Using the principal component analysis to assess some properties of agro-gray soil and bonitet

A Ruchkina¹, R Ushakov¹* and O Zaharova¹

¹Ryazan State Agrotechnological University Named after P.A. Kostychev, 1 Str. Kostychev, Ryazan 390044, RF

*E-mail: r.ushakov1971@mail.ru

Abstract. To understand all the structural complexity of soil fertility, one can use the principal component analysis of factor analysis. Factor analysis, with all its shortcomings to identify structural relationships between variables, can be used to understand the complexity of soil fertility, by some assumptions. It can be estimated through bonitet. The work was carried out on the basis of materials from an agrochemical survey of agro-gray soils. Exchangeable acidity, mobile forms of phosphorus and potassium, humus, the sum of exchangeable bases and bonitet calculated in accordance with them were analyzed. The sample number was 68. Group 0 included bonitet with values less than 90 units and group 1 had bonitet with values more than 90 units. It was found that the contribution of exchangeable acidity and phosphorus to the first main component, which accounted for about 39% of the total dispersion, was maximum and equal to 0.7 and 0.84 units, respectively. Humus and potassium were positively associated with the second component (29% of dispersion) and their contribution was 0.7 and 0.81 units, respectively.

1. Introduction

Most of the processes occurring in the soil are interconnected. They are reflected in soil properties, between which both direct and indirect relationships can be traced. Bonitet is used for the integral expression of properties and for evaluating the properties. Thus, the bonitet changes under the influence of various factors. Factor analysis (FA) allows to identify these indicators, analyze them and study the degree of influence. FA investigates the structure of correlation or covariance matrices.

Formalization of knowledge about the bonitet of soils and the role (place) of soil properties in its formation, presented in the form of structural links, is essential to understand soil fertility issues, to solve practical problems of assessing the quality of land. One of the methods for solving these issues is the principal component analysis, which is aimed not only at extracting the necessary information from all its diversity, but also at determining the structural relationships. When presented graphically, there is a transition from the original feature coordinate system to the orthogonal coordinate system of the main components.

The essence of the method is to search for a hyperplane of a given dimension in the original space. The hyperplane is chosen, having the minimal data design error in the sense of the sum of squares of deviations [1, 2]. This method finds its application for processing information on soils [3, 4, 5], other objects [6, 7]

The purpose of the research is to determine the structure of the relationship between soil parameters and bonitet.
2. Materials and methods

The work was carried out based on the materials of agrochemical survey on agro-gray soils.

The object of research was the data on some agrochemical indicators (variables) of agro-gray medium loamy soil (exchangeable acidity, mobile forms of phosphorus and potassium, humus and exchangeable bases) and bonitet calculated in accordance with them. The sample number was 68. Group 0 included bonitet with values less than 90 units and group 1 had bonitet with values more than 90 units. To determine the structure of relationships between soil parameters and bonitet, i.e. the classification, the principal component analysis was used (PCA - orthogonal linear transformation). Thus, the bonitet was characterized by five variables. The meaning of the principal component analysis is that a certain proportion of the total dispersion of the original data (load) is associated with each principal component. Dispersion, being a measure of data variability, can reflect their informative value. It is based on correlation matrices. Indicators related to one vector factor are correlated with each other. Projections on the factorial plane indicate some positive correlation in one direction, a negative correlation of indicators in the opposite direction and their independence in the perpendicular direction. Software product STATISTICA was used as a toolkit.

3. Results and discussion

One of the conditions for conducting factor analysis is the independence of variables and their normal distribution. In this case, the analysis involved independent variables: exchangeable acidity, mobile forms of phosphorus and potassium, humus, the sum of exchangeable bases.

Table 1 shows that the data on humus and bonitet do not obey the law of normal distribution, since multimodality is observed. Most often (9 times) there is a very high value for phosphorus - 254 mg/kg (mode), with significantly lower but high values of the average (198.2 mg/kg) and median (203.5 mg/kg). For potassium, the mode (100 mg/kg) corresponds to the middle class of supply, the average value (135 mg/kg) and median (124 mg/kg) correspond to the increased supply. This testifies to their contrasting spatial variability at the farm. This can also be judged by the frequency distribution of the data. So, in 36.8% of cases, the value of mobile phosphorus is in the range from 240 to 260 mg/kg. Hence the corresponding unreasonable overestimated bonitet values. In 54% of cases they are in the range from 95 to 100 units. This cannot correspond to the true state of affairs, because, for example, the humus content in 50% of cases is below 3% with such bonitet values.

From a statistical point of view, this may reflect the absence of normal distribution, which is a condition for performing some statistical analyses, in particular factorial, since it is based on correlations. Therefore, a check for the normal distribution of data is given below.

Any normal distribution is strictly symmetric about its center, therefore, its asymmetry (As) is equal to zero. Asymmetry characterizes the bias measure of the histogram. Table 1 presents data on asymmetry, which is left-sided and right-sided. Data for potassium (As = 0.93), humus (As = 1.56) and slightly less for pH (As = 0.41) are biased to the right (positive value) while those for bonitet and phosphorus are biased to the left and values in both cases are not close to zero. In general, the data is asymmetric.

| Parameter | Average | Median | Mode | Min value | Max value | Standard deviation | Coefficient of variation | Standard error | Asymmetry |
|-----------|---------|--------|------|-----------|-----------|-------------------|------------------------|--------------|-----------|
| pH        | 3.2     | 3.1    | 3.1  | 4.3       | 6.1       | 0.43              | 8.2                    | 0.052        | 0.41      |
| P2O5      | 198.2   | 203.5  | 254  | 77.0      | 257.0     | 53.07             | 26.8                   | 6.435        | -0.49     |
| K2O       | 135.1   | 124.0  | 100  | 65.0      | 251.0     | 46.39             | 34.3                   | 5.625        | 0.93      |
| Humus     | 3.2     | 3.1    | Several modes | 2.2  | 5.6  | 0.64              | 20.1                   | 0.078        | 1.56      |
| Sum of bases | 17.8   | 19.6  | 21.7  | 11.4      | 26.3      | 4.57              | 25.7                   | 0.555        | 0.04      |
| Bonitet   | 87.9    | 90.6   | Several modes | 54.7 | 99.7 | 10.53             | 12.0                   | 1.276        | -1.21     |
Since factor analysis is based on a correlation matrix, it is necessary to check the condition for the normal distribution of data. The Kolmogorov-Smirnov test and chi-square were used for the assessment. According to these criteria, only the pH data have a normal distribution, since the actual level of significance (p) is greater than the theoretical one (0.05). For other soil parameters, the normal distribution is absent (although according to Kolmogorov-Smirnov, the hypothesis of the normal distribution can be accepted), because the chi-square value of the significance level is much less than the theoretical one. Considering the histogram of the frequency distribution, an adjustment was made based on differentiated (categorical) distribution diagrams. According to the value the data array is divided as it follows: 245 mg/kg for mobile phosphorus, 215 mg/kg for mobile potassium, 4% for humus and 15 mg eq/100 g for the sum of bases. The results are shown in Figure 1. The Shapiro-Wilk criterion (SW-W) was used. For phosphorus, potassium and humus, the values, respectively, up to 245 mg/kg, 215 mg/kg are distributed normally (p is more than 0.05). In terms of the amount of exchange bases, it is performed both before and after 15 mg eq/100 g. Due to the large asymmetry (-1.21) for bonitet, there are no conditions for normal distribution.
Figure 1. Distribution histograms.

It is believed that it is possible to use for statistical processing the initial data of soil variables up to the specified values. This is not critical for phosphorus and potassium, because very high values, which do not fit into the general trend of the dynamics of the state of soil fertility of agro-gray soils in the RF nonblack earth zone, are cut out. An increase in humus in the noted soils up to 4% or more can be considered unlikely.

To build a more accurate model of structural relationships between soil parameters, it was decided to remove the sum of exchange bases from the analysis.

Factor analysis is based on understanding that relationships between variables should be approximately linear, or at least not clearly curvilinear. Therefore, first it is necessary to check the presence or absence of correlation. Calculation of the correlation matrix established the presence of
some positive relationship between bonitet and pH, mobile forms of phosphorus and potassium. There is no connection with humus (Table 2). Positive relationships are noted between pH and mobile phosphorus, and negative relationships are found between the latter and humus. This is due to the intensive use of phosphorus and potash fertilizers on the farm, against the background of which it is impossible to identify the relationship between nutrients and humus. There is no connection between some properties.

Table 2. Correlation matrix.

| Value        | pH    | P₂O₅  | K₂O  | Humus | Sum of bases |
|--------------|-------|-------|------|-------|--------------|
| pH           | 1.0   | 0.4   | 0.1  | -0.1  | 0.5          |
| P₂O₅         | 0.4   | 1.0   | 0.2  | -0.3  | 0.6          |
| K₂O          | 0.1   | 0.2   | 1.0  | 0.2   | 0.7          |
| Humus        | -0.1  | -0.3  | 0.2  | 1.0   | -0.1         |
| Sum of bases | 0.5   | 0.6   | 0.7  | -0.1  | 1.0          |

At first approximation, it can be seen that the calculation of the bonitet is based mainly on nutrient data. The role of humus is diminished and / or hidden. More detailed analysis is needed.

The analysis of correlations so far only allows to assume that there are two relatively independent factors (two types of factors) reflected in the correlation matrix: one relates to mineral nutrition and acidity, and the other one is connected with the organic part of the soil or humus.

It is known that the extraction of factors consists in the selection of interacting variables, whose cross-correlation determines the largest share of the total dispersion. These variables form the first factor. Table 3 presents the extracted factors. The first two factors reflect the maximum variability and have the value of more than 1, which together explain about 68% of the dispersion (Table 3). The third factor has the value of about 1, and its share in the dispersion is about 19%.

Table 3. Eigenvalue and contribution of factors (components) to the dispersion.

| Factor | Eigenvalue | %, of total dispersion | Cumulative eigenvalue (CE) | Resulting, % |
|--------|------------|------------------------|---------------------------|--------------|
| 1      | 1.6        | 38.8                   | 1.6                       | 38.8         |
| 2      | 1.2        | 29.2                   | 2.7                       | 68.0         |
| 3      | 0.8        | 18.9                   | 3.5                       | 86.9         |
| 4      | 0.5        | 13.1                   | 4.0                       | 100.0        |

Table 4 shows that the load or contribution (correlation coefficient) of exchangeable acidity and mobile forms of phosphorus to the first main component is maximum and approximately the same (0.7 and 0.84 units). Consequently, they determine the spread of bonitet along the abscissa axis (projection of the horizons onto the first main component, Figure 2).

Humus and potassium are positively associated with the second component (29% of the dispersion) and their contribution is 0.7 and 0.81 units, respectively. The third and fourth factors have eigenvalues of 0.83 and 0.62 units, respectively.

Table 4. Factor loads (critical less than 0.7).

| Parameter | Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|-----------|----------|----------|----------|----------|
| pH        | 0.7      | 0.2      | 0.6      | 0.2      |
| P₂O₅      | 0.8      | -0.0     | -0.2     | -0.5     |
| K₂O       | 0.3      | 0.8      | -0.4     | 0.3      |
| Humus     | -0.5     | 0.7      | 0.3      | -0.4     |
4. Conclusion
Exchangeable acidity, mobile forms of phosphorus and potassium, humus are of equal importance for bonitet, therefore, for soil fertility. Factor analysis, with all its shortcomings, in identifying structural relationships between variables, under certain assumptions, can be used to understand the complexity of soil fertility. It can be estimated through bonitet. In this case, exchangeable acidity and phosphorus play a decisive, but not exclusive role in the formation of bonitet. This is followed by potassium and humus. Despite the increased and high average values of potassium and phosphorus, the humus content is not low for agro-gray soils and their combinations in the soil can be considered not always optimal. In quantitative terms, the soil parameters are not in optimal proportions with each other, that does not correspond to the agricultural law of equivalence.

References
[1] Kolemaev V A 2014 Econometrics: textbook for university students in specialty 061800 “Mat. methods in economics” (Moscow, INFRA-M)
[2] Volkov D A and Navoev V V 2014 Overcoming the problem of multicollinearity in solving problems of multidimensional calibration in expert systems. Theoretical and practical problems of the development of modern science. Collection of materials of the 6th international scientific and practical conference pp 36-38
[3] Fedorov B A 2007 Bulletin of Tver State Technical University 12 pp 148-150
[4] Rozhkov V A 2013 Bulletin of Soil Science Institute Named after V. V. Dokuchaev 72 pp 47-64
[5] Lipatov D N, Scheglov, A I, Manakhov D V, Zavgorodnyaya Yu A, Rozanova M S and Brekhov P T 2017 Soil Science 7 pp 874-885
[6] Clayden N J, Lehnert R J and Turnock S 1997 Analytica Chimica Acta 344, 3 pp 261-269
[7] Bakulina G, Fedoskin V, Pikushina M., Kukhar V and Kot E 2020 International Journal of Circuits, Systems and Signal Processing 14 pp 232-240