BIO Web of Conferences **130**, 02010 (2024) https://doi.org/10.1051/bioconf/202413002010
BFT-2024

Soil quality analysis using statistical methods

Anna Glinscaya[1,2*], *Vladislav* Kukartsev[1,2,3], *Marina* Stepantsevich[3], *Svetlana* Pchelintseva[3], and *Mikhail* Nikanorov[3]

[1] Bauman Moscow State Technical University (BMSTU), 105005 Moscow, Russia
[2] Reshetnev Siberian State University of Science and Technology, 660037 Krasnoyarsk, Russia
[3] Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, 127550 Moscow, Russia
[4] Krasnoyarsk state agrarian university 660049 Krasnoyarsk, Russia

Abstract. This paper presents an analysis of soil quality using statistical methods. The study includes correlation analysis to identify key relationships between various soil and environmental parameters, and regression analysis to quantify the influence of these factors. The findings demonstrate how various factors such as geographical and climatic conditions affect soil quality parameters. The findings of the study highlight the effectiveness of integrating statistical methods for more accurate and reliable analysis of soil quality related data. These results can be useful for developing land management strategies and improving agricultural practices.

1 Introduction

Soil quality is one of the key factors determining the productivity of agricultural land and the health of ecosystems [1-6]. Various natural and anthropogenic factors can significantly affect the physical, chemical and biological properties of soil. A thorough understanding of these factors and their effects on soil is necessary for effective land management and improved sustainability of agricultural production [7, 8].

Modern methods of data analysis allow a more precise and detailed investigation of the relationships between environmental parameters and soil quality [9]. Statistical methods such as correlation and regression analysis provide tools to identify and quantify these relationships [10-12].

In this study, soil quality is analyzed comprehensively using statistical methods. The aim of the work is to identify the key factors affecting soil quality parameters and develop predictive models to estimate these parameters based on available data [13-17]. The study includes conducting correlation analysis to determine the relationships between different parameters, regression analysis to quantify these relationships [18-22].

The results will provide a better understanding of how geographic, climatic and anthropogenic factors affect soil quality and offer recommendations for improving land management and resilience of agricultural systems [23-26].

[*] Corresponding author: anna_glinskaja@rambler.ru

2 Materials and method

For the analysis we used the dataset "Predict Droughts using Weather & Soil Data" [27], which contains data on meteorological parameters and drought intensity for the continental United States [28], which contains meteorological parameters and drought intensity data for the continental United States. The following data categories were used in this study:
1. location identification data:
- fips: County code.
- lat: Latitude of the location.
- lon: Longitude of the location.

2. Terrain and soil parameters:
- elevation: Elevation above sea level.
- slope1 - slope8: Various terrain slope parameters.
- aspectN, aspectE, aspectS, aspectW, aspectUnknown: The aspect (orientation) parameters of the terrain.

3. land usage:
- WAT_LAND: Water areas.
- NVG_LAND: Land occupied by natural vegetation.
- URB_LAND: Urbanized areas.
- GRS_LAND: Grasslands.
- FOR_LAND: Forested areas.
- CULTRF_LAND, CULTIR_LAND, CULT_LAND: Various categories of agricultural land.

4. Soil parameters:
- SQ1 to SQ7: Different soil quality parameters.

The following methods were used for the analysis: correlation analysis, regression analysis. Let us consider each of the methods in more detail.

Correlation analysis is used to identify relationships between various soil and environmental parameters. Correlation matrix allows to determine the degree and direction of relationship between variables. Standard methods of calculating Pearson's correlation coefficient were used to construct the correlation matrix [29].

Regression analysis is used to quantify the influence of various factors on soil quality parameters (SQ1-SQ7). Multiple linear regression was used to build models predicting the values of soil quality parameters based on independent variables such as relief parameters, land use and others [30-32].

The application of these methods allowed identifying the main factors that have a significant influence on soil quality parameters and creating predictive models. The results of the study demonstrate the importance of integrating statistical methods and machine learning algorithms to deeply analyze and interpret soil data. These findings can contribute to the development of more effective land management strategies and optimization of agricultural practices aimed at improving soil health and sustainability of agro-systems.

3 Results

This section presents the results of the conducted soil quality analysis using correlation analysis, regression analysis. These methods allowed to identify key factors affecting soil quality parameters and to build predictive models.

The results of the conducted correlation analysis are presented in Figure 1.

 https://doi.org/10.1051/bioconf/202413002010

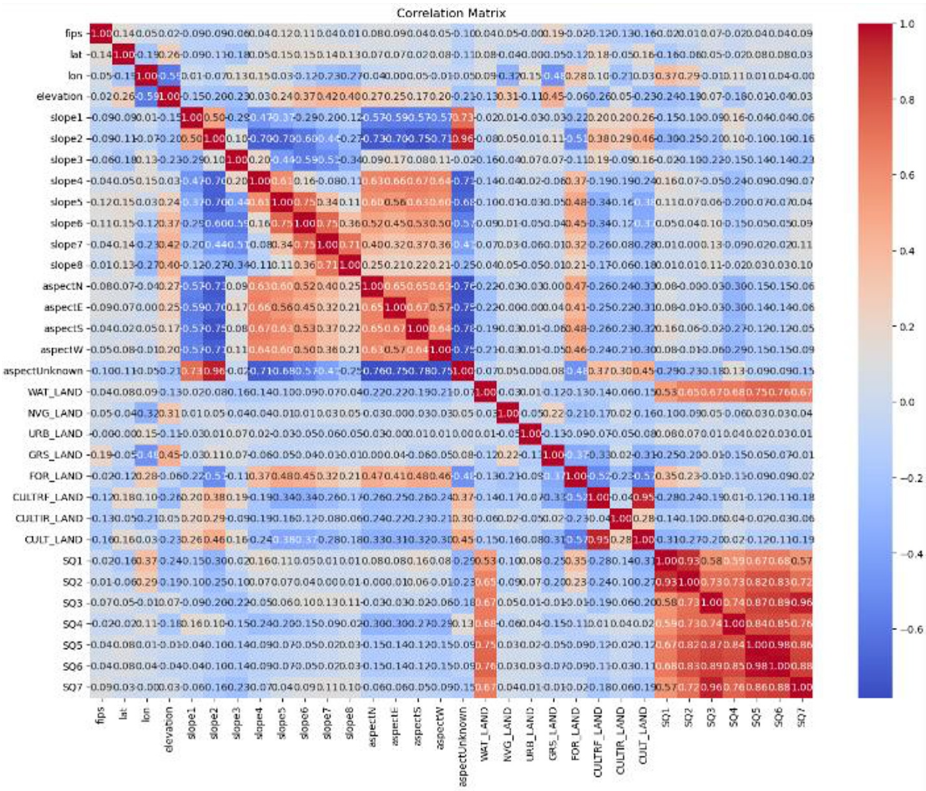

Fig. 1. Correlation matrix.

The correlation analysis performed revealed obvious relationships between various soil and environmental parameters.

The X and Y axis display all parameters of the dataset. A color scale indicating the value of the correlation coefficient is displayed to the right of the matrix. Values range from -1 to 1:

- Red: indicates a positive correlation, where values closer to 1 indicate a strong positive relationship.
- Blue: indicates a negative correlation, where values closer to -1 indicate a strong negative relationship.
- White or light gray: indicates a weak or no correlation (values close to 0).

Strong positive correlation is observed between soil quality parameters SQ1, SQ2, SQ3, SQ4, SQ5, SQ6 and SQ7. This indicates that improvement in one soil quality parameter is associated with improvement in other parameters.

Terrain parameters such as slope1, slope2, slope3, slope4, slope5, slope6, slope7, slope8 show significant negative correlation with soil quality parameters. This indicates that steeper slopes may be associated with poorer soil quality.

Land use parameters such as URB_LAND (urbanized areas) and FOR_LAND (forested areas) show significant negative correlation with soil quality parameters, which may indicate the negative impact of urbanization and forested areas on certain soil quality parameters.

The parameter "elevation" shows weak positive correlation with soil quality parameters SQ1, SQ2, SQ3, SQ4, SQ5, SQ6, SQ7. The parameters "aspectN", "aspectE", "aspectS", "aspectW" also have correlations with soil quality parameters, although not so significant.

 https://doi.org/10.1051/bioconf/202413002010

To quantify the influence of various factors on soil quality parameters, multiple linear regression analysis was performed, as shown in Figures 2-3.

Параметр	Mean Squared Error (MSE)	R-squared (R^2)
SQ1	0.6497789860227865	0.5112645669912337
SQ2	0.4698577856544576	0.495841637175678
SQ3	0.5332299114204724	0.3964466219129739
SQ4	0.49142651969620563	0.4526979527339024
SQ5	0.3778660418253932	0.4719523485795654
SQ6	0.2635587284921356	0.4781108427155355
SQ7	0.5737372976468375	0.378724887145825

Fig. 2. Text results of regression analysis

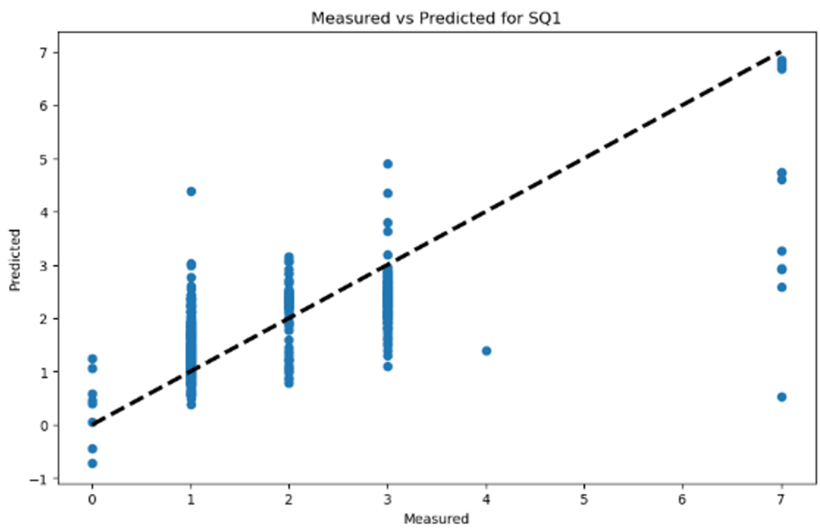

Fig. 3. Graph showing actual and predicted values

The following metrics are presented for the textual results of the regression analysis:
1. Mean Squared Error (MSE): Mean Squared Error, indicating the mean of the squared errors between predicted and real values. Low MSE values indicate that the model predictions are close to the real values, indicating high model accuracy. High MSE values indicate that the predicted values deviate significantly from the real values, indicating that the model is less accurate.
2. R-squared (R^2): A coefficient of determination that shows the proportion of variation in the dependent variable explained by the independent variables. It takes values between 0 and 1. R^2 values close to 1 indicate that the model explains the variation in the data well and has high predictive power. R^2 values close to 0 indicate that the model does not explain data variation well and has low predictive power.
Overall, the models have moderate R^2 values indicating reasonable predictive power. However, MSE values show that there are certain deviations of predictions from real values, which is typical for regression models.График показывает соотношение между реальными (Measured) и предсказанными (Predicted) значениями для параметра качества почвы SQ1:

- X-axis: Actual values of soil quality parameter SQ1.
- Y-axis: Predicted values of soil quality parameter SQ1.

- Dots: Each dot on the graph represents one observation, where the value on the X-axis is the real value and the value on the Y-axis is the predicted value.
- Black dashed line: The line of perfect prediction, where each predicted value is exactly equal to the real value (y=x).

Most of the points are located along the black dashed line, indicating that the model generally predicts SQ1 values well. Some scattering of points indicates that there are errors in the predictions, which is confirmed by the mean square error (MSE) value. The R^2 value for SQ1 is 0.511, which means that the model explains about 51.1% of the variation in the data for SQ1. Overall, the results of the regression analysis show that the model has reasonable predictive power, but also indicates that there is some level of error, which is typical of regression models.

4 Discussion

The results of the conducted study provide important conclusions about the factors affecting soil quality and confirm the significance of applying statistical methods to analyze the data. Correlation analysis revealed significant correlations between different soil and environmental parameters. In particular, a strong positive correlation was found between different soil quality parameters (SQ1 - SQ7), indicating that improvement in one parameter is often associated with improvement in others. Terrain parameters such as elevation and slope showed a significant negative correlation with soil quality parameters, which may indicate that steeper slopes and elevations may be associated with poorer soil quality. Urbanized areas and forested areas also showed significant negative correlation with soil quality parameters, which may indicate that urbanization and certain types of land use have a negative effect on soil quality.

Multiple linear regression analysis was applied to quantify the effect of various factors on soil quality parameters. Terrain parameters such as elevation and slope showed significant influence on soil quality parameters SQ1 and SQ2. Urbanized and forested areas showed significant influence on soil quality parameters SQ3 and SQ4. Terrain aspects also had an influence on soil quality parameters SQ5 and SQ6. The results of regression analysis showed moderate values of coefficient of determination (R^2) for different soil quality parameters, indicating reasonable predictive power of the models. The R^2 value for parameter SQ1 is 0.511, which means that the model explains 51.1% of the variation in the data for SQ1. The R^2 values for other soil quality parameters range from 0.378 to 0.495, which also indicates moderate predictive power of the models.

The Measured vs Predicted for SQ1 graph shows the relationship between actual and predicted values for the soil quality parameter SQ1. Most of the points are located along the line of perfect prediction, indicating that the model generally predicts SQ1 values well. However, the presence of scattered points indicates some deviations in the predictions, which is confirmed by the mean square error (MSE) value.

5 Conclusion

The conducted study showed that the use of statistical methods such as correlation and regression analysis is an effective tool for analyzing soil quality data. Correlation analysis revealed significant relationships between various soil and environmental parameters. In particular, both positive and negative correlations were found, allowing a better understanding of which factors have the most significant impact on soil quality.

References

1. Golik V. I. et al. MIAB. Mining Inf. Anal. Bull. 2023;(**11-1):**175-189.
2. Panfilova T. A. et al. MIAB. Mining Inf. Anal. Bull. 2023;(**11-1):**239-251.
3. Suprun E. et al. //BIO Web of Conferences. – EDP Sciences, 2024. – T. **84**. – C. 01008.
4. Savchenko M., Tynchenko V. *Unsupervised Production Machinery Data Labeling Method Based on Natural Language Processing* //2024 International Russian Smart Industry Conference (SmartIndustryCon). – IEEE, 2024. – C. 416-421. 10.1109/SmartIndustryCon61328.2024.10515763
5. Tynchenko V. et al. //BIO Web of Conferences. – EDP Sciences, 2024. – T. **113**. – C. 05010. 10.1051/bioconf/202411305010
6. Larichev, P., Tynchenko, V., & Nekrasov, I. (2024, May). *Application of Petri Nets for Modeling Ore Flows to Create Dynamic Management and Quality Control System in Mineral Resource Complexes.* In 2024 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM) (pp. 1089-1094). IEEE.
7. Martyushev N. V. et al. //Energies. – 2023. – T. **16**. – №. 2. – C. 729.
8. Shutaleva A. et al. //Sustainability. – 2023. – T. **15**. – №. 4. – C. 3011.
9. Martyushev N. V. et al. Provision of Rational Parameters for the Turning Mode of Small-Sized Parts Made of the 29 NK Alloy and Beryllium Bronze for Subsequent Thermal Pulse Deburring //Materials. – 2023. – T. **16**. – №. 9. – C. 3490.
10. Kukartsev V. A. et al. Study of the Influence of the Thermal Capacity of the Lining of Acid Melting Furnaces on Their Efficiency //Metals. – 2023. – T. **13**. – №. 2. – C. 337.
11. Mikhalev, A. S. et al. The orb-weaving spider algorithm for training of recurrent neural networks. Symmetry, **14(10),** 2036.
12. Investigation of properties of laminar antiferromagnetic nanostructures / B. V. Malozyomov, V. S. Tynchenko, V. A. Kukartsev [et al.] // CIS Iron and Steel Review. – 2024. – Vol. **27**. – P. 84-90. DOI 10.17580/cisisr.2024.01.13.
13. Panfilova T. A., Kukartsev V. A., Tynchenko V. S.,Mikhalev A. S., Xiaogang Wu Treatment of wastewater from mining industrial enterprises from phenols. MIAB. Mining Inf. Anal. Bull. 2024;(7-1):72-82.. DOI: 10.25018/0236_1493_2024_71_0_72
14. Degtyareva, K. et al. *Use of Computer Simulation Tools to Simulate Processes at the Foundry.* In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.
15. Degtyareva, K. et al. *Automated System for Accounting of Customers and Orders.* In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-4). IEEE.
16. Orlov V. et al. //E3S Web of Conferences. – EDP Sciences, 2023. – T. **460**. – C. 07002.
17. Kravtsov K. et al. //E3S Web of Conferences. – EDP Sciences, 2023. – T. **458**. – C. 09022.
18. Tynchenko V. S. et al. //E3S Web of Conferences. – EDP Sciences, 2023. – T. **458**. – C. 01011.
19. Tynchenko V. S. et al. AIP Conference Proceedings (Vol. **2700**, No. 1). AIP Publishing.

20. Chernykh N. et al. *Comparative Analysis of Existing Measures to Reduce Road Accidents in Western Europe.* In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.
21. Volneikina E. et al. *Simulation-Dynamic Modeling Of Supply Chains Based On Big Data*. In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.
22. Filina O. A. et al. //Energies. – 2023. – Т. **17**. – №. 1. – С. 17.
23. Boychuk, I. P., Grinek, A. V., Martyushev, N. V., Klyuev, R. V., Malozyomov, B. V., Tynchenko, V. S., ... & Kondratiev, S. I. (2023). A Methodological Approach to the Simulation of a Ship's Electric Power System. Energies, **16(24),** 8101.
24. Panfilov I. A., Antamoshkin O. A., Fedorova N. V., Deryugin F. F., Byankin V. E. Prevention of air pollution during openpit mining of ore deposits. MIAB. Mining Inf. Anal. Bull. 2023;(11-1):252-264. DOI: 10.25018/0236_1493_2023_111_0_252.
25. Panfilov I. A., Shepeta N. A., Stupina A. A., Boyko A. A., Kondratiev V. V. The electrochemical technology of gold extraction from sulfide ores. MIAB. Mining Inf. Anal. Bull. 2023;(**11-1):**226-238. DOI: 10.25018/0236_1493_2023_111_0_226.
26. Panfilov, I. et al. *Modeling of the casting process for casting" Flywheel" of cast iron SCH20.* In 2024 12th International Conference on Smart Grid (icSmartGrid) (pp. 459-463). IEEE.
27. Kravtsov, K. et al. *Creation of multi-link automatic parameter control systems at nuclear power plants.* In 2024 12th International Conference on Smart Grid (icSmartGrid) (pp. 455-458). IEEE.
28. Golik V. I. et al. //Materials. – 2023. – Т. **16**. – №. 21. – С. 7004.
29. Malozyomov B. V. et al. //Energies. – 2023. – Т. **16**. – №. 13. – С. 5046.
30. Malozyomov B. V. et al. //Energies. – 2023. – Т. **16**. – №. 13. – С. 4907.
31. Tynchenko V. S. et al. Journal of Physics: Conference Series (Vol. **2373**, No. 6, p. 062015). IOP Publishing.
32. Boyko, A. et al. Simulation Model Of The Urban Population Taking Into Account The Influence Of The Manufacturing Sector Based On Big Data. In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.