The application of factor analysis to determine the functional dependencies of hydrological characteristics in aquatic ecosystems

S V Sarkisov, I I Pivovarova*, VA Vakunenkov, V I Musatov and A A Sorokin

Military Institute (engineering) of the Military Academy logistics support named after Army General A. V. Hrulev, 22 Zakharyevskaya, St. Petersburg, 196121, Russian Federation

*E-mail: inivpivovarova@gmail.com

Abstract. The article discusses methods of factor analysis to identify heterogeneity in the conditions of spatial distribution and flow formation in aquatic ecosystems. Various factor analysis algorithms were studied to take into account the conjugate action of many factors. Examples of ranking the established dependencies by the value of their own significance criteria are shown. Deficiencies of factor analysis methods in certain software applications were identified because of a possible incorrect interpretation of the results due to the complexity of the technology. The possibility of using factor analysis tools to describe the spatial structures of hydrological data and identify risk zones associated with the instability of the hydrological regime was identified.

1. Introduction
A comprehensive assessment of the characteristic factors of the hydrological regime of aquatic ecosystems can be carried out using factor analysis. Recently, factor analysis algorithms have been widely used for processing hydrochemical and hydrophysical data arrays [1-3]. Multivariate statistical analysis is a convenient tool necessary for a better understanding of the nature of flow formation, analysis of the long-term variability of hydroecological indicators of ground and underground waters [4]. Factor analysis helps to determine spatial changes associated with climatic features and water quality, which is necessary to improve the management of ground and surface waters [5].

This type of analysis allows to solve the following problems: to describe the subject of measurement compactly and at the same time comprehensively. Using factor analysis, it is possible to identify the factors responsible for the presence of linear statistical correlations between the observed variables. There are four main objectives of factor analysis:

- finding hidden, but objectively existing patterns that are determined by the impact of internal and external causes on the studied process;
- compression of information by describing the process using common factors or main components, the number of which is significantly less than the number of initially taken attributes;
identification and study of statistical relationship of attributes with factors or main components [6];
allocation of homogeneous spatio-temporal data structures.

2. Materials and methods
There are different methods for selecting factors from a data set. The chosen method will depend on the sample size, the number of variables and most importantly, of course, the current task.

Principal component analysis (PCA). As the most common form of factor analysis, PCA searches for a linear combination of variables that extracts maximum variance from variables. Then it removes this variance and searches for a second linear combination that explains the maximum proportion of the remaining variance, and so on. This is called the Principal axis method and results in orthogonal (non-correlating) factors [7].

Maximum likelihood factorization (MLF) relies on a linear combination of variables for the formation of factors, where parametric estimates are the ones that most likely lead to the observed correlation matrix using MLF methods (i.e. Maximum likelihood estimation) and allowing multidimensional data normality. Correlations are weighted according to the uniqueness of each variable. MLF implements a Chi-square quality assurance test. We can increase the number of factors by one each time, until the method shows a satisfactory quality of compliance.

The work examined data on 37 hydrological stations evenly distributed over the territory of the Oka River basin. A factor analysis was performed on a set of 6 attributes (latitude, longitude, drain coefficient, modulus of flow, coefficient of variation, frequency of occurrence of hydrological hazards) using the two described above methods in the SPSS Statistics software application.

It should be noted that the results of factorization are often difficult to interpret. Therefore, in order to solve the problem of distribution of variables by factors, all methods of factor analysis provide for the rotation of factors relative to attributes [8]. Rotation helps to present the results in a more convenient way and usually facilitates the interpretation of factors. Rotation does not affect the sum of the eigenvalues, but it changes the factor loadings and the eigenvalues (and the percentage of explained variance) of specific factors.

The most common rotation methods:

- Varimax - criterion – the index of complexity of each factor, which is proportional to the number of variables associated with this factor. The “Varimax” method maximizes the spread of load squares for each factor, which leads to an increase in large and a decrease in small values of factor loadings. As a result, a simple structure is obtained for each factor separately;
- Quartimax - criterion – factor complexity of the variable, proportional to the number of factors associated with it;
- Equimax – a rotation method that combines Varimax, simplifying factors, and Quartimax, simplifying variables. The number of variables with large factor loadings and the number of factors required to explain the variable are minimized.

When performing the analysis, it should be noted that in SPSS Statistics the absence of rotation is set by default, but it is still better to choose some method of rotation, Varimax is used in the work. The original (non-rotating solution), for example, in Principal component analysis maximizes the sum of the squares of the factor loadings, creating a set of factors that explain the maximum amount of variance in the original variables. This explained quantity is reflected in the sum of the eigenvalues of all factors. Such decisions are difficult to interpret because variables tend to be loaded on many factors.

3. Results and discussion
Thus, the procedure of factor analysis in the work consisted of four main stages:

- Calculation of the correlation matrix for all variables involved in the analysis.
- The extraction of factors.
- The choice of factors and rotation of factors for the creation of simplified structure.
- Interpretation and analysis of results.

### 3.1 «Principal component» method

Hydrological data were processed using the "Principal component" method, the "Varimax" rotation, in all six parameters. The initially obtained matrices confirmed the assumption of compliance of the source data with the criteria of factor analysis (there are clear correlation dependences). The tables showed the suitability of the available data for factor analysis in general (table 1, table 2). The results of the KMO test led to the conclusion that the constructed factor model can be used to describe the data structure. Bartlett’s test also showed a satisfactory result. In theory, the significance of Bartlett’s test tests the hypothesis of the absence of functional relationships between variables. If this test is positive (variables are not correlated), factor analysis should be considered unacceptable and other statistical methods should be used. The statistics that determine the suitability of the factor analysis on the Bartlett’s test is significance. At an acceptable level of significance (below 0.05) factor analysis is considered acceptable for the analysis of the sample. In our case, the test showed the significance equal to 0.000, which implies the conclusion about the possibility of using factor analysis to solve the problem.

#### Table 1. Correlation matrix.

|               | Latitude | Longitude | Modulus of flow | Coefficient of variation | Drain coefficient | Hydrological hazards |
|---------------|----------|-----------|-----------------|------------------------|------------------|---------------------|
| Latitude      | 1.000    | 0.159     | 0.648           | -0.223                 | 0.822            | 0.004               |
| Longitude     | 0.159    | 1.000     | -0.419          | -0.271                 | 0.153            | -0.199              |
| Modulus of flow | 0.648   | -0.419    | 1.000           | -0.194                 | 0.470            | 0.120               |
| Coefficient of variation | -0.223 | -0.271 | -0.194          | 1.000                  | -0.335           | 0.270               |
| Drain coefficient | 0.822 | 0.153    | 0.470           | -0.335                 | 1.000            | -0.066              |
| Hydrological hazards | 0.004 | -0.199    | 0.120           | 0.270                  | -0.056           | 1.000               |

#### Table 2. KMO and Bartlett’s test.

| Kaiser-Meyer-Olkin Measure of Sampling Adequacy | 0.473 |
|------------------------------------------------|-------|
| Bartlett’s Test of Sphericity                  |       |
| Approx. Chi-Square (Approximately)             | 95.270|
| Df. (Degrees of freedom)                       | 15    |
| Sig. (Significance)                            | 0.000 |

The following tables and graph made it possible to determine how many factors should be determined and which variables should be loaded on them (table 3, table 4).

#### Table 3. Component matrix.

|               | Factor 1 | Factor 2 | Factor 3 |
|---------------|----------|----------|----------|
| Latitude      | 0.935    | 0.141    | 0.152    |
| Drain coefficient | 0.892    | -0.071   | 0.021    |
| Modulus of flow  | 0.608    | 0.583    | -0.171   |
Table 4. Rotated component matrix.

|                    | Factor 1 | Factor 2 | Factor 3 |
|--------------------|----------|----------|----------|
| Latitude           | 0.949    | -0.027   | 0.125    |
| Drain coefficient  | 0.845    | -0.257   | 0.142    |
| Modulus of flow    | 0.720    | 0.197    | -0.426   |
| Hydrological hazards | 0.118    | 0.826    | -0.037   |
| Coefficient of variation | -0.243   | 0.752    | -0.119   |
| Longitude          | 0.73     | -0.091   | 0.953    |

In the methodology of factor analysis, the graph of eigenvalues shows the dependence of the eigenvalues of factors on their numbers in the allocation order. There are various ways to determine the number of factors for further factor analysis [9]. In the work, Cattell’s scree test was used. The scree graph depicts the components as the X-axis, and the corresponding eigenvalues as the Y-axis. When moving to the right on this graph, the eigenvalues initially sharply fall. When this drop stops and the curve makes a bend towards a less sharp drop, Cattell’s scree test suggests discarding all further components after the bend starts [10]. The graph of eigenvalues (figure 1) allowed to take the number of factors equal to three.

As a result, after setting the required number of factors, determining their loading variables and carrying out the necessary calculation algorithms, numerical indicators of the correspondence of hydrological stations to each of the three factors are obtained. Data was transferred to the map and processed in the GIS application (Global Mapper GIS) (figure 2). The result is the zoning of the river basin (in our case, three regions) by the nature of the formation of the hydrological regime. We were
able to identify the cumulative dependencies of the parameters and their influence on the overall picture of the distribution of hydrological characteristics in the basin. And most importantly, to identify risk zones associated with the possibility of the occurrence of dangerous hydrological hazards. In our case, it is the southwestern region and the section of the central part of the catchment (factor 1).

![Spatial distribution of factors: Factor 1, Factor 2, Factor 3.](image)

**Figure 2.** Spatial distribution of factors: ▲ Factor 1 ▶ Factor 2 ◆ Factor 3.

### 3.2 The «Maximum likelihood» method

The source data were also processed using the maximum likelihood method. However, the method did not give the expected results. The distribution of factors could not reflect the real heterogeneity in the distribution of hydrological indicators. If, using the Principal components method, three main regions were clearly identified, reflecting local features of flow characteristics, in this case, the patterns of hydrological regime formation were practically not revealed. This is apparently due to the fact that in large samples even a very small improvement in the explained variance may turn out to be significant when testing for compliance quality. There is also a general drawback of factor analysis — alternative rotations can explain the same variance (have the same total eigenvalue), but lead to different factor loadings.

### 4. Conclusion

In the course of the work, a factor analysis was carried out, implemented by two methods: the “Principal components” method and the “Maximum likelihood” method. For a better interpretation of the obtained factors, each method was used for a set of hydrological data with a georeferenced location, processed over a ten-year period. During the analysis, factors characterizing the local conditions for the formation of river flow were identified. The possibility of constructing an information model of flow in the study area with the allocation of zones of maximum risk is shown.

Factor analysis methods allow, with some approximation, to solve one of the most common tasks of scientific research - the task of a compact description of the phenomenon based on processing large information arrays. Factors combine in one group correlated variables, and if they can be interpreted, then in the case of an ecosystem, this means that they can be directly or indirectly related to some specific source of chemical substances entering the watercourse or the process by which they are combined. Therefore, it is advisable to use the described technologies to study the mechanisms of formation of not only hydrological components, but also the quality of river waters. And due to the fact that at present very often the anthropogenic factor acts as a structurally forming element in the formation of the water regime along with natural geochemical and biological processes, it is concluded that it is promising to use the described technology to analyze hydrochemical data in the ecosystem.
References

[1] Ohwoghere-Asuma O, Aweto K E and Ugbe C F 2019 Lithofacies Identification and Multivariate Analysis of Groundwater Chemistry in Coastal Aquifers in Koko Area of the Western Niger Delta *Hydrology* 6(2) 31

[2] Qian J et al 2016 Multivariate statistical analysis of water chemistry in evaluating groundwater, geochemical evolution and aquifer connectivity near a large coalmine, Anhui, China *Environ Earth Sci.* 75 747

[3] Ashley R P and Lloyd J W 1978 An example of the use of factor analysis and cluster analysis in groundwater chemistry interpretation. *J. Hydrol* 39 441–4

[4] Pivovarova I I 2019 Cluster analysis for structuring spatial data *International Journal of Mechanical Engineering and Technology* 10(2) 1085-109

[5] Kuzmin V and Pivovarova I 2018 Monitoring and long-term probabilistic assessment of hydrological risks in the conditions of climate change and variable anthropogenic burden on river basins *International Journal of Mechanical Engineering and Technology* 9(11) 1652-61

[6] Silveira L. 1997 Multivariate analysis in hydrology: the factor correspondence analysis method applied to annual rainfall data *Hydrological sciences journal-Journal des sciences hydrologiques* 42(2) 215-24

[7] Jackson J E 1991 *User’s Guide of Principal Component* (New York: Wiley-Interscience) p 587

[8] O’Connor B P 2000 SPSS and SAS programs for determining the number of components using parallel analysis and Velicer’s MAP test *Behavior Research Methods, Instrumentation, and Computers* 32 396-402

[9] Lu G and Wong D 2008 An adaptive inverse-distance weighting spatial interpolation technique, *Computer&Geosciences* 34 1044–55

[10] Lance C E Butts M M and Michels L C 2006 The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods* 9(2) 202–20