting plant growth and development (Vacca et al., 2016; Camenzind et al., 2017; Chagnon et al., 2017). Numerous studies have attempted to quantify and emphasize the importance of changes in soil nutrient conditions and nutrient availability (Grove et al., 2017; Murphy et al., 2017; Bassaco et al., 2018). For this reason, an accurate evaluation of soil fertility is essential for effective planning of restoration and management measures.

The selection of soil fertility indicators and their statistical analysis for deriving a soil quality index varies considerably across studies (Ebrahimi et al., 2017). According to the reported findings (Velasquez et al., 2007), the identification of relevant indicators and agrochemical parameters that reflect different aspects of soil quality is of particular significance. To minimize subjectivity in expert assessments, statistical tools such as principal component analysis, multiple correlation, factor analysis, and cluster analysis can be employed to compute soil quality indices. These approaches allow various soil properties to be integrated into a single composite index, thus ensuring an objective and data-driven evaluation (Roudier et al., 2017; Bachmann and Kinzel, 1992; Doran and Parkin, 1996).

Despite the recognized importance of soil quality assessment, a universally accepted framework for its definition has not yet been established. In the United States, the concept of soil quality incorporates soil fertility, productivity, resource sustainability, and environmental quality, whereas in Canada and much of Western Europe, soil contamination forms the central focus of the soil quality paradigm (Kawamura et al., 2017).

The authors (Singer and Ewing, 1998; Sojka and Upchurch, 1999) suppose that creating a universal and simple soil quality index is virtually impossible. The selection of indicators or composite indices always depends on the specific objectives of ecosystem management. For example, if the goal is to achieve agroecosystem sustainability, the soil quality index becomes only one component within the broader sustainability hierarchy. Moreover, management objectives may differ depending on the interests of various groups involved in agricultural activities.

Soil quality indices (SQI) integrate the factors and processes that determine soil quality (Zhao et al., 2009; Bindraban et al., 2000). Quantifying changes in soil nutrient stocks is essential for identifying problematic land-use systems. The soil nutrient balance indicator integrates two key characteristics: the dynamics (rate of change) of nutrient levels and their overall content. Thus, the SQI is a tool that integrates multiple types of data into a single value that can be used to compare one soil with another, improve understanding of soil processes, and inform measures required for soil improvement or restoration (Zhang et al., 2017; Bhardwaj et al., 2011). This tool helps assess how soil quality changes under different land-use systems (Masto et al., 2008). SQI is used to evaluate the potential effects of crop cultivation (Maksymenko et al., 2022), different agricultural practices, and soil management strategies. Its calculation is a quantitative procedure that requires identifying a minimum set of essential soil attributes, assessing them, and integrating these parameters into a unified index (Forkuor et al., 2017). The methods used to select a representative subset of soil properties from a large number of available characteristics are generally referred to as the minimum data set.

Soil organic carbon is the foundation of soil quality assessment because it regulates fertility, structure, water-holding capacity, and biological activity. High-quality crop production depends on an adequate amount of organic carbon, which supports the fundamental ecological functions of the soil (Prakash, 2018; Bentsen et al., 2019).

There is currently no universally accepted dataset for monitoring soil quality or interpreting relevant indicators. Previous studies have shown that multivariate statistical analysis is an effective tool that facilitates the identification of soil quality parameters and the interpretation of correlated variables in a joint analytical framework (Wander

and Bollero, 1999; Brejda et al., 2000a, 2000b; Schipper and Sparling, 2000; Chaudhury et al., 2005; Govaerts et al., 2006).

According to the findings presented in (Schipper and Sparling, 2000), there is no need to measure all soil parameters if some of them are highly correlated, which allows the construction of a more rational minimum data set. A predictive model developed to estimate the content of key agrochemical elements in the soil, including nitrogen (N), phosphorus (P), and potassium (K), used multiple linear regression (MLR) and demonstrated sufficient accuracy in capturing the relationship between these nutrient concentrations. This is important for assessing soil fertility. The application of the MLR model for predicting the levels of N, P, and K produced an accuracy of approximately 78%, indicating that the model can serve as an effective tool for agrochemical soil evaluation. This approach can be considered a dependable method for assessing soil fertility and supporting the efficient management of agricultural production systems. In (Yu et al., 2020), machine learning techniques such as MLR and classification algorithms were used to compare their performance in analyzing the concentrations of elements (Cd and Se) in soils and plants.

Accurate and detailed spatial information about soil is essential for environmental modeling, risk assessment, and decision-making. The study (Forkuor et al., 2017) examined the use of high-resolution satellite imagery together with terrain and climate data, as well as laboratory-analyzed soil samples, to map the spatial distribution of six soil properties. Four statistical prediction models were tested, and their accuracy was evaluated to determine their effectiveness in predictive tasks.

Agricultural analysis employs a wide range of machine learning models, including regression and classification of algorithms that enable high-precision prediction. Sensor technologies and devices designed using intelligent optimization methods are also widely applied.

Quantitative assessment of soil nutrient status relies on both process-based and empirical models. For example, MLR has been used to predict soil organic matter stocks during spatial downscaling (Ebrahimi et al., 2017). Such models offer a distinct advantage due to their simplicity and ease of practical implementation (Du, 2016).

MLR can be used to establish relationships between soil organic carbon stocks and other soil properties (Meersmans et al., 2008; Liu et al., 2015; John et al., 2020). Many studies employing MLR focus on modelling organic carbon stocks across large geographical areas, including entire countries or continents (Forkuor et al., 2017). In this context, multivariate linear regression is applied to examine the relationships between labile carbon fractions and other soil characteristics such as texture, pH, nutrient content, and fertilizer application rates.

The study in (Shukla et al., 2006) aimed to determine the SQI using factor analysis. Factor analysis reduced a large number of measured soil variables to a smaller set of principal factors, which simplified the soil quality monitoring process and improved its efficiency and effectiveness.

Despite the considerable number of studies devoted to modeling soil organic carbon content using MLR, as well as analyzing soil physical properties and microelements, other important agrochemical characteristics remain insufficiently explored. The lack of an integrated approach to the quantitative evaluation of these indicators limits the potential for effective management of soil resources in agricultural production. This highlights the need to develop new predictive models that incorporate multiple agrochemical parameters and enable the identification of their relationships.

The aim of the study is to assess soil quality using the soil quality index and to construct an optimized multiple linear regression model, which, based on agrochemical indicators, allows identifying key indicators of fertility and improving the effectiveness of decision-making regarding soil resource management.

## MATERIALS AND METHODS

### Study area description

The assessment of soil conditions in the Rozsoshansk community of the Khmelnytskyi region (Ukraine) was carried out in the area between the villages of Skarzhyntsi and Perehonka, within the Vovk River basin. The study area is characterized by a mosaic soil cover, shaped by the complexity of the local terrain. Elevation differences, small depressions, and slope breaks create heterogeneous conditions for the water and air regimes of the soil profile. Variations in altitude and slope exposure give rise to diverse hydrothermal conditions, which in turn determine spatial heterogeneity in

humus accumulation, carbonate leaching, and podzolization processes within the soil cover.

The dominant soils in the study area are podzolized chernozems formed on loess-like loams. These soils represent a transitional type between classical chernozems and dark-grey podzolized soils. They combine agronomic characteristics typical of both chernozems and grey forest soils, such as a high content of organic carbon and increased susceptibility to the leaching of mineral components (Polupan and Velychko, 2019; Natsionalnyi atlas Ukrainy, 2009).

The location of the study plots ensures representative coverage of slope exposures, elevation gradients, and agricultural land types, allowing the spatial variability of agrochemical indicators to be captured ($pH\ (KCl)$), $C_{org}$, $NH_4^+$, $NO_3^-$, $P_2O_5$, $Ca^{2+}$, $K^+$). (Figure 1).

Classification approaches applied to these soils differ depending on the classification framework. The international FAO/WRB system and national soil classification systems of individual countries often assign these soils to different categories. Despite these discrepancies, field observations and laboratory analyses confirm that podzolized chernozems are complex formations that combine features of both podzolized and chernozem soils. This combination of properties requires adaptive agricultural technologies and soil management practices that account for local variations in moisture, microclimate, and nutrient regimes (FAO, 2014).

## Soil sampling and preparation

A total of 192 soil samples were collected for the study in accordance with the requirements of the State Standard of Ukraine (DSTU 4287:2004, 2005). Soil intended for laboratory analysis was sampled at a moisture level that facilitated sieving. The collected material was air-dried to a constant weight and passed through a 2-mm mesh sieve to remove stones and plant residues. Any surface plant material, visible roots, large fragments of herbaceous or woody vegetation, and visible soil fauna were removed to minimise contamination by fresh organic carbon. The prepared samples were placed in paper bags with labels indicating the farm name and address, field number, soil type, and sampling date (DSTU ISO 11464:2007, 2009; DSTU EN ISO/IEC 17065, 2016).

The samples were transported to the laboratory under conditions that prevented wetting or contamination. After delivery, they were stored in the dark at a temperature of (4 ± 2) °C with free air circulation.

## Laboratory analyses and soil quality index

Combining several soil characteristics into a single index requires careful consideration of sampling procedures and the variability of individual soil properties. Agrochemical indicators were selected because they provide a basis for assessing the chemical fertility of the soil. The


**Figure 1.** Location of the study sites

present study is focused on key agrochemical parameters, including soil exchange acidity (*pH* (*KCl*)), organic carbon ($C_{org}$) , ammonium nitrogen ($NH_4^+$) nitrate nitrogen ($NO_3^-$), available phosphorus expressed as $P_2O_5$, exchangeable calcium ($Ca^{2+}$), and potassium ($K^+$).

Nitrogen content was determined in two forms, ammonium nitrogen and nitrate nitrogen, according to the method described in (DSTU 4725:2007, 2008). Available forms of phosphorus and potassium were measured using the Chirikov method (DSTU 4115:2002, 2003). Calcium content was analysed according to (DSTU 7861:2015, 2016), exchange acidity (DSTU ISO 10390:2001, 2002), and total organic matter using the Tyurin method (DSTU 4289:2004, 2005). All chemical reagents were of the "osch" (especially pure) or "hch" (chemically pure) grade.

The proposed methodological framework makes it possible to develop models capable of comparing soil quality across regions with relatively homogeneous soils and climatic conditions. In our study area, where sampling sites were distributed across several square kilometers, this approach provides a reasonable basis for valid comparison. The SQI assessment procedure consisted of three steps: selection of indicators, interpretation of indicator values, and integration of the results into a single soil quality index. Converting agrochemical indicators into scores allowed the quantitative evaluation of soil quality using a unified measurement scale (TsINAO, 1994). The assessment used a scale from 0 to 44. Low soil quality corresponded to scores from 1 to 15, medium quality to scores from 16 to 22, and high quality to scores from 23 to 44. The scoring procedure was performed using macros developed in MS Excel.

Figure 2 illustrates the distribution of soil pH, organic carbon content, and ammonium nitrogen.

The average soil pH value is 7.4, indicating a slightly alkaline (near-neutral) reaction of the soil environment. Such conditions are favorable for the uptake of macro- and micronutrients by most crops and support the activity of soil microbiota that facilitate the mineralisation of organic residues. The organic carbon content is 3.8%, which characterizes the soil as having an elevated level of organic matter. Values exceeding 3% reflect a stable humus profile, improved physicochemical properties, and enhanced buffering capacity of the soil. The mean ammonium nitrogen content was only 5.0 mg/kg, which corresponds to a low level. This can be attributed to the high mobility of ammonium and its rapid conversion to nitrate under the activity of nitrifying microorganisms. Low concentrations of ammonium nitrogen may limit the growth of crops that more efficiently absorb ammonium forms, such as rice and certain legumes (Marschner, 2012; Robertson and Groffman, 2015).

Figure 3 illustrates the distribution of nitrate nitrogen, available phosphorus, exchangeable calcium, and potassium.

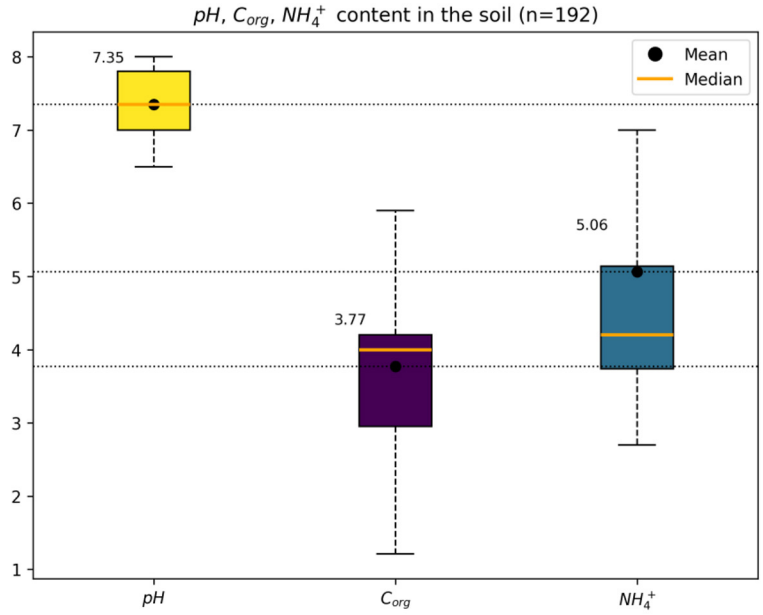


**Figure 2.** Variation in soil pH, organic carbon content, and ammonium nitrogen (mg/kg)

The concentrations of soil nutrients varied across a wide range. The mean nitrate nitrogen content was 31.3 mg/kg, which corresponds to a high level of nitrogen availability. This indicates intensive mineralization of organic matter or the application of nitrogen fertilizers. The average content of available phosphorus was approximately 64.5 mg/kg, which corresponds to a medium level of supply. The considerable variability in phosphorus values may be attributed to differences in fertilizer use and soil properties. Phosphorus availability strongly depends on soil acidity. In acidic soils, it may be immobilized by $Al^{3+}$ and $Fe^{3+}$ oxides, whereas in calcareous soils it becomes fixed by $Ca^{2+}$ compounds (Hinsinger, 2001). Exchangeable calcium had an average value of 10.4 mg/kg, indicating a high level of supply. This is favorable for crops sensitive to soil acidity. Exchangeable potassium averaged 12.9 mg/kg, which corresponds to a low level of availability. Such a deficiency may limit the productivity of crops with high potassium demand, including potatoes, maize, and sugar beet (Rengel and Damon, 2008). Overall, the studied soils exhibited an excess of nitrate nitrogen and exchangeable calcium, moderate reserves of available phosphorus, and a deficiency of potassium. This pattern underscores the need to adjust fertilization strategies with greater emphasis on potassium fertilizers for selected crops.

The combination of optimal soil acidity and adequate organic carbon reserves indicates a high level of potential soil fertility, which is an important factor for sustaining the productivity of agroecosystems.

## Dataset and regression analysis

The analysis was conducted using a dataset comprising 192 samples collected within the Rozsoshansk community of the Khmelnytskyi region. The dataset contains agrochemical indicators that serve as key measures of soil fertility. The SQI was used as the dependent variable. The regression model was constructed in the form of MLR, and the parameters were estimated using the Ordinary Least Squares (OLS) method implemented in Python (statsmodels package) ("Ordinary Least Squares - statsmodels", n.d.). The independent variables (predictors) included in the regression equation were defined as follows:

- $x1$ – exchangeable acidity of the soil ($pH$);
- $x2$ – organic carbon content ($C_{org}$);
- $x3$ – ammonium nitrogen ($NH_4^+$);
- $x4$ – nitrate nitrogen ($NO_3^-$);
- $x5$ – available phosphorus ($P_2O_5$);
- $x6$ – exchangeable calcium ($Ca^{2+}$);
- $x7$ – exchangeable potassium ($K^+$).

The general form of the model is expressed as (Mohr et al., 2022):

$$\bar{y}_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \\ +\beta_3 x3_i + \cdots + \beta_j xj_i \quad (1)$$

where: $\bar{y}_i$ – represents the SQI value for the $i$-th sample, $\beta_0$ is the intercept, $\beta_1$ ... $\beta_j$ are the

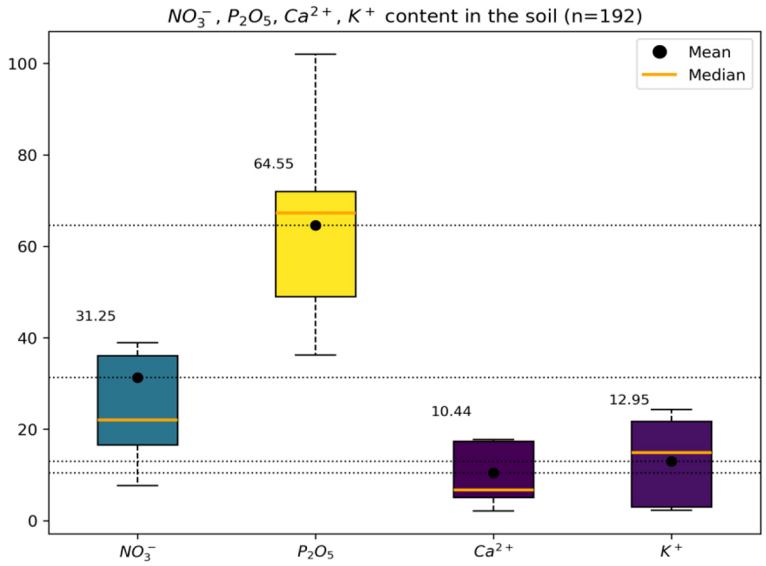


**Figure 3.** Variation in nitrate nitrogen, available phosphorus, exchangeable calcium, and potassium in soils (mg/kg)

regression coefficients describing the contribution of each indicator to SQI, and *j* denotes the number of predictors.

The OLS method estimates the model parameters $\beta_j$ by minimizing the sum of squared differences between the observed values of the dependent variable ($y_i$) and the predicted values ($\bar{y}_i$) obtained from the model. The objective of OLS is to identify the coefficient vector $\beta$ that yields the smallest possible sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \quad (2)$$

where: $\hat{\beta}$ is the estimated parameter vector obtained by OLS, and *n* is the number of samples.

## Preliminary statistical analysis

Before conducting the correlation analysis and constructing the regression model, all input variables were examined for conformity to the normal distribution. To ensure high reliability of the assessment, three complementary statistical tests were applied: the Shapiro-Wilk test, the Lilliefors test (Kolmogorov–Smirnov with estimated parameters) (GeeksforGeeks, 2025), and the Cramér-von Mises test (Zhang and Xu, 2024).

The Shapiro-Wilk test is one of the most sensitive and statistically robust tests for small and medium-sized samples. It evaluates the agreement between the empirical distribution and the standard normal distribution. A p-value < 0.05 indicates a deviation from normality. The Lilliefors test is suitable for cases in which the mean and standard deviation of the investigated variable are unknown and must be estimated from the sample. This makes the test particularly appropriate for real experimental data. A p-value < 0.05 also indicates a departure from normality. The Cramér–von Mises test measures the integrated squared distance between the empirical cumulative distribution function and the corresponding normal cumulative distribution function. It is sensitive to deviations in both the central part of the distribution and the tails, providing an additional layer of reliability when assessing normality. As with the previous tests, p < 0.05 denotes non-normality.

Predictor selection was performed to identify the most informative variables and reduce redundancy among explanatory factors. To achieve this, several complementary statistical procedures were applied, combining correlation analysis, multicollinearity diagnostics, and analysis of variance.

The choice of correlation method was based on the statistical properties of the input data, particularly the results of the normality assessment. If both variables under consideration conformed to the normal distribution, demonstrated a linear relationship, and contained no substantial outliers, the Pearson correlation coefficient was used to evaluate the strength of association. This coefficient is sensitive to violations of normality and heteroscedasticity; therefore, its application is justified only when key parametric assumptions are satisfied. In situations where at least one variable did not follow the normal distribution, or when the relationship was nonlinear, monotonic, or contained pronounced outliers, the Spearman rank correlation coefficient was applied. Unlike Pearson's coefficient, Spearman's method is nonparametric and is based on ranked values, which ensures robustness to skewed distributions, nonlinear patterns, and other departures from parametric assumptions.

The methodological principle of the Pearson correlation applied in this study was used only when normality of both variables was confirmed by all applied tests. When deviations from normality were detected in at least one of the tests, the Spearman coefficient served as the primary analytical tool (DataScientest, n.d.).

Multicollinearity among the predictors was assessed using the variance inflation factor (VIF) (Team, 2010). This measure reflects the extent to which the variance of a regression coefficient increases due to correlations with other variables. The VIF for the *j*-th predictor is defined as:

$$VIF_j = \frac{1}{1 - R_j^2} \quad (3)$$

where: $R_j^2$ is the coefficient of determination obtained from an auxiliary regression in which the variable *xj* is treated as the dependent variable and all other predictors serve as independent variables.

A VIF value close to 1 indicates the absence of multicollinearity. Values below 5 are generally considered acceptable and correspond to low or moderate multicollinearity. Values in the range of 5–10 indicate problematic multicollinearity that requires further examination. When VIF exceeds 10, strong multicollinearity is present, making the

regression results unreliable and suggesting that the corresponding variables should be removed or transformed. Identifying predictors with high VIF helps reduce redundancy in the model and ensures more robust and stable regression estimates.

To evaluate the statistical significance of predictors within the regression models, analysis of variance (ANOVA) was applied (Ståhle and Wold, 1989). This method partitions the total variability of the dependent variable into explained and unexplained components, making it possible to determine whether the inclusion of a given predictor significantly improves the model. Low p-values ($p < 0.05$) indicate that a predictor contributes meaningfully to the model, whereas high p-values suggest limited explanatory power.

Thus, the combined use of correlation analysis, VIF, and ANOVA provides a comprehensive framework for predictor selection. This approach allows for the identification of relevant variables, detection of multicollinearity, and assessment of the statistical significance of factors, enabling the construction of parsimonious and interpretable regression models.

### Model performance evaluation

A set of standard statistical metrics was used to quantify the accuracy of the regression models (scikit-learn, n.d.). These measures allow for a comprehensive evaluation of the average prediction error, relative accuracy, and explanatory power of the model.

The mean absolute error (MAE) quantifies the average magnitude of deviations between predicted and observed values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i| \tag{4}$$

This metric is interpreted as the average error expressed in the same units as the predicted variable. The mean squared error (MSE) reflects the average of the squared prediction errors:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{5}$$

This metric assigns greater weight to large deviations, making it more sensitive to outliers than MAE.

The mean absolute percentage error (MAPE) characterizes the relative accuracy of the model in percentage terms:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \bar{y}_i}{y_i}\right| \tag{6}$$

The coefficient of determination ($R^2$) expresses the proportion of variance in the dependent variable that is explained by the model:

$$R^2 = 1 - \frac{SSE}{SST} \tag{7}$$

where: the sum of squared errors $SSE = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ represents the unexplained portion of variability, and the Total Sum of Squares $SST = \sum_{i=1}^{n}(y_i - \tilde{y})^2$ captures the overall variability of the dependent variable relative to its mean value $\tilde{y}$.

The adjusted coefficient of determination ($R^2_{adj}$) is a modified version of $R^2$ that accounts for the number of predictors in the model:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-j-1} \tag{8}$$

Adjusted $R^2$ provides a more objective assessment of model quality. Unlike the standard $R^2$, which always increases or remains unchanged when additional predictors are included, $R^2_{adj}$ incorporates model complexity and penalizes variables that do not improve explanatory power. This prevents model overfitting and enables an accurate comparison of regression models with differing numbers of predictors. Therefore, it serves as a more reliable criterion for selecting an optimal regression model and assessing the true contribution of predictors to the explained variability of the dependent variable.

## RESULTS AND DISCUSSION

### Assessment of data normality and correlation analysis

Before conducting the correlation and regression analyses, a statistical assessment was performed to evaluate whether the input agrochemical indicators conformed to the normal distribution. Three independent and complementary tests were applied: the Shapiro-Wilk test, the Lilliefors

test, and the Cramér-von Mises test. Their combined use made it possible to comprehensively assess the agreement between the empirical distribution and the theoretical model by considering both distributional shape and deviations occurring in the central and tail regions.

The results consistently indicated statistically significant deviations of all examined variables from normality. The p-values obtained from the Shapiro-Wilk test ranged from $2.29 \cdot 10^{-19}$ to $2.49 \cdot 10^{-6}$, which are far below the critical threshold of $\alpha = 0.05$ and clearly reject the hypothesis of normality. The Lilliefors test produced similarly significant results, with p-values of $p = 0.001$ for all variables. Likewise, the Cramér-von Mises test yielded consistently low p-values in the range of $7.45 \cdot 10^{-9}$ to $2.84 \cdot 10^{-2}$, confirming substantial discrepancies between the empirical and theoretical cumulative distribution functions. None of the applied criteria provided evidence to support the assumption of normality for any of the input variables. This indicates that the data exhibit skewness, heavy tails, or other structural features that contradict the assumption of a normal distribution.

Given the non-normal distributional characteristics of the input variables, the correlation analysis was performed using the Spearman rank correlation coefficient, which does not require assumptions regarding the underlying distribution. The obtained results (Figure 4) reflect both the strength and the direction of monotonic relationships among the agrochemical soil indicators, as well as their statistical significance.

The significance levels on the heatmap in Figure 4 are marked as follows:

- *, $p<0.05$ indicates a statistically significant relationship with an error probability below 5%;
- **, $p<0.01$ denotes a highly significant relationship with a 1% error probability;
- ***, $p<0.001$ reflects an extremely significant relationship with an error probability below 0.1%;
- no marker ($p \geq 0.05$) indicates a non-significant relationship that may be attributed to random variation.

A strong positive association was identified between *pH* and $K^+$ ($p$=0.69***). This relationship suggests that an increase in soil pH is accompanied by higher potassium concentrations, which may be associated with enhanced mobility of this element under more neutral conditions, a pattern observed in the present study. Potassium is less available in acidic soils due to leaching and competition with $H^+$ and $Al^{3+}$ ions (Mengel and Kirkby, 2001).

A stable negative association was found between pH and $Ca^{2+}$ content ($p$=−0.65***), and an

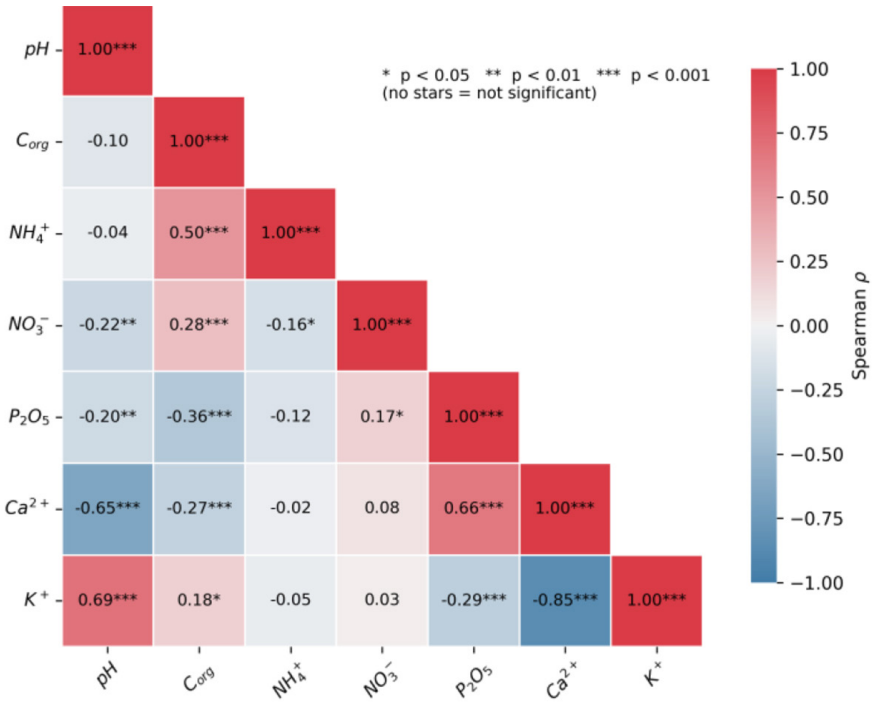


**Figure 4.** Heatmap of spearman correlation coefficients

exceptionally strong negative relationship between $Ca^{2+}$ and $K^{+}$ ($p$=−0.85$^{***}$). This indicates that these elements exhibit opposing trends: rising calcium concentrations coincide with decreasing potassium levels. This effect may result from ion-exchange processes within the soil absorption complex (SAC), potassium leaching under conditions of high calcium content, or competition with $Mg^{2+}$. The strongest associations, particularly between $Ca^{2+}$ i $K^{+}$, further emphasize the robustness of these findings. This pattern is supported by the buffering properties of soils: calcium regulates acidity, creating conditions under which potassium is more effectively retained, and both ions participate in the soil adsorption complex, as they are cations bound to soil colloids (Kuśmierz et al., 2023).

A significant correlation was also observed between $C_{org}$ and $NH_4^+$ ($p$=0.50$^{***}$). Higher organic matter content corresponded to increased ammonium nitrogen concentrations. In most soils, the ratio of organic carbon to total nitrogen (TN) ranges from approximately 10.1 to 12.1 (occasionally 8–15). During humus formation, nitrogen participates in the formation of stable organo-mineral complexes with carbon. Soil microorganisms use carbon as an energy source and nitrogen for protein synthesis (Yang et al., 2024).

Importantly, the p-values confirmed the reliability of the detected relationships: all strongest correlations had $p < 0.001$, minimizing the likelihood of random effects. Weaker coefficients, even when statistically significant ($p < 0.05$), require cautious interpretation because their biological relevance may be limited.

Overall, the correlation analysis provided a comprehensive assessment of the relationships among the agrochemical indicators. The correlation matrices revealed the presence of very strong associations between several soil parameters. Such high correlation levels indicate potential multicollinearity among predictors, which may cause instability in the estimated regression coefficients and complicate model interpretation.

## Evaluation of the MLR model performance

The constructed MLR model demonstrated a high level of explanatory power in predicting the SQI. The coefficient of determination ($R^2$ = 0.904) and the adjusted coefficient (Adjusted $R^2$ = 0.900) indicate that approximately 90% of the variation in SQI is explained by the predictors included in the model. The high F-statistic (F = 277.6, $p < 0.001$) confirms the overall statistical significance of the model. Prediction accuracy was further evaluated using standard error metrics. The obtained values, MSE = 1.85 and MAE = 1.16, indicate a low prediction error, and the MAPE = 5.72% additionally confirms the high accuracy of the model.

To improve interpretability, the errors were also normalized relative to the mean value and the range of SQI variation. The results showed that the Normalized MAE (mean-based) was 5.95%, the Normalized MSE (mean-based) was 0.48%, the Normalized MAE (range-based) was 7.77%, and the Normalized MSE (range-based) was 0.82%. These values confirm that the prediction errors are low both relative to the mean SQI level and to the range of SQI variation. Therefore, the model may be considered preliminarily suitable for soil quality assessment.

The regression coefficients and their statistical significance for all seven predictors are presented in Table 1.

Among the predictors, the most significant positive contributors were organic carbon ($p < 0.001$), calcium ($p < 0.001$), and potassium ($p < 0.001$), all of which increased soil quality. Nitrate nitrogen exhibited a strong statistically significant negative effect (t = -20.0, $p < 0.001$), while the magnitude of this effect was relatively small but stable. This may indicate nitrate over-accumulation in the soil, which can disrupt cation balance and slightly acidify the soil. Ammonium nitrogen and pH also had negative, though less pronounced, effects. In contrast to the other predictors, available phosphorus was statistically insignificant ($p = 0.673$), suggesting no meaningful contribution to SQI variation. Therefore, excluding this predictor is reasonable to improve model stability and avoid unnecessary complexity without reducing explanatory power.

However, even after removing non-significant variables, the model estimates remained unstable due to high multicollinearity. This is confirmed by the large condition number (Cond. No. ≈ 2830). High condition numbers indicate that the predictor matrix is poorly conditioned, which leads to inflated standard errors, reduced reliability of statistical inferences, and difficulties in interpreting the contribution of individual factors. Consequently, further assessment of multicollinearity was performed using variance inflation factors.

**Table 1.** Estimated coefficients of the MLR model with seven predictors

| Predictor | Coefficient | Std. Error | t-value | p-value | CI (2.5%) | CI (97.5%) |
|---|---|---|---|---|---|---|
| Intercept | 11.4347 | 3.934 | 2.907 | 0.004 | 3.673 | 19.196 |
| x1 ($pH$) | –1.2904 | 0.418 | –3.087 | 0.002 | –2.115 | –0.466 |
| x2 ($C_{org}$) | 1.3342 | 0.097 | 13.813 | <0.001 | 1.144 | 1.525 |
| x3 ($NH_4^+$) | –0.1163 | 0.050 | –2.339 | 0.020 | –0.214 | –0.018 |
| x4 ($NO_3^-$) | –0.0732 | 0.004 | –20.044 | <0.001 | –0.080 | –0.066 |
| x5 ($P_2O_5$) | 0.0057 | 0.013 | 0.422 | 0.673 | –0.021 | 0.032 |
| x6 ($Ca^{2+}$) | 1.0367 | 0.115 | 9.053 | <0.001 | 0.811 | 1.263 |
| x7 ($K^+$) | 0.3299 | 0.062 | 5.325 | <0.001 | 0.208 | 0.452 |

## Multicollinearity diagnostics and variance analysis

To quantitatively assess the degree of multicollinearity, an analysis of the VIF was conducted. This approach made it possible to identify variables characterized by excessive mutual correlation and to evaluate their potential impact on the stability and interpretability of the regression model parameters. The obtained VIF values allowed the identification of predictors with elevated multicollinearity and supported preliminary decisions regarding their inclusion in the optimized model. The results are presented in Figure 5.

As it is presented in Figure 5, the VIF values for most variables do not exceed the threshold level (VIF < 4), indicating their relative independence. For $P_2O_5$ (VIF = 5.4), borderline multicollinearity is observed, which requires additional consideration. The most problematic variables are $Ca^{2+}$ (VIF = 42.72) and $K^+$ (VIF = 30.37), whose variation is almost entirely explained by other predictors ($R^2 > 0.96$). This suggests redundancy in the model and indicates a potential substantial increase in the standard errors of the corresponding regression coefficients.

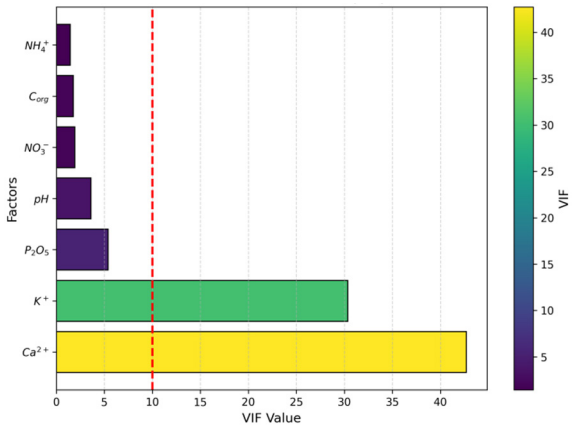


**Figure 5.** Results of the VIF analysis for the seven predictors

Given the identified multicollinearity, an additional assessment of each predictor's contribution to explaining SQI variation was performed. For this purpose, analysis of variance (ANOVA) was applied, providing a quantitative evaluation of the significance of each independent variable. The use of ANOVA made it possible to identify the predictors with the strongest influence on SQI and to develop reasoned recommendations regarding their further inclusion or removal from the regression equation to optimize the model. Table 2 represents the ANOVA results for all seven predictors.

The results indicate that the largest contributions to SQI variation are provided by $pH$, $P_2O_5$ та $Ca^{2+}$, which exhibit extremely high F-statistics and minimal p-values ($p << 0.001$). Significant effects were also observed for $C_{org}$, $NO_3^-$, and $K^+$, whereas $NH_4^+$ did not demonstrate statistical significance ($p = 0.111$).

## Optimization of the regression model structure

The combined results of the regression model with seven predictors, ANOVA, and VIF analysis provide a well-supported basis for optimizing the regression model. The ANOVA results showed that $NH_4^+$ ($p = 0.111$) has no statistically significant effect on soil quality. Therefore, its removal as a non-significant predictor is justified. In addition, the OLS regression results indicated that $P_2O_5$ is also statistically insignificant ($p = 0.673$), further supporting its exclusion to improve model stability and avoid unnecessary complexity. A comparison of $Ca^{2+}$ and $K^+$ demonstrated that

**Table 2.** ANOVA results for seven predictors

| Predictor | Degrees of freedom | Sum of squares | Mean square | F-statistic | p-value |
|---|---|---|---|---|---|
| x1 ($pH$) | 1 | 2242.8535 | 2242.8535 | 1162.2931 | $1.92 \cdot 10^{-81}$ |
| x2 ($C_{org}$) | 1 | 17.8350 | 17.8350 | 9.2425 | 0.002709 |
| x3 ($NH_4^+$) | 1 | 4.9497 | 4.9497 | 2.5651 | 0.110964 |
| x4 ($NO_3^-$) | 1 | 81.8878 | 81.8878 | 42.4359 | $6.78 \cdot 10^{-10}$ |
| x5 ($P_2O_5$) | 1 | 522.7790 | 522.7790 | 270.9149 | $5.17 \cdot 10^{-38}$ |
| x6 ($Ca^{2+}$) | 1 | 415.1927 | 415.1927 | 215.1615 | $9.09 \cdot 10^{-33}$ |
| x7 ($K^+$) | 1 | 52.4203 | 52.4203 | 27.1653 | $4.98 \cdot 10^{-7}$ |
| Residual | 184 | 355.0611 | 1.9297 | - | - |

although both predictors are statistically significant according to ANOVA ($p < 0.001$), they are characterized by critically high VIF values (>30), indicating strong multicollinearity. At the same time, the contribution of $Ca^{2+}$ to explaining SQI variation is substantially greater ($F = 215.16$) than that of $K^+$ ($F = 27.17$). Based on these findings, retaining $Ca^{2+}$ in the model and removing $K^+$ is the most appropriate approach to reduce multicollinearity and enhance the stability of the regression equation.

The coefficients of the optimized multiple linear regression model and their statistical significance for the four retained predictors are presented in Table 3.

The multiple linear regression model for predicting SQI using the four key agrochemical predictors is expressed as:

$$SQI = 16.4685 - 0.8623 \cdot x1 + \\ +1.1442 \cdot x2 - 0.0434 \cdot x4 + 0.6208 \cdot x6 \quad (9)$$

The optimized model demonstrated slightly lower yet still high explanatory power compared with the full model, with $R^2 = 0.856$ and Adjusted $R^2 = 0.853$. This indicates that the model explains approximately 86% of the variation in SQI. The F-statistic (278.9, $p < 0.001$) confirms the overall statistical significance of the model.

The coefficient analysis showed that calcium has a strong and stable positive effect on SQI. The t-value of 16.963 and the narrow confidence interval (0.549; 0.693) indicate high estimation precision and a substantial contribution of this indicator to soil quality formation. Calcium plays a key role in soil buffering capacity and contributes to structural stability. Organic carbon is one of the strongest positive predictors in the model. A t-value of 11.285 and a very narrow confidence interval (0.944; 1.344) confirm the stability and strength of its effect, reflecting its essential role in maintaining soil fertility (Wei et al., 2025). Nitrate nitrogen exhibits a negative but relatively small effect on SQI. An increase in nitrate nitrogen content by 1 mg/kg reduces SQI by only 0.043 points. However, this relationship is statistically significant ($t = -8.770$, $p < 0.001$), indicating the stability of the effect. The confidence interval (-0.053; −0.034) is narrow, which demonstrates that the model estimates this effect with high precision. Excess nitrate nitrogen may reflect excessive application of nitrate-based fertilisers or intensive mineralisation of organic matter. Soil pH exhibits a negative effect, meaning that an increase in pH is associated with a decrease in the SQI value. In the studied area, where podzolized chernozems predominate, the soil reaction is typically slightly acidic or close to neutral (pH 6.0-7.0). In our

**Table 3.** Estimated coefficients of the MLR model with four predictors.

| Predictor | Coefficient | Std. error | t-value | p-value | CI (2.5%) | CI (97.5%) |
|---|---|---|---|---|---|---|
| Intercept | 16.4685 | 3.625 | 4.543 | <0.001 | 9.317 | 23.620 |
| x1 ($pH$) | -0.8623 | 0.434 | -1.986 | 0.048 | -1.719 | -0.006 |
| x2 ($C_{org}$) | 1.1442 | 0.101 | 11.285 | <0.001 | 0.944 | 1.344 |
| x4 ($NO_3^-$) | -0.0434 | 0.005 | -8.770 | <0.001 | -0.053 | -0.034 |
| x6 ($Ca^{2+}$) | 0.6208 | 0.037 | 16.963 | <0.001 | 0.549 | 0.693 |

dataset, only about 30% of the samples fall within this range of soil reaction. The mean pH value was 7.4, which is not entirely characteristic of this soil type. As a result, higher pH values contribute to a reduction in soil quality in the model, as reflected by the regression coefficient (-0.8623). However, the magnitude of this effect is moderate, since a one-unit increase in pH reduces SQI by only about 0.86 points. The p-value (p = 0.048) lies close to the threshold of statistical significance, which is consistent with the confidence interval (-1.719; -0.006) and indicates limited stability in the strength of this effect.

The accuracy of the optimized model was additionally evaluated using prediction error metrics. The values MSE = 2.76 and MAE = 1.51 indicate relatively low deviations between predicted and observed SQI values. The MAPE = 7.97% confirms high percentage-based predictive accuracy. The normalized error metrics – Normalized MAE (mean-based) = 7.69%, Normalized MSE (mean-based) = 0.72%, Normalized MAE (range-based) = 10.04%, and Normalized MSE (range-based) = 1.23% – further demonstrate the stability of predictions both relative to the mean and across the full SQI range.

The VIF analysis has proved that all predictors fall within acceptable limits (VIF < 4), indicating no critical redundancy among variables and confirming model stability (Figure 6).

As it is demonstrated in Figure 6, the VIF values for all variables do not exceed the threshold level (VIF < 4), which indicates their relative independence.

ANOVA revealed the highest contributions from $Ca^{2+}$ (F = 287.7, p << 0.001) and $pH$ (F = 791.1, p << 0.001), while $C_{org}$ (F = 6.29, p = 0.013) and $NO_3^-$ (F = 30.43, p = 0.001) demonstrated a weaker, yet statistically significant effect (Table 4).

Overall, the optimized model demonstrates improved stability and interpretability compared with the full model while retaining strong explanatory

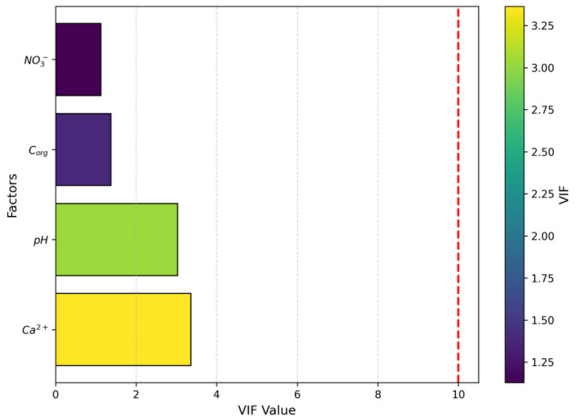


**Figure 6.** Results of the VIF analysis for the four predictors

capability. Based on the obtained results, the key indicators of soil quality within the study area are $C_{org}$, $Ca^{2+}$, $pH$ and $NO_3^-$, whereas $NH_4^+$, $P_2O_5$ and $K^+$ required critical assessment due to low significance ($NH_4^+$, $P_2O_5$) or high multicollinearity ($K^+$).

## Diagnostic assessment of the regression model

To evaluate the validity of the constructed regression model and the reliability of the obtained results, a comprehensive diagnostic assessment of the key statistical assumptions of the OLS method was conducted. Each applied test relies on formal decision rules: if the p-value is below the threshold α = 0.05, the null hypothesis is rejected, indicating that the corresponding assumption has been violated.

The normality of residuals was examined using the Jarque–Bera (GeeksforGeeks, 2024) and Shapiro–Wilk tests. The Jarque–Bera test revealed substantial deviations from the normal distribution (JB = 18.0491, p = 0.0001204), consistent with the observed skewness = 0.1090 and kurtosis = 1.5139,, both of which differ from the theoretical parameters of a normal distribution

**Table 4.** ANOVA results for four predictors

| Predictor | Degrees of freedom | Sum of squares | Mean square | F-statistic | p-value |
|---|---|---|---|---|---|
| x1 ($pH$) | 1 | 2242.8535 | 2242.8535 | 791.0609 | $4.26 \cdot 10^{-69}$ |
| x2 ($C_{org}$) | 1 | 17.8350 | 17.8350 | 6.2904 | 0.012989 |
| x4 ($NO_3^-$) | 1 | 86.2804 | 86.2804 | 30.4313 | $1.14 \cdot 10^{-7}$ |
| x6 ($Ca^{2+}$) | 1 | 815.8188 | 815.8188 | 287.7416 | $1.10 \cdot 10^{-39}$ |
| Residual | 187 | 530.1912 | 2.8352 | - | - |

(skewness = 0, kurtosis = 3). These results indicate a platykurtic distribution, meaning that the residuals are flatter than a normal distribution. The Shapiro–Wilk test results (W = 0.9003, p = $4.796 \cdot 10^{-10}$) further confirmed a significant violation of the normality assumption.

Homoscedasticity was tested using the Breusch–Pagan (Breusch and Pagan, 1979) and White (White, 1980) tests. Both tests indicated the presence of heteroscedasticity. For the Breusch–Pagan test, LM = 38.9433 (p = $7.157 \cdot 10^{-8}$) and F = 11.8949 (p = $1.245 \cdot 10^{-8}$). For the White test, LM = 91.3433 (p = $2.114 \cdot 10^{-13}$) and F = 11.4731 (p = $1.59 \cdot 10^{-18}$). These results show that the variance of the residuals depends on the predictor values, indicating a violation of the homoscedasticity assumption.

To mitigate the influence of the detected violations, an additional evaluation was performed using OLS with HC3 (Long and Ervin, 2000) robust standard errors (Table 5) and robust linear modeling (RLM) with the HuberT estimator (Robust Linear Models – statsmodels, n.d.) (Table 6).

A comparison of the results from the classical OLS regression (Table 2), the model with HC3 robust standard errors (Table 5), and the robust regression model estimated using RLM with the HuberT function (Table 6) demonstrates a high degree of stability in the estimated coefficients. This consistency indicates the structural reliability of the model, even under violations of the classical OLS assumptions, namely non-normality and heteroscedasticity of residuals, both of which were confirmed by the diagnostic tests.

To further evaluate the compliance of the constructed regression model with OLS assumptions, a graphical analysis of the residuals was performed (SixSigma.us, n.d.). Such visual inspection makes it possible to detect potential nonlinearity, heteroscedasticity, and structural patterns that may not be captured by statistical tests alone. Figure 7 presents the residuals versus fitted values plot, which is typically used to assess the presence of heteroscedasticity.

The points are distributed without any noticeable curvature or increasing variance pattern, indicating that the spread of residuals does not systematically depend on the fitted values. Despite the statistically confirmed heteroscedasticity, the visual assessment does not reveal sharply defined irregularities, which is consistent with the conclusion that applying robust standard errors (HC3) effectively compensates for this violation.

To further assess linearity and detect any potential patterns in the residual structure, scatter plots of residuals against each individual predictor were generated (Figure 8).

The analysis of the plots revealed the monotonic trends. The distribution of points appears

**Table 5.** Estimated coefficients of the optimized MLR model with four predictors using OLS with HC3 robust standard errors

| Predictor | Coefficient | Std. Error | t-value | p-value | CI (2.5%) | CI (97.5%) |
|---|---|---|---|---|---|---|
| Intercept | 16.4685 | 3.826 | 4.304 | <0.001 | 8.920 | 24.017 |
| x1 ($pH$) | -0.8623 | 0.453 | -1.904 | 0.049 | -1.756 | -0.003 |
| x2 ($C_{org}$) | 1.1442 | 0.093 | 12.262 | <0.001 | 0.960 | 1.328 |
| x4 ($NO_3^-$) | -0.0434 | 0.004 | -10.203 | <0.001 | -0.052 | -0.035 |
| x6 ($Ca^{2+}$) | 0.6208 | 0.041 | 15.199 | <0.001 | 0.540 | 0.701 |

**Table 6.** Estimated coefficients of the optimized regression model obtained using robust linear modeling (RLM, HuberT)

| Predictor | Coefficient | Std. Error | z-value | p-value | CI (2.5%) | CI (97.5%) |
|---|---|---|---|---|---|---|
| Intercept | 16.4685 | 3.625 | 4.543 | <0.001 | 9.363 | 23.574 |
| x1 ($pH$) | -0.8623 | 0.434 | -1.986 | 0.047 | -1.713 | -0.011 |
| x2 ($C_{org}$) | 1.1442 | 0.101 | 11.285 | <0.001 | 0.945 | 1.343 |
| x4 ($NO_3^-$) | -0.0434 | 0.005 | -8.770 | <0.001 | -0.053 | -0.034 |
| x6 ($Ca^{2+}$) | 0.6208 | 0.037 | 16.963 | <0.001 | 0.549 | 0.692 |

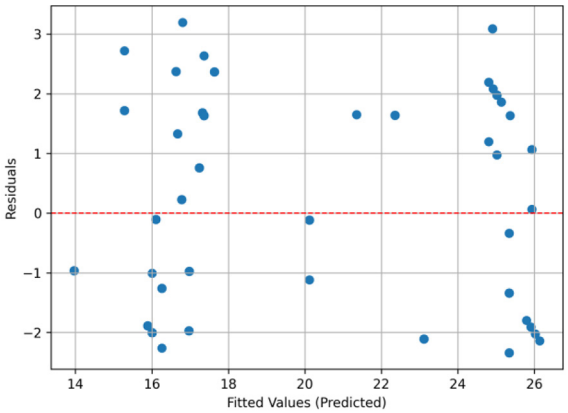


**Figure 7.** Residuals versus fitted values plot for the optimized MLR model

random and does not form systematic shapes such as S-curves or clustered segments, which confirms the absence of systematic residual shifts associated with the value of any individual predictor. This behavior is consistent with the ANOVA results, which indicated independent contributions of each predictor included in the model.

To assess the normality of residuals, a QQ-plot was generated (Figure 9), allowing for a comparison between the empirical quantiles and the theoretical quantiles of a normal distribution.

The concentration of points along the diagonal indicates that the majority of residuals are close to a normal distribution. At the same time, deviations in the tails, expressed as a systematic upward shift of points above the theoretical line on the left side and their downward shift on the right side, confirm the presence of platykurtic behavior in the distribution. This pattern is consistent with the estimated kurtosis value (kurtosis = 1.514). Such tail deviations are typical of residuals that are departed from normality and were also detected by the Jarque–Bera and Shapiro–Wilk tests. Nevertheless, these deviations do not substantially affect the coefficient estimates, as evidenced by the stability of the parameter values in the OLS-HC3 and RLM models.

The results of the graphical diagnostics confirm the absence of critical residual patterns that would call into question the correctness of the model specification. Although the statistical tests indicate certain violations of classical assumptions, the visual assessment demonstrates that the model adequately captures the structure of the data, and the application of HC3 and RLM ensures robust and interpretable parameter estimates.

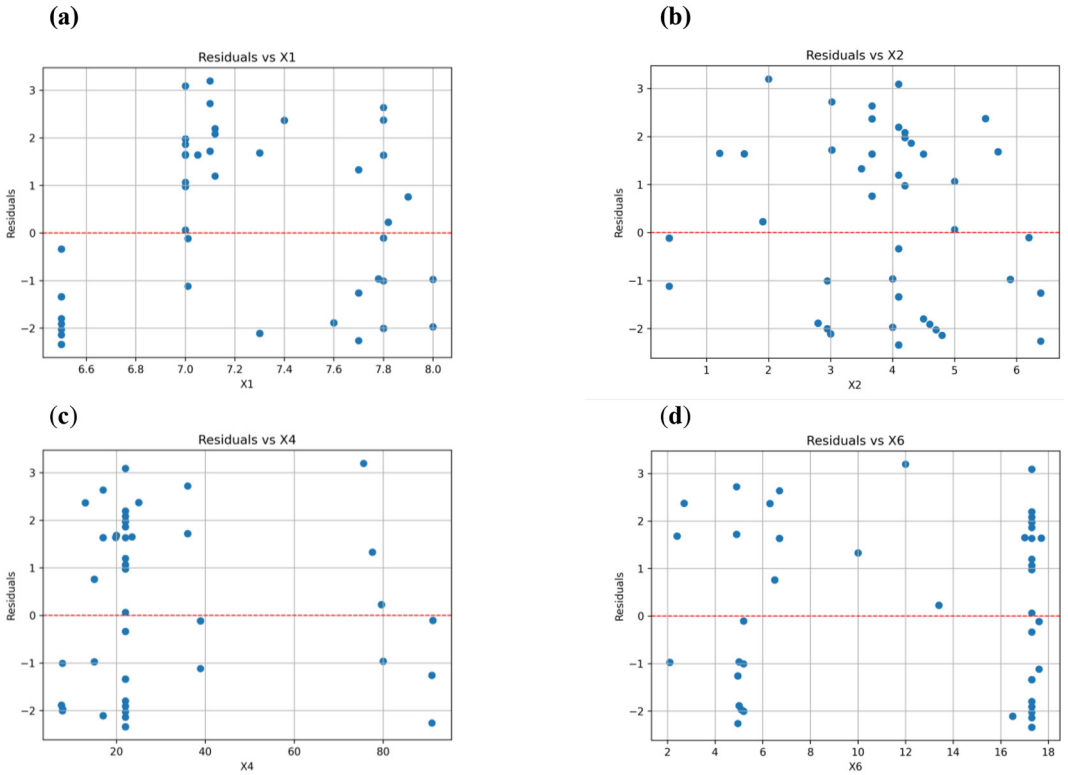


**Figure 8.** Residuals versus individual predictors: (a) x1 ($pH$ ); (b) x2 ($C_{org}$); (c) x4 ($NO_3^-$); (d) x6 ($Ca^{2+}$)

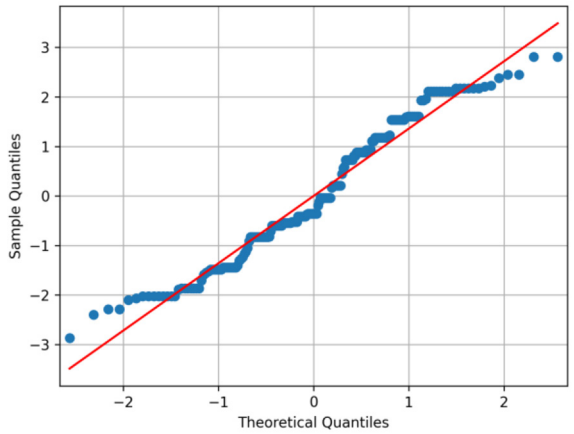


**Figure 9.** Normal Q-Q plot of residuals or the optimized MLR model

## CONCLUSIONS

In this research, the agrochemical properties of soils in the Rossoshanska community of the Khmelnytskyi region were analyzed to assess soil quality through the SQI and to construct a statistically sound predictive model.

The initial regression model with seven predictors demonstrated high explanatory power ($R^2 = 0.904$, Adjusted $R^2 = 0.900$), but exhibited substantial multicollinearity, confirmed by the elevated VIF values for $Ca^{2+}$ and $K^+$. Based on the ANOVA results and the statistical significance of the coefficients, $NH_4^+$ and $P_2O_5$ were excluded as nonsignificant predictors, while $K^+$ was removed due to critical multicollinearity. The optimized model with four predictors ($pH$, $C_{org}$, $NO_3^-$, and $Ca^{2+}$) maintained strong explanatory capability ($R^2 = 0.856$, Adjusted $R^2 = 0.853$), while reducing multicollinearity to an acceptable level (VIF < 4) and preserving parameter stability.

Residual diagnostics (Jarque–Bera, Shapiro–Wilk, Breusch–Pagan, White) revealed deviations from normality and homoscedasticity; however, the application of HC3 robust standard errors and the RLM (HuberT) model confirmed the structural robustness of the estimates. The coefficients remained stable and statistically significant, supporting the reliability of the conclusions. Graphical diagnostics (residuals versus fitted values, residuals versus predictors, QQ-plots) did not reveal any critical patterns or nonlinearities that could undermine the correctness of the model specification.

Overall, the optimized multiple linear regression model is an interpretable and statistically grounded tool for assessing soil quality using SQI. The key indicators of soil condition in the study area are organic carbon, exchangeable calcium, soil acidity, and nitrate nitrogen content. These findings may support the planning of agrochemical interventions, the optimization of fertilization strategies, and the enhancement of agroecosystem sustainability.

## REFERENCES

1. Bachmann, G., Kinzel, H. (1992). Physiological and ecological aspects of the interactions between plant roots and rhizosphere soil. *Soil Biology and Biochemistry*, *24*, 543–552.
2. Bassaco, M. V. M., Motta, A. C. V., Pauletti, V., Prior, S. A., Nisgoski, S., Ferreira, C. F. (2018). Nitrogen, phosphorus, and potassium requirements for Eucalyptus urograndis plantations in southern Brazil. *New Forests*, *49*(5), 681–697. https://doi.org/10.1007/s11056-018-9658-0
3. Bentsen, N., Larsen, S., Stupak, I. (2019). Sustainability governance of the Danish bioeconomy – The case of bioenergy and bio-materials from agriculture. *Energy Sustainability and Society*, *9*, 40. https://doi.org/10.1186/s13705-019-0210-3
4. Bhardwaj, A. K., Jasrotia, P., Hamilton, S. K., Robertson, G. P. (2011). Ecological management of intensively cropped agro-ecosystems improves soil quality with sustained productivity. *Agriculture, Ecosystems & Environment*, *140*(3–4), 419–429. https://doi.org/10.1016/j.agee.2011.01.005
5. Bindraban, P., Stoorvogel, J., Jansen, D., Vlaming, J., Groot, J. (2000). Land quality indicators for sustainable land management: Proposed method for yield gap and soil nutrient balance. *Agricultural Ecosystems & Environment*, *81*, 103–112.
6. Brejda, J., Karlen, D., Smith, J., Allan, D. (2000). Identification of regional soil quality factors and indicators: II. northern Mississippi loess hills and Palouse prairie. *Soil Science Society of America Journal*, *64*(6), 2125–2135.
7. Brejda, J., Moorman, T., Karlen, D., Dao, T. (2000). Identification of regional soil quality factors and indicators: I. Central and southern high plains. *Soil Science Society of America Journal*, *64*(6), 2115–2124.
8. Breusch, T. S., Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287. https://doi.org/10.2307/1911963
9. Camenzind, T., Hättenschwiler, S., Treseder, K. K., Lehmann, A., Rillig, M. C. (2017). Nutrient limitation of soil microbial processes in tropical forests. *Ecological Monographs*, *88*(1), 4–21. https://doi.org/10.1002/ecm.1279

10. Chagnon, P.-L., Brown, C., Stotz, G. C., Cahill, J. F. (2017). Soil biotic quality lacks spatial structure and is positively associated with fertility in a northern grassland. *Journal of Ecology*, *106*(1), 195–206. https://doi.org/10.1111/1365-2745.12844
11. Chaudhury, J., Mandal, U., Sharma, K., Ghosh, H., Mandal, B. (2005). Assessing soil quality under long-term rice-based cropping system. *Communications in Soil Science and Plant Analysis*, *36*, 1–21.
12. DataScientest. (2024, January 19). *Pearson and spearman correlations: A guide to understanding and applying correlation methods*. https://datascientest.com/en/pearson-and-spearman-correlations-a-guide-to-understanding-and-applying-correlation-methods
13. Doran, J. W., Parkin, T. B. (1996). Quantitative indicators of soil quality: A minimum data set. In J. W. Doran & A. J. Jones (Eds.), *Methods for assessing soil quality 49*, 25–37. Soil Science Society of America.
14. DSTU 4115:2002. (2003). Soils. Determination of mobile phosphorus and potassium compounds by the modified Chirikov's method.
15. DSTU 4287:2004. (2005). Soil quality. Sampling.
16. DSTU 4289:2004. (2005). Soil quality. Methods for determination of organic matter.
17. DSTU 4725:2007. (2008). Soil quality. Potassium, ammonium, nitrate and chloride ion activity determination by potentiometric method.
18. DSTU 7861:2015. (2016). Soil quality. determination of exchanges calcium, magnesium, sodium and potassium in soil according to shollenberger in NSC ISSAR named after sokolovsky modification.
19. DSTU EN ISO/IEC 17065. (2016). Issued by the system for environmental certification and ecolabelling.
20. DSTU ISO 10390:2001. (2002). Soil quality. determination of ph.
21. DSTU ISO 11464:2007. (2009). Soil quality. pretreatment of samples for physico-chemical analyses.
22. Du, Z. X. (2016). *The study and the evaluation of the nutrient element content in the liao river estuary wetland soil under the papermaking wastewater irrigation* [Unpublished Master's thesis]. Shenyang Agricultural University.
23. Ebrahimi, M., Sinegani, A., Sarikhani, M., Mohammadi, S. (2017). Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses. *Computers and Electronics in Agriculture*, *140*, 409–421. https://doi.org/10.1016/j.compag.2017.06.019
24. FAO. (2014). *World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps (update 2015)*. World Soil Resources.
25. Forkuor, G., Hounkpatin, O., Welp, G., Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: A comparison of machine learning and multiple linear regression models. *PLoS ONE*, *12*, e0170478. https://doi.org/10.1371/journal.pone.0170478
26. GeeksforGeeks. (2024, April 19). *How to conduct a jarque-bera test in R*. https://www.geeksforgeeks.org/r-machine-learning/how-to-conduct-a-jarque-bera-test-in-r/
27. GeeksforGeeks. (2025, July 29). *Lilliefors test*. https://www.geeksforgeeks.org/data-science/lilliefors-test/
28. Govaerts, B., Sayre, K. D., Deckers, J. (2006). A minimum data set for soil quality assessment of wheat and maize cropping in the highlands of Mexico. *Soil Tillage Research*, *87*(1–2), 163–174. https://doi.org/10.1016/j.still.2005.03.005
29. Grove, S., Parker, I. M., Haubensak, K. A. (2017). Do impacts of an invasive nitrogen-fixing shrub on Douglas-fir and its ectomycorrhizal mutualism change over time following invasion? *Journal of Ecology*, *105*(6), 1687–1697. https://doi.org/10.1111/1365-2745.12764
30. Hinsinger, P. (2001). Bioavailability of soil inorganic P in the rhizosphere as affected by root-induced chemical changes: A review. *Plant and Soil*, *237*, 173–195.
31. John, K., Isong, I., Kebonye, N., Ayito, E., Agyeman, P., Afu, S. (2020). Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land*, *9*(12), 487. https://doi.org/10.3390/land9120487
32. Kawamura, K., Tsujimoto, Y., Rabenarivo, M., Asai, H., Andriamananjara, A., Rakotoson, T. (2017). Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar. *Remote Sensing*, *9*(10), 1081. https://doi.org/10.3390/rs9101081
33. Kuśmierz, S., Skowrońska, M., Tkaczyk, P., Lipiński, W., Mielniczuk, J. (2023). Soil organic carbon and mineral nitrogen contents in soils as affected by their pH, texture and fertilization. *Agronomy*, *13*(1), 267. https://doi.org/10.3390/agronomy13010267
34. Liu, S., An, N., Yang, J., Dong, S., Wang, C., Yin, Y. (2015). Prediction of soil organic matter variability associated with different land use types in mountainous landscape in southwestern Yunnan province, China. *Catena*, *133*, 137–144. https://doi.org/10.1016/j.catena.2015.05.015
35. Long, J. S., Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

36. Madhumathi, R., Arumuganathan, T., Iyer, R., Shruthi, R., Shruthhi, K. (2020). Soil nutrient analysis using machine learning techniques. In *Proceedings of the national e-conference communication, computation, control and automation (CCCA-2020)*. Sri Ramakrishna Engineering College. http://www.ijsred.com/
37. Maksymenko, N. V., Baliuk, S. A., Kucher, A. V., Peresadko, V. A. (2022). Regional differences of soils of Ukraine to assess the cost of ecosystem services. *Ukrainian Geographical Journal*, *2022*(2), 19–31. https://doi.org/10.15407/ugz2022.02.019
38. Malato, G. (n.d.). *An introduction to the shapiro-wilk test for normality*. Built In. https://builtin.com/data-science/shapiro-wilk-test
39. Marschner, H. (2012). *Marschner's mineral nutrition of higher plants* (3rd ed.). Academic Press.
40. Masto, R., Chhonkar, P., Singh, D., Patra, A. (2008). Alternative soil quality indices for evaluating the effect of intensive cropping, fertilisation and manuring for 31 years in the semi-arid soils of India. *Environmental Monitoring and Assessment*, *136*, 419–435.
41. Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M. (2008). A multiple regression approach to assess the spatial distribution of soil organic carbon (SOC) at the regional scale (Flanders, Belgium). *Geoderma*, *143*, 1–13. https://doi.org/10.1016/j.geoderma.2007.08.025
42. Mengel, K., Kirkby, E. (2001). *Principles of plant nutrition* (5th ed.). Kluwer Academic Publishers.
43. Mohr, D. L., Wilson, W. J., Freund, R. J. (2022). Multiple regression. In *Statistical methods* (pp. 351–444). Elsevier. https://doi.org/10.1016/b978-0-12-823043-5.00008-4
44. Murphy, C. J., Baggs, E. M., Morley, N., Wall, D. P., Paterson, E. (2017). Nitrogen availability alters rhizosphere processes mediating soil organic matter mineralisation. *Plant and Soil*, *417*(1-2), 499–510. https://doi.org/10.1007/s11104-017-3275-0
45. *Natsionalnyi atlas ukrainy*. (2009). DNUVP "Kartohrafiia".
46. statsmodels. (n.d.). *Ordinary least squares*. https://www.statsmodels.org/stable/examples/notebooks/generated/ols.html
47. Pichura, V., Potravka, L., Kyrylov, Y., Domaratskiy, Y., Dudiak, N., Skrypchuk, P., Biedunkova, O., Breus, D., Rybak, V., Statnyk, I., Meištininkas, R., Pedišius, N., Žaltauskaitė, J., Dyudyayeva, O., Stroganov, O., Skrypchuk, M., Chata, R., Rutta, O., Biloshkurenko, O. (2023). *Sustainable agriculture in Ukraine*. Baltija Publishing. https://doi.org/10.30525/978-9934-26-359-0
48. Polupan, M., Velychko, V. (2019). *Ukrainian agronomic soil science. part 2: Soil classification, agrosoil potential, ecological zoning*. Agrarna nauka.
49. Prakash, S., Sharma, A., Sahu, S. S. (2018). Soil moisture prediction using machine learning. In *Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICICCT.2018.8473260
50. Rengel, Z., Damon, P. (2008). Crops and genotypes differ in efficiency of kalium uptake and use. *Physiologia Plantarum*, *133*, 624–636.
51. Robertson, G. P., Groffman, P. M. (2015). Nitrogen transformations. In E. A. Paul (Ed.), *Soil microbiology, ecology and biochemistry* (pp. 421–446). Academic Press.
52. Roudier, P., Malone, B., Hedley, C., Minasny, B., McBratney, A. (2017). Comparison of regression methods for spatial downscaling of soil organic carbon stocks maps. *Computers and Electronics in Agriculture*, *142*(A), 91–100.
53. Schipper, L., Sparling, G. (2000). Performance of soil condition indicators across taxonomic groups and land uses. *Soil Science Society of America Journal*, *64*, 300–311.
54. Scikit-learn. (n.d.). *Metrics and scoring: Quantifying the quality of predictions*. https://scikit-learn.org/stable/modules/model_evaluation.html#model-evaluation
55. Shchesniak, A. O., Bosak, P. V., Kovalchuk, N. P., Sokolov, S. O. (2025). Environmental safety assessment of soils in Khmelnytskyi region based on chemical composition and acidity analysis. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, (1), 63–69. https://doi.org/10.33271/nvngu/2025-1/063
56. Shukla, M. K., Lal, R., Ebinger, M. (2006). Determining soil quality indicators by factor analysis. *Soil Tillage Research*, *87*, 194–204. https://doi.org/10.1016/j.still.2005.03.017
57. Singer, M., Ewing, S. (1998). Soil quality. In M. Sumner (Ed.), *Handbook of soil science* (pp. G271—G278). CRC Press.
58. SixSigma.us. (n.d.). *Everything to know about residual analysis*. https://www.6sigma.us/six-sigma-in-focus/residual-analysis/
59. Sojka, R., Upchurch, D. (1999). Reservations regarding the soil quality concept. *Soil Science Society of America Journal*, *63*(5), 1039–1054.
60. Ståhle, L., Wold, S. (1989). Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems*, *6*(4), 259–272. https://doi.org/10.1016/0169-7439(89)80095-4
61. statsmodels. (n.d.). *Robust linear models – statsmodels 0.14.4*. https://www.statsmodels.org/stable/rlm.html
62. Team, T. I. (2010, August 14). *Variance inflation factor (VIF): Definition and formula*. https://www.investopedia.com/terms/v/variance-inflation-factor.asp

63. Tsentralnyi Instytut Ahronimichnoho Obsluhovuvannia (TsINAO). (1994). *Metodychni vkazivky z ahronimichnoi otsinky vmistu osnovnykh elementiv mineralnoho kharchuvannia roslyn u gruntakh Ukrainy*.
64. Vacca, A., Aru, F., Ollesch, G. (2016). Short-term impact of coppice management on soil in a quercus ilex L. stand of sardinia. *Land Degradation & Development*, *28*(2), 553–565. https://doi.org/10.1002/ldr.2551
65. Velasquez, E., Lavelle, P., Andrade, M. (2007). GISQ, a multifunctional indicator of soil quality. *Soil Biology and Biochemistry*, *39*, 3066–3080.
66. Wander, M., Bollero, G. (1999). Soil quality assessment of tillage impact in Illinois. *Soil Science Society of America Journal*, *63*, 961–971.
67. Wei, X., Others. (2025). Enhancing soil fertility and organic carbon stability with high-nitrogen biogas slurry: Benefits and environmental risks. *J. Environ. Manage.*, *384*, 125584. https://doi.org/10.1016/j.jenvman.2025.125584
68. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817. https://doi.org/10.2307/1912934
69. Yang, M., Zhou, D., Hang, H., Chen, S., Liu, H., Su, J., Lv, H., Jia, H., Zhao, G. (2024). Effects of balancing exchangeable cations Ca, Mg, and K on the growth of tomato seedlings (*Solanum Lycopersicum* L.) based on increased soil cation exchange capacity. *Agronomy*, *14*(3), 629. https://doi.org/10.3390/agronomy14030629
70. Yu, T., Hou, W., Hou, Q., Ma, W., Xia, X., Li, Y., Yan, B., Yang, Z. (2020). Safe utilization and zoning on natural selenium-rich land resources: A case study of the typical area in Enshi County, China. *Environmental Geochemistry and Health*, *42*, 2803–2818. https://doi.org/10.1007/s10653-019-00487-w
71. Zhang, D., Xu, M. (2024). A high-dimensional Cramér–von mises test. *Mathematics*, *12*(22), 3467. https://doi.org/10.3390/math12223467
72. Zhang, D., Zhang, W., Huang, W., Hong, Z., Meng, L. (2017). Upscaling of surface soil moisture using a deep learning model with VIIRS RDR. *ISPRS International Journal of Geo-Information*, *6*(5), 130. https://doi.org/10.3390/ijgi6050130
73. Zhao, Z., Chow, T., Rees, H., Yang, Q., Xing, Z., Meng, F. (2009). Predict soil texture distributions using an artificial neural network model. *Computers and Electronics in Agriculture*, *65*(1), 36–48.