RESEARCH PAPER

STATISTICAL METHOD FOR REDUCING THE NUMBER OF CLIMATIC PREDICTORS IN SPECIES DISTRIBUTION MODELING

Igor O. Popov[1,2*], Elena N. Popova[2]

[1]Yu. A. Israel Institute of Global Climate and Ecology, Glebovskaya str., 20B, Moscow,107258, Russia

[2]Institute of Geography, Russian Academy of Sciences, Staromonetniy pereulok, 29/4, Moscow,119017, Russia

***Corresponding author:** igor_o_popov@mail.ru**

Received: November 16th 2024 / Accepted: July 21st 2025 / Published: October 1st 2025

https://doi.org/10.24057/2071-9388-2025-3734

ABSTRACT. Nineteen bioclimatic parameters from BIOCLIM are widely used in Species Distribution Modeling (SDM). To improve modeling quality, it is essential to reduce the number of parameters. Several approaches have been proposed to solve this challenge, but each has its own limitations. In this study, we aimed to develop an effective statistical method based on identifying correlation groups of parameters and selecting the least correlated ones. Several statistical techniques were used to ensure a reliable parameter selection: simple correlation matrix analysis, cluster analysis (HDBSCAN), and factor analysis (varimax and quartimax). As an example, bioclimatic parameter values for the period 1991–2020 were analyzed for the whole globe. The results obtained using different methods show good consistency. Several correlation groups were identified, ranging from four to five, depending on the interpretation of the negative correlations. One group of two parameters, BIO14 and BIO17, can also be identified based on the results of the varimax factor analysis, although this correlation group was not identified by other methods. Finally, six bioclimatic parameters were selected (BIO2, BIO5, BIO7, BIO14, BIO15, and BIO18), one from each group that demonstrated the minimum average value of the correlation coefficient with parameters from other groups. The average correlation between the selected parameters was significantly lower than in the case of using previously applied methods with the same number of selected parameters.

KEYWORDS: species distribution modeling, data dimension, cluster analysis, factor analysis, HDBSCAN

CITATION: I. O. Popov, E. N. Popova (2025). Statistical Method For Reducing The Number Of Climatic Predictors In Species Distribution Modeling. Geography, Environment, Sustainability, 3 (18), 19-31
https://doi.org/10.24057/2071-9388-2025-3734

ACKNOWLEDGEMENTS: The studies were supported by the grant of the Ministry of Science and Higher Education of Russian Federation (agreement № 075-15-2024-554 of 24.04.2024).

Conflict of interests: The authors reported no potential conflict of interests.

INTRODUCTION

Living organisms, as open systems, are affected by the environment. Climatic factors, particularly ambient temperature, are the most significant abiotic factors determining the existence and reproduction of individuals and populations. For terrestrial organisms, humidity is also an important factor (Bonan 2008; Schimel 2013). Climate change has various effects on land and marine ecosystems, including their structure, species composition, and relationships between components. The most significant issue is the impact of climate and climate change on species distribution, including shifts in their ranges (McCarty 2001; Gilman et al. 2010; Post 2013).

The assessment of potential changes in species distribution, particularly those important for economic activity and human health, presents a significant challenge for modern science. Currently, the main methodological approach to this issue is Species Distribution Modeling (SDM), which is a rapidly evolving field at the intersection of ecology, biogeography, applied climatology, and information technology (Franklin 2009; Peterson et al. 2011; Araújo et al. 2019; Srivastava et al. 2019). Various algorithms are used to construct these models, including general-purpose machine learning techniques such as support vector machines, logistic regression, and neural networks, as well as specialized methods designed for habitat modeling, the most commonly used of which is MaxEnt (Phillips et al. 2004; Phillips et al. 2006).

Although a wide variety of environmental factors, both abiotic and biotic, can be used as predictors of species distribution in these models, climate variables play a major role in almost all models, as they have a fundamental limiting effect on organism ranges (Popova and Popov 2013; Popova and Popov 2019). Obviously, it is possible to design a huge, if not infinite, number of such variables. However, not all variables will correlate well with distribution data or be significant for range formation, and not all will be convenient for projecting models to other regions of the world.

In 1984, BIOCLIM was proposed as one of the first methods for constructing Species Distribution Models (Nix 1986; Busby 1991). This software package included a set of 12 climatic parameters, specifically designed to be

biologically significant for most species and suitable for projecting models across hemispheres. The package was developed by a group of Australian scientists and was initially used to assess the invasive potential of different species. In 1996, a new version of this software package was presented, with the number of bioclimatic parameters increased to 19 (Booth 2018). Their list is given in Table 1. The names of these parameters begin with the prefix BIO, followed by a number from 1 to 19 (BIO1-BIO19).

As shown in Table 1, the first 11 parameters (BIO1-BIO11) are related to temperature, while the remaining 8 (BIO12-BIO19) reflect a precipitation regime. There is no specific mention of a particular month or season. Instead, periods of the year with the highest or lowest temperatures, or the highest or lowest precipitation, are used. This makes it easy to move models between regions with different annual climatic variation, like hemispheres. In addition, four parameters (BIO8, BIO9, BIO18, and BIO19) are "mixed", reflecting the values of climatic factors of one type over a period determined by factors of another type. Such an arrangement can be useful for modeling the ranges of certain species, but it can also cause some problems in certain cases. For instance, they can have a very high gradient of spatial variability in some regions, particularly in equatorial and tropical areas. Some researchers recommend avoiding the use of these parameters or using them with extreme caution (Booth 2022).

The design of the BIOCLIM parameters has been so successful that they are widely used in SDM and other areas of ecological modeling. This set was further popularized with the release of the WorldClim database in 2005 and its second version in 2017[1]. This database contains values for six continents and is interpolated onto a spatial grid with a step of up to 30" (Hijmans et al. 2005; Fick and Hijmans 2017). According to the study (Bradie and Leunig 2017), the BIOCLIM parameters have been used significantly more often than other climate variables in the modeling of nearly 1900 species in about 2000 publications.

However, using a large number of potential predictors has several disadvantages. First, it introduces a challenge known as the "curse of dimensionality" in machine learning. As the number of independent variables increases, so does the distance between samples in feature space. That can result in inaccuracies in the classification of virtual space (Hastie et al. 2009) and lead to overfitting of models, when a model that fits too well to the training data classifies new data with a high error rate. Additionally, a large number of variables can significantly increase the computational load, especially when analyzing large amounts of data.

In addition to the above-mentioned problems, climate variables have a fairly strong correlation between each other, which can also influence the performance of several algorithms (for instance, in the case of MaxEnt). Furthermore, when it is necessary to assess the predictor significance for classification, which in SDM may be linked to their biological significance for a particular species, the presence of strongly correlated variables may lead to an inaccurate assessment of their significance, especially when using ensemble techniques based on decision trees such as "random forest" or gradient boosting.

One possible approach to reducing the number of predictors is to create new variables based on linear or non-linear combinations of the original variables. These new variables should retain as much information as

Table 1. Bioclimatic parameters

BIO1	annual mean temperature
BIO2	mean diurnal range (mean of monthly (max temp - min temp))
BIO3	isothermality (BIO2/BIO7) (×100)
BIO4	temperature seasonality (standard deviation ×100)
BIO5	max temperature of warmest month
BIO6	min temperature of coldest month
BIO7	temperature annual range (BIO5-BIO6)
BIO8	mean temperature of wettest quarter
BIO9	mean temperature of driest quarter
BIO10	mean temperature of warmest quarter
BIO11	mean temperature of coldest quarter
BIO12	annual precipitation
BIO13	precipitation of wettest month
BIO14	precipitation of driest month
BIO15	precipitation seasonality (coefficient of variation)
BIO16	precipitation of wettest quarter
BIO17	precipitation of driest quarter
BIO18	precipitation of warmest quarter
BIO19	precipitation of coldest quarter

[1] *https://www.worldclim.org*

possible while being significantly smaller in number. Common methods for such reduction include various versions of Principal Component Analysis (PCA), Locally-Linear Embedding (LLE) and Multidimensional Scaling (MDS), among others (Roweis and Saul 2000). In particular, the study (Dinnage 2023) used a neural network Variable Autoencoder (VAE) to reduce the set of WorldClim variables to 5 synthetic variables without significant information loss. These synthetic variables are nonlinear combinations of the original 19 parameters. However, the disadvantage of this approach is that the obtained variables are artificial. It complicates a biological interpretation of the results.

An alternative approach is to identify correlation groups of the actual variables, i.e., groups with a higher correlation within than between them. From these groups, we can select variables that either have the lowest correlation with the other groups or are particularly significant for a specific study. Typically, this approach eliminates variables that demonstrate a high level of correlation with each other; for example, if the value of a correlation coefficient is above a certain threshold (Bellard et al. 2013; Petrosyan et al. 2023; Zhang et al. 2023). However, such simultaneous pairwise reduction may result in the loss of several important variables since a variable that is highly correlated with one or more variables may also be weakly correlated with other variables. In addition, the choice of a selection threshold is not always clear.

As an alternative to the strategies described, we propose using statistical methods to identify correlation groups. This approach involves using algorithms that allow for the identification of fine structures and groups in data based on various types of relationships between its elements. For this purpose, we used a modern, highly effective clustering algorithm called HDBSCAN. Two methods of factor analysis, varimax and quartimax, were also used as an alternative approach to verifying the clustering results. These three algorithms were used for the first time to solve this problem.

After identifying the correlation groups, our approach involves selecting one parameter from each group with the least mean correlation to parameters from other groups. The identification of correlation groups allows us to determine the optimal number of selected parameters. This number balances the minimization of the correlation between parameters with their minimum sufficient quantity.

The aim of this study was to evaluate the effectiveness of the proposed approach to reducing the number of SDM predictors using 19 bioclimatic parameters calculated for the entire globe as an example.

MATERIALS AND METHODS

Climate data

The climate data source used in this study was the CRU TS 4.05 database (Harris et al. 2020), which contains the results of meteorological observations with a monthly resolution, interpolated onto a regular spatial grid with a step of 0.5°. This database is widely used in SDM. In particular, it forms the basis for the popular bioclimatic database WorldClim, which was discussed in the introduction. The fact that CRU is based on meteorological observations affords it several advantages over reanalysis, such as ERA5. Many studies have found that reanalysis often produces erroneous results, especially with respect to precipitation data, which is of special importance for SDM (Purnadurga et al. 2019; Bodjrènou et al. 2025; Fatolahzadeh et al. 2024).

In total, this grid contains 67,420 nodes with values, as the nodes over the seas, oceans, and Antarctica do not have climate variables' values. Nineteen bioclimatic parameters were calculated according to their description in Table 1 for the entire globe, using temperature variables and monthly precipitation amounts. These values were averaged over the period 1991-2020 for each node in the spatial grid.

As a result of the calculations, each of the 67,420 spatial nodes was characterized by 19 bioclimatic parameters. Based on this data, linear correlation coefficients were calculated for each pair of parameters to form a correlation matrix with a size of 19×19.

All calculations in this work were performed using the Python 3 programming language. A Python 3 module for the calculation of bioclimatic parameters is available in the repository[2]. Jupyter notebooks containing the calculations and some additional materials are available in the repository[3].

Cluster analysis

To identify correlation groups among bioclimatic parameters, cluster analysis was used. This method allows the identification of groups of objects (in this study, sets of bioclimatic parameter values) that are closer together than other objects. In other words, it helps to detect areas of increased density in the space of objects. Cluster analysis can use different metrics to measure the distance between objects. In this study we used metrics based on the linear correlation coefficient to measure the distance between the values of bioclimatic parameters. This allows us to determine groups of parameters that have a higher correlation with each other than with other parameters.

Currently, there are many methods of cluster analysis (Wierzchoń and Kłopotek 2018). In this work, we used the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm, which is an evolution of the DBSCAN and OPTICS methods (Campello et al. 2013; McInnes and Healy 2017). A special feature of this method is that it can independently determine the number of clusters and identify noise points – samples that do not belong to any cluster and can be considered as single-sized clusters. Furthermore, it does not require access to the original data but only a matrix of distances between the analyzed samples.

In its modern form, the HDBSCAN algorithm includes several stages of data processing:

1. Transformation of the original sample space to better select areas of increased density, using the method described and justified in the paper (Eldridge et al. 2015).
2. Construction of a graph where the vertices are the samples, and the edge weights are equal to the distance between the samples. The graph is then transformed into a minimum spanning tree, which is a graph where each vertex has at least one connection to other vertices, and the total weight of all the edges is minimized.
3. Construction of a hierarchical cluster tree based on the obtained minimum connected tree.
4. Transformation of the hierarchical cluster tree into a flat cluster system. At this stage, both user-defined hyperparameters (minimum cluster size and ε – minimum allowable distance between clusters) and several parameters calculated directly from the data are used. This distinguishes the HDBSCAN method from DBSCAN, which only identifies clusters based on the specified hyperparameters.

[2]https://doi.org/10.5281/zenodo.13913422
[3]https://doi.org/10.5281/zenodo.13970876

When analyzing a small number of samples, as in this study, it is recommended to set the minimum cluster size to 2. In this case, ε becomes the only hyperparameter that needs to be optimized to find the optimal value that provides the best quality of cluster selections. (Malzer and Baum 2020).

The distance between bioclimatic parameters was determined using two different metrics. These metrics differ in their assessment of negative correlations. Negative correlation, like positive correlation, implies the presence and duplication of information about one variable in another variable, albeit in a different sense. This type of correlation can also negatively affect the quality of the modeling.

The first metric, d1, considers negative correlation values as an indicator of a greater distance between parameters. It is calculated using the Eq. 1:

$$d_1 = 1 - r \tag{1}$$

where r is the linear correlation coefficient.

This metric ranges from 0 (for parameters with a perfect positive correlation) to 2 (for parameters with a perfect negative correlation).

The second metric, d_2, considers negative correlation as equivalent to positive correlation. It is calculated using the absolute value of the correlation coefficient (Eq. 2):

$$d_2 = 1 - |r| \tag{2}$$

This metric ranges from 0, where the parameters have correlation coefficients of 1 or -1, to 1, where there is a complete lack of correlation between the parameters.

To select the optimal value for the hyperparameter ε, the average value of the silhouette coefficients was used (Rousseeuw 1987). This is one of the most commonly used metrics for evaluating clustering quality. The implementation of the HDBSCAN algorithm from the scikit-learn machine learning library[4] was used in this study.

Factor analysis

Another alternative approach that we used to identify correlation groups is factor analysis. This method is used in conjunction with cluster analysis to increase the reliability and validity of the results.

Factor analysis is based on the assumption that there are a small number of latent variables (called factors) underlying the observed variables. Observed variables can be expressed as linear or non-linear combinations of factors (Mulaik 2009; Gorsuch 2014). The most common model currently used is the linear model for the relationship between factors and observed variables. It can be expressed mathematically as (Eq. 3):

$$X = AP + U + E \tag{3}$$

where X is a matrix of observed values with m rows and n columns, corresponding to n observed variables and m samples. P is a matrix of factor scores. It has a size of $k \times m$ ($k << n$) and contains columns with the coordinates of the observed variables in the new space of k factors. U is a matrix of deviations from the mean of the observed values, and E is an error matrix. A is called a factor matrix of size $n \times k$. Its elements are called factor loadings, which are the coordinates of the factor space basis and reflect the influence of the factors on the observed variables (Reyment and Jöreskog 1996).

Before conducting factor analysis, it is common to standardize the values of the observed variables. This process leads to the matrix U becoming a zero matrix. This simplifies further analysis.

The goal of further calculations is to determine a matrix A in which the factor loadings of each variable for different factors are as distinct as possible, while minimizing the number of factors and the values of the elements in the error matrix E. The most common approach to solving this problem is the so-called "rotation". It involves rotating the initial basis or its subset in the space of observed variables by a certain angle. This operation is done to satisfy the criteria mentioned above. The resulting factors can be either orthogonal or non-orthogonal, depending on the specific rotation method used.

These methods are based on a specific criterion for optimally choosing the factor matrix. Historically, the first criterion was the quartimax method proposed in the work (Ferguson 1954). This criterion corresponds to the maximization of the criterion q_4, which is the sum of the factor loadings aij in the fourth power (Eq. 4):

$$q_4 = \sum a_{ij}^4 \rightarrow maximum \tag{4}$$

A feature of this method is that it tends to produce factors that are too general. The number of factors produced is too small, and each of these factors has too great an influence on several observed variables simultaneously. Nevertheless, this criterion is still in use today.

As a development of the quartimax method, the varimax method was proposed in the work (Kaiser 1958). According to it, the criterion to be maximized is

$$V_{varimax} = \sum_i \left[\sum_j \left(\frac{a_{ij}}{\sqrt{\sum_j aij^2}} \right)^4 - \frac{\left(\sum_j \left(\frac{a_{ij}}{\sqrt{\sum_j a_{ij}^2}} \right)^2 \right)^2}{n} \right] \rightarrow maximum \tag{5}$$

where a_{ij} are the elements of the factor matrix A and n is the number of observed variables.

There are several other rotation methods available, both orthogonal (such as oblimax, equimax, and parsimax) and non-orthogonal (including promax and quartimin). Each of these methods has its set of benefits and drawbacks. However, varimax and quartimin, which are both orthogonal, are currently the most commonly used in factor analysis.

As can be seen, the search for a factor matrix is reduced to solving an optimization problem of the corresponding criterion. Currently, several methods are used for this purpose, the most effective of which is recognized as the Gradient Projection Algorithm (GPA) (Jennrich 2001; Jennrich 2004).

In this study, two rotation methods were used to identify correlation groups among bioclimatic parameters: quartimax and varimax. Their implementation in the scikit-learn package was used. Before the analysis, the values of the bioclimatic parameters were standardized.

[4] *https://scikit-learn.org/stable*

According to the accepted approach, it was considered that the identified factor could be defined as the main factor for a certain parameter (in other words, the parameter could be attributed to a certain correlation group) if its loading value was maximum (since loadings can have negative values, we will talk hereinafter about their absolute values) and exceeded the values of the loadings of other factors by at least 30%. If there were one or more factors with a lower loading value and the difference in loadings did not exceed 30% of the maximum value, then a conclusion was drawn about the influence of several main factors on the bioclimatic parameter (Mulaik 2009).

RESULTS

Correlation matrix analysis

Fig. 1 shows a heatmap of the correlation matrix for all 19 bioclimatic parameters. This matrix contains Pearson linear correlation coefficients *r*. As can be seen, all bioclimatic parameters can be divided into several groups.

Firstly, two groups of parameters are distinguished, containing temperature (BIO1-BIO11, excluding BIO4 and BIO7) and humidity (BIO12-BIO19, excluding BIO15) factors. Within these groups, the correlations are significantly higher than those between parameters from different groups. Within the first group, the correlation coefficients ranged from 0.53 to 0.99, with an average of 0.829. In the second group, they ranged from 0.45 to 0.99, averaging 0.712. The correlation coefficients between the groups ranged from -0.05 to 0.63, with an average of 0.32.

Secondly, two factors stand out among the temperature parameters: BIO4 and BIO7. These factors characterize the annual temperature range and have a fairly strong negative correlation with most other parameters, except for BIO7 and BIO2. There is also a strong positive correlation between BIO4 and BIO7 ($r = 0.97$).

The BIO2 parameter also stands out, having a fairly weak positive correlation with the temperature parameters (ranging from 0.08 to 0.29), and a weak negative correlation with the humidity parameters (ranging from -0.14 to -0.3), except for BIO15 ($r = 0.38$).

The parameter BIO15, in turn, also stands out among the other humidity parameters. It has a negative or very weak positive (with BIO13) correlation with other humidity parameters and a low positive correlation (from 0.17 to 0.38) with most temperature parameters, except for the above-described BIO4 and BIO7, with which it has *r* values equal to -0.12 and -0.03, respectively.

Thus, five correlation groups can be distinguished already at the stage of simple analysis of the correlation matrix of bioclimatic parameters:

1. BIO1, BIO3, BIO5, BIO6, BIO8-BIO11
2. BIO12-BIO14, BIO16-BIO19
3. BIO4, BIO7
4. BIO2
5. BIO15.

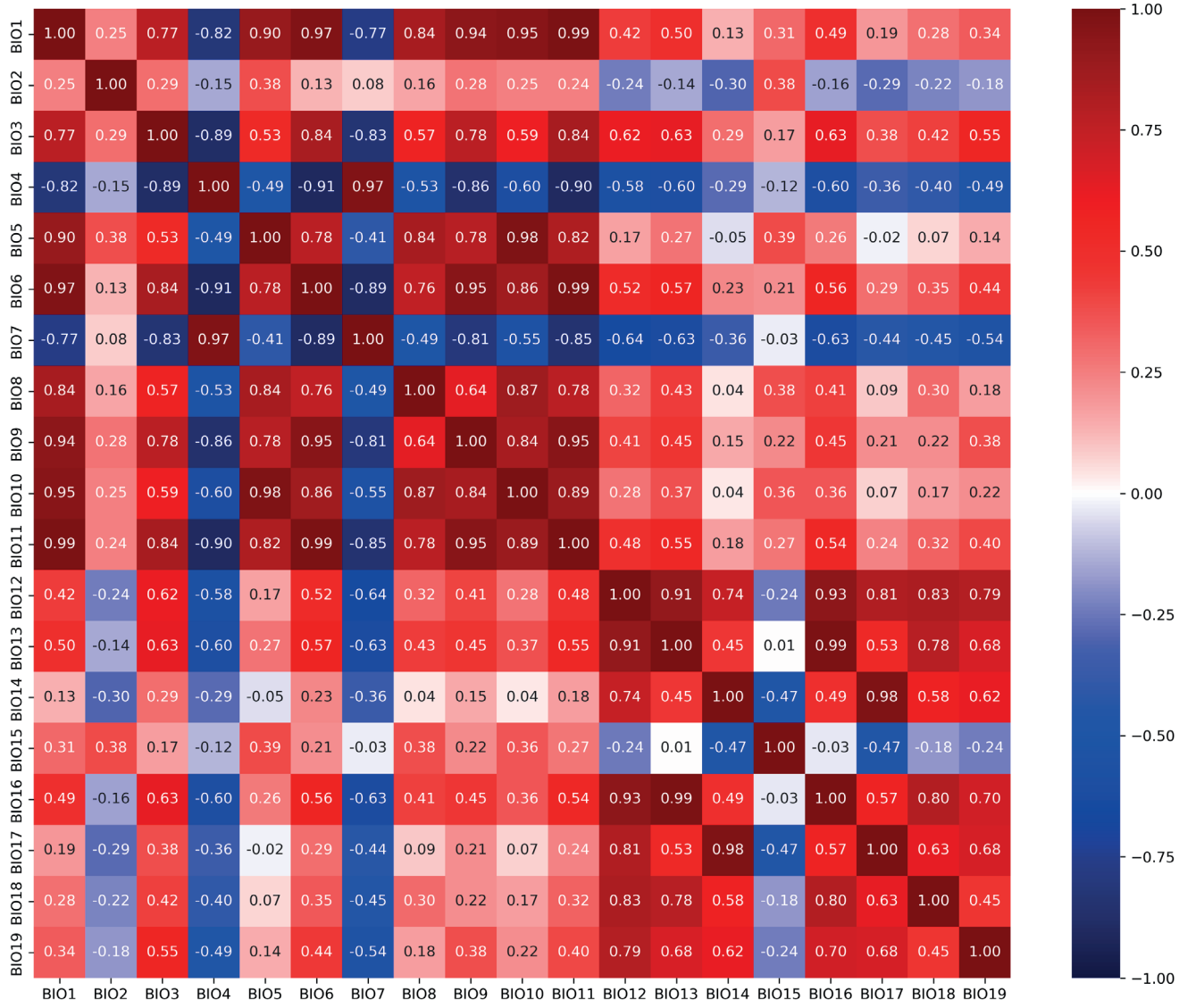

Fig. 1. Heatmap of the correlation matrix of bioclimatic parameters

Results of the cluster analysis

The optimal value of the hyperparameter ε for the HDBSCAN algorithm was found by simply enumerating its possible values in the range from 0.01 to 0.5, with a step size of 0.01. For the metric d_1, the optimal ε value was found to be in the range of 0.19-0.36, giving an average silhouette coefficient of 0.6336. For the metric d_2, the same range of values (0.19-0.36) was found to provide the optimal ε, with an average silhouette coefficient of 0.5531.

Table 2 shows the obtained distribution of bioclimatic parameters by clusters for the two distance metrics used. The noise points are marked with a value of -1. As can be seen, when using the d_1 metric, the HDBSCAN algorithm identified three clusters and two noise points. In the case of the d_2 metric, two clusters and two noise points were identified. In both cases, the BIO2 and BIO15 parameters were identified as noise points. Cluster 1 was completely the same for both metrics. Cluster 0, obtained for the d_2 metric, when using the d_1 metric, was divided into two clusters: 0 and 2. In this case, cluster 2 contained the parameters BIO4 and BIO7.

As can be seen, the results of the cluster analysis coincide completely with the results of a simple analysis of the correlation matrix. The noise points (BIO2 and BIO15 parameters) were previously assigned to groups 4 and 5, respectively. Cluster 1 corresponds to group 2, and cluster 0 (d_2 metric) includes groups 1 and 3. When the d_1 metric is used, clusters 0 and 2 coincide completely with these groups.

Results of the factor analysis

Table 3 shows the results of the factor analysis (factor matrix and identified main factors) conducted using the varimax method. As can be seen, varimax identifies 5 factors. Meanwhile, for most bioclimatic parameters, it can be concluded that there is only one main factor.

The temperature parameters BIO1 and BIO3-BIO11 are influenced by the main factor 1, which is consistent with the results of the correlation matrix analysis and cluster analysis.

The parameters BIO4 and BIO7 are also influenced by the main factor 3. This conclusion is consistent with the results of the cluster analysis, which allocated them to cluster 2 when using the metric d_1 and combined them with cluster 0, corresponding to factor 1 when using the metric d_2. In this case, the loadings of factor 1 for these parameters are positive, unlike the loadings of the other temperature parameters, for which they are negative. This means a different nature of the influence of factor 1 on these parameters, and corresponds to the negative correlation of the parameters BIO4 and BIO7 with the other temperature parameters (except BIO2). These circumstances allow us to allocate the parameters BIO4 and BIO7 to a separate group, if we take into account the nature of their correlation with other temperature parameters, or to combine them if the sign of the correlation coefficient is considered to be unimportant.

The BIO2 parameter has one main factor, 5, which is not the main factor for any other parameter. This corresponds to the allocation of this factor to a separate group 4 and to a separate noise point.

Table 2. Belonging of the studied bioclimatic parameters to the selected clusters according to two metrics

Bioclimatic parameter	Cluster number	
	metric d_1	metric d_2
BIO1	0	0
BIO2	-1	-1
BIO3	0	0
BIO4	2	0
BIO5	0	0
BIO6	0	0
BIO7	2	0
BIO8	0	0
BIO9	0	0
BIO10	0	0
BIO11	0	0
BIO12	1	1
BIO13	1	1
BIO14	1	1
BIO15	-1	-1
BIO16	1	1
BIO17	1	1
BIO18	1	1
BIO19	1	1

Table 3. Factor matrix of the bioclimatic parameters obtained using the varimax method (loadings of the main factors are highlighted)

Parameter	Factor loadings					Main factor
	1	2	3	4	5	
BIO1	**-0.9393**	0.2066	-0.2091	0.0183	-0.0338	1
BIO2	-0.2896	-0.2246	-0.0365	-0.2483	**-0.5180**	5
BIO3	**-0.6146**	0.4194	-0.4777	0.0983	-0.1950	1
BIO4	**0.6481**	-0.3408	**0.6066**	-0.1174	0.0304	1.3
BIO5	**-0.9362**	0.0151	0.1118	-0.0779	-0.0902	1
BIO6	**-0.8679**	0.2764	-0.3608	0.0902	0.0142	1
BIO7	**0.5878**	-0.3902	**0.6036**	-0.1867	-0.0852	1.3
BIO8	**-0.8371**	0.2314	0.1058	-0.0916	0.0079	1
BIO9	**-0.8548**	0.1491	-0.3749	0.0748	-0.0605	1
BIO10	**-0.9579**	0.0986	0.0285	-0.0312	-0.0168	1
BIO11	**-0.8915**	0.2533	-0.3324	0.0422	-0.0432	1
BIO12	-0.2235	**0.8269**	-0.1694	0.4163	0.0242	2
BIO13	-0.3051	**0.8701**	-0.1482	0.1011	0.0163	2
BIO14	-0.0141	0.4509	-0.0585	**0.7711**	0.0280	4
BIO15	-0.3547	-0.0657	0.0470	**-0.5679**	-0.1909	4
BIO16	-0.2923	**0.8753**	-0.1525	0.1422	0.0178	2
BIO17	-0.0488	0.5216	-0.1043	**0.7473**	0.0183	4
BIO18	-0.1147	**0.7889**	-0.0327	0.2531	0.0673	2
BIO19	-0.1780	**0.5949**	-0.2218	0.4326	-0.0451	2

The humidity parameters BIO12, BIO13, BIO16, BIO18, and BIO19 have one main factor, 2, which corresponds to their assignment to cluster 1 and correlation group 2.

The parameters BIO14, BIO15, and BIO17 are influenced by one main factor, 4, which distinguishes them from the other parameters. At first glance, they could be combined into one group on this basis. However, the values of the loadings of factor 4 for these parameters have a peculiarity: the loading of the parameter BIO15 is negative, and that of BIO14 and BIO17 is positive. This distinction means that this factor determines these parameters in different senses: it has a negative relationship with BIO15 and a positive relationship with BIO14 and BIO17. This difference can also be seen in the correlation matrix: BIO14 and BIO17 have a strong positive correlation with each other (r = 0.98) and a moderate negative correlation with BIO15 (r = -0.47 for both parameters). In addition, BIO14 and BIO17 have quite large positive loadings for factor 2. The loadings of the other humidity bioclimatic parameters, for which this factor is the main one, are also positive. At the same time, the loading of factor 2 for BIO15 is very low. These results allow us to single out the parameter BIO15 as a separate group, as well as the parameters BIO14 and BIO17, but this group has a relative proximity to the parameters that are influenced by factor 2, as by the main one.

Table 4 shows the factor matrix obtained as a result of applying the quartimax method.

All temperature parameters, except BIO2, are influenced by factor 1. Simultaneously, the parameters BIO4 and BIO7 have loadings of the main factor with signs opposite to the signs of loadings for the other parameters. This finding is consistent with the negative correlation between these groups of parameters. A similar situation was observed when using the varimax method, as well as cluster analysis, which, when using the d_1 metric, singled out these parameters into a separate group, and when using the d_2 metric, combined them with other temperature parameters.

The parameter BIO2 is influenced by one main factor, 4, for which it is the only parameter with a significant load value.

All the humidity parameters are influenced by the main factor 2. At the same time, the parameters BIO14 and BIO17 do not differ from other parameters, as was the case with the varimax method. But the parameter BIO15 is determined not only by the factor 2 but also by the main factor 3, which does not influence any other parameter. The loading value of factor 2 for BIO15 also has a different sign from the sign of the loadings of this factor for other humidity parameters. These results obtained on the basis of the quartimax method allow us to single out the parameter BIO15 into a separate group and to combine the other humidity parameters. This data is consistent with the results of the correlation matrix analysis, cluster analysis, and partly with the results of using the varimax method.

In general, it is possible to note the consistency of the results obtained from all applied methods for identifying correlation groups. At the same time, factor analysis is

Table 4. Factor matrix of bioclimatic parameters obtained using the quartimax method (loadings of the main factors are highlighted)

Parameter	Factor loadings				Main factor
	1	2	3	4	
BIO1	**-0.9782**	0.0875	0.0013	0.0039	1
BIO2	-0.2851	-0.3667	-0.0766	**-0.4936**	4
BIO3	**-0.7772**	0.3583	-0.0554	-0.1790	1
BIO4	**0.8274**	-0.3108	-0.0107	0.0122	1
BIO5	**-0.8659**	-0.1456	0.0005	-0.0456	1
BIO6	**-0.9579**	0.2029	0.0282	0.0449	1
BIO7	**0.7688**	-0.3988	-0.0406	-0.0978	1
BIO8	**-0.7973**	0.0454	-0.1254	0.0440	1
BIO9	**-0.9367**	0.0872	0.0772	-0.0274	1
BIO10	**-0.9161**	-0.0458	0.0047	0.0257	1
BIO11	**-0.9731**	0.1523	-0.0006	-0.0096	1
BIO12	-0.3533	**0.9003**	-0.0335	0.0069	2
BIO13	-0.4409	**0.7659**	-0.3197	0.0093	2
BIO14	-0.0623	**0.7781**	0.4386	0.0009	2
BIO15	-0.3410	**-0.3957**	**-0.4259**	-0.1593	2, 3
BIO16	-0.4294	**0.7930**	-0.2879	0.0092	2
BIO17	-0.1186	**0.8247**	0.3872	-0.0083	2
BIO18	-0.2088	**0.7885**	-0.1648	0.0506	2
BIO19	-0.2984	**0.7201**	0.0901	-0.0608	2

distinguished by greater complexity in interpreting the results, although it allows the detection of some subtle properties of the data not revealed by other methods.

Selection of parameters from the identified correlation groups

On the basis of the above results, it is possible to identify five groups of bioclimatic parameters, the correlation within which is higher than the correlation with parameters from other groups. The composition of these groups is presented in Table 5.

In case it is assumed that the negative correlation has the same value as the positive one, it is possible to combine groups 5 and 1. Also, from the results of the factor analysis using the varimax method, it follows that the parameters BIO14 and BIO17 can be separated, if necessary, from group 2 into a separate group 6 (for example, if it is known that they are of particular importance for modeling the distribution of the species under study).

Next, a final selection of parameters was carried out, one from each identified group that demonstrated minimal correlation with parameters from other groups. For this purpose, the average values of the corresponding linear correlation coefficients and their absolute values were calculated (Table 6).

Based on the results presented in Tables 5 and 6, a list of selected bioclimatic parameters can be proposed as follows:

1. BIO2 (mean diurnal range (mean of monthly (max temp - min temp)))
2. BIO5 (max temperature of warmest month)
3. BIO7 (temperature annual range BIO5-BIO6)
4. BIO14 (precipitation of driest month)

Table 5. Identified correlation groups of bioclimatic parameters

Group	Bioclimatic parameters
1	BIO1, BIO3, BIO5, BIO6, BIO8-BIO11
2	BIO12-BIO14, BIO16-BIO19
3	BIO2
4	BIO15
5	BIO4, BIO7

Table 6. Average values of correlation coefficients *r* and average absolute values of correlation coefficients |*r*| between bioclimatic parameters and parameters from other groups (the minimum values in each group are highlighted)

Bioclimatic parameter	Average value	
	r	| *r* |
Group 1		
BIO1	0.121	0.409
BIO3	0.206	0.519
BIO5	**0.064**	**0.242**
BIO6	0.136	0.464
BIO8	0.117	0.303
BIO9	0.100	0.405
BIO10	0.087	0.296
BIO11	0.134	0.452
Group 1 (including BIO4 and BIO7)		
BIO1	0.324	0.324
BIO3	0.443	0.443
BIO4	-0.396	0.396
BIO5	0.179	**0.194**
BIO6	0.367	0.367
BIO7	**-0.404**	0.422
BIO8	0.256	0.256
BIO9	0.308	0.308
BIO10	0.234	0.234
BIO11	0.358	0.358
Group 2		
BIO12	0.127	0.409
BIO13	0.201	0.429
BIO14	**-0.034**	**0.211**
BIO16	0.190	0.427
BIO17	-0.008	0.253
BIO18	0.073	0.282
BIO19	0.100	0.342
Group 2 (without BIO14 and BIO17)		
BIO12	0.219	0.462
BIO13	0.242	0.438
BIO16	0.239	0.442
BIO18	**0.149**	**0.328**
BIO19	0.179	0.386

Group 5		
BIO4	-0.544	**0.554**
BIO7	**-0.563**	0.563
Group 6		
BIO14	**0.145**	**0.389**
BIO17	0.184	0.434

5. BIO15 (precipitation seasonality (coefficient of variation))

6. If it is necessary to separate group 6 from group 2, BIO18 (precipitation of the warmest quarter) can be added to this list, but this should be done with caution due to the mixed nature of this parameter and the possible negative effects associated with it when constructing species distribution models (see Introduction).

If the same meaning of positive and negative correlations is accepted, the parameter BIO7 can be removed from the list due to the merging of groups 2 and 5. Scatter plots of the mutual dispersion of these six parameters and the values of their correlation coefficients r are presented in Fig. 2.

As can be seen in Fig. 2, the maximum value of the correlation coefficient between the selected parameters is 0.389 (BIO5 and BIO15). In absolute value, it is -0.582 (BIO14 and BIO18). Generally, the correlation between these selected parameters is quite low.

To compare the results obtained, we selected parameters using a method based on pairwise correlation threshold. Only those parameters were selected that had values of the linear correlation coefficient *r* below a certain value. As a result, only two parameters were selected at the threshold of 0.7 (BIO2 and BIO15), three parameters were selected at the threshold of 0.8 (BIO2, BIO15 and BIO19), five parameters at the threshold of 0.85 (BIO2, BIO3, BIO15, BIO18, and BIO19) and six parameters at the threshold of 0.9 (BIO 2, BIO 3, BIO8, BIO15, BIO18 and BIO19). Different

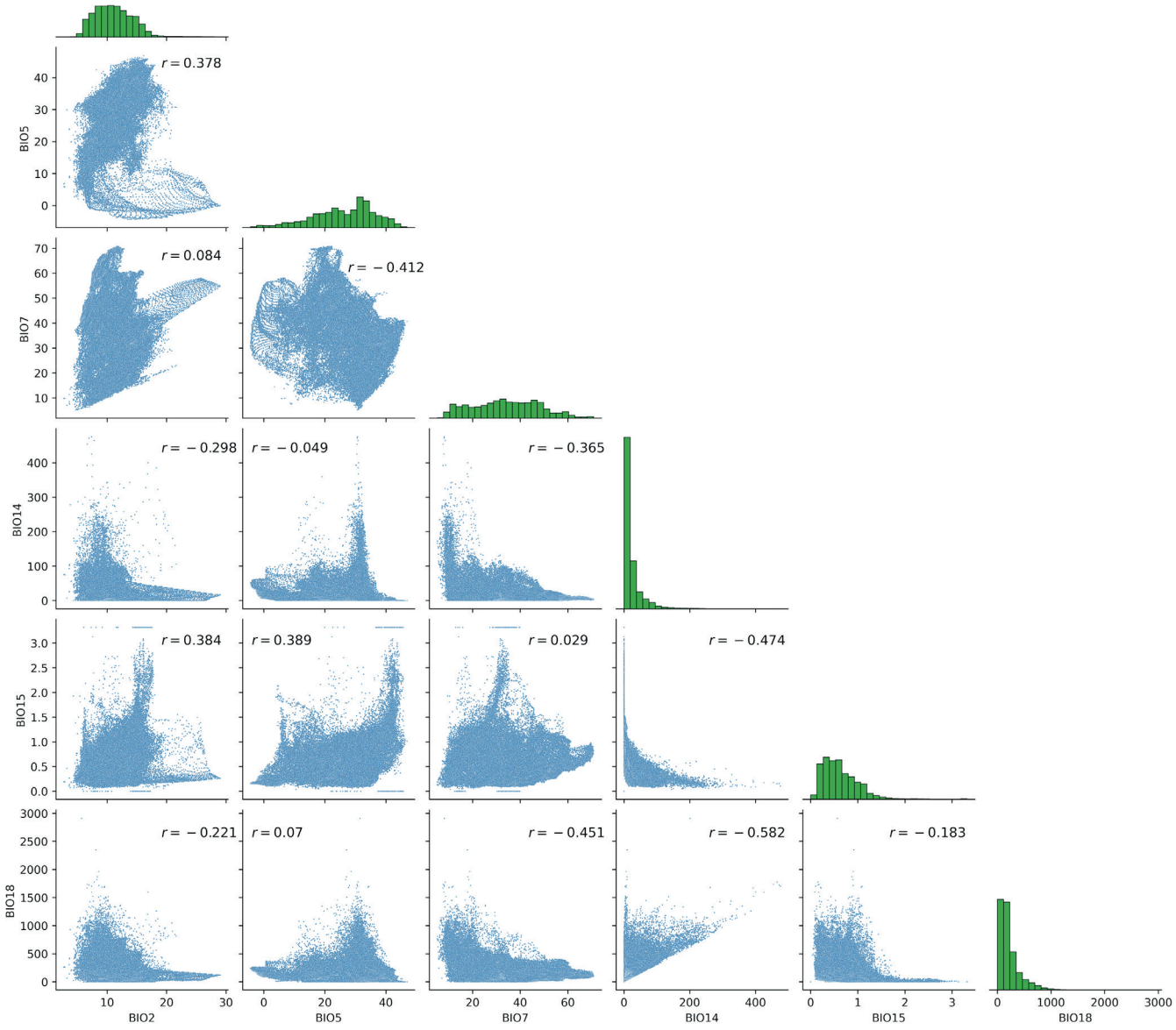

Fig. 2. Scatter plots, linear correlation coefficients r and histograms of distributions (on the diagonal) for six selected bioclimatic parameters

threshold values lead to different numbers of selected parameters. What threshold must be used is unclear. At threshold 0.85 maximum value of the linear correlation coefficient is 0.575 (BIO3 and BIO8), which is significantly higher than the maximum value for the method we used (0.389). The number of parameters with low correlation coefficients with other selected parameters were lost. As can be seen, this comparative study indicates that the approach we used for the analyzed data is more effective than the commonly used method based on selection by correlation threshold.

DISCUSSION

As noted in the introduction, the problem of reducing the number of predictors in SDM, as in any classification problem, is an important step in reducing the overfitting of the constructed models. The resource for this reduction is the presence of redundant information in the initial set of predictors, expressed in a high level of correlation between them.

As can be seen in the results of this study, the statistical approach we proposed made it possible to reduce the pairwise correlation to a low level. At the same time, the number of selected predictors (5 or 6), as experience shows, is sufficient to build effective species distribution models. It can be noted that the number of main correlation groups of bioclimatic parameters identified in this study coincides with the number of synthetic variables obtained as a result of using a neural network of the Variable Autoencoder type in the paper (Dinnage 2023), which was mentioned in the introduction.

The use of the HDBSCAN cluster analysis algorithm to identify correlation groups in our study showed its effectiveness. With its help, a fairly large number of clusters with a good level of difference between them were identified. At the same time, the technology of its application and, importantly, the interpretation of the obtained results are easy to use and can be applied routinely.

The results of factor analysis, in general, with the exception of some nuances, corresponded to the results of the cluster analysis. This fact confirms the reliability of the results of the cluster analysis. The assignment of a number of parameters to several main factors is quite consistent with the presence of a high negative correlation between the parameters. When using factor analysis, it is important to pay attention to the sign of the loading. However, it should be noted that the sufficient complexity and ambiguity of the interpretation of the factor analysis results make it less preferable for routine use in SDM practice compared to cluster analysis.

Our proposed approach to the final selection of parameters from correlation groups is not the only possible one. Firstly, it is possible to select them based on the special significance of any parameter for the vital activity of the organism, known in advance from physiological or ecological studies. Secondly, it is possible to make a selection based on the results of a preliminary distribution modeling using an unreduced set of predictors, followed by analysis of their importance for model construction. Approaches based on the jackknife principle, with successive elimination of parameters or modeling using only one parameter, can be applied. Thirdly, the approach used in our work can also estimate the correlation in the final set of predictors in another way. For example, we can use multiple correlation metrics, such as the variance inflation factor (VIF).

In this study, the values of 19 bioclimatic parameters were analyzed across the globe for the period of 1991–2020. Obviously, even when analyzing this set of parameters for a narrower geographic area or for a different time period, different results can be obtained. The degree and nature of the correlation between these variables vary in time and space, and also depend on the spatial scale of their calculation (Dormann et al. 2012).

Reducing the number of predictors while preserving the information they contain as much as possible is a common problem in machine learning and predictive systems, as noted in the introduction. The approach proposed in this work can be applied to a wide variety of areas related to modeling and forecasting, including both classification and regression. First of all, it can be useful for climatological and meteorological studies, since meteorological and climatological parameters tend to strongly correlate with each other.

CONCLUSIONS

In the course of the conducted studies, using several methods, it was shown that, for the period 1991-2020, for the entire territory of the Earth, it is possible to identify 4-6 correlation groups of bioclimatic parameters, depending on the interpretation of the negative correlation. From these groups, it is possible to select six bioclimatic parameters that demonstrate a minimum average correlation with parameters from other groups. The obtained results are an illustration of the proposed method for reducing bioclimatic parameters and focusing on the selected time period and geographical area. They are of a recommendatory nature. The developed approach to reduce the number of predictors can be used in various areas of statistical modeling and forecasting, both in classification and in regression analysis.

REFERENCES

Araújo M.B., Anderson R.P., Barbosa M.A., Beale C.M., Dormann C.F., Early R., Garcia R.A., Guisan A., Maiorano L., Naimi B., O'Hara R.B., Zimmermann N.E., and Rahbek C. (2019). Standards for distribution models in biodiversity assessments. Science advances, 5(1), DOI: 10.1126/sciadv.aat4858.

Bellard C., Thuiller W., Leroy B., Genovesi P., Bakkenes M., and Courchamp F. (2013). Will climate change promote future invasions? Global Change Biology, 12(19), 3740–3748, DOI: 10.1111/gcb.12344.

Bodjrènou R., Sintondji L., N'Tcha Y., Germain D., Azonwade F., Sohindji F., Hounnou G., Amouzouvi E., Kpognin A., and Comandan F. (2025). Assessment of Hydrologic Data Estimates From ERA5 Reanalyses in Benin, West Africa. Geoscience Data Journal, 12(1), 1-16, DOI: 10.1002/gdj3.288

Bonan G.B. (2008). Ecological Climatology. 2nd ed. Cambrige: Cambrige University Press, DOI: 10.1017/CBO9780511805530.

Booth T.H. (2018). Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. Austral Ecology, 43(8), 852-860, DOI: 10.1111/aec.12628.

Booth T.H. (2022). Checking bioclimatic variables that combine temperature and precipitation data before their use in species distribution models. Austral Ecology, 47(7), 1506-1514, DOI: 10.1111/aec.13234.

Bradie J. and Leunig B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. Journal of Biogeograhy, 44(6), 1344–1361, DOI: 10.1111/jbi.12894.

Busby J. R. (1991). BIOCLIM – A bioclimate analysis and prediction system. Plant Protection Quarterly, 6(1), 8–9.

Campello R.J.G.B., Moulavi D., and Sander J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In: J. Pei, V.S. Tseng, L. Cao, H. Motoda, and G. Xu, eds., Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science, 7819. Berlin, Heidelberg: Springer, 160-172, DOI: 10.1007/978-3-642-37456-2_14.

Dinnage R. (2023). How many variables does Wordclim have, really? Generative A.I. unravels the intrinsic dimension of bioclimatic variables. bioRxiv preprint, DOI: 10.1101/2023.06.12.544623.

Dormann C.F., Elith J., Bacher S., Buchmann C., Carré G., Marquéz J.R.G., Gruber B., Lafourcade B., Leitão P.J., Münkemüller T., McClean C., Osborne P.E., Reineking B., Schröder B., Skidmore A.K., Zurell D., and Lautenbach, S. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27–46, DOI: 10.1111/j.1600-0587.2012.07348.x.

Eldridge J., Belkin M., and Wang Y. (2015). Beyond Hartigan Consistency: Merge Distortion Metric for Hierarchical Clustering. Proceedings of The 28th Conference on Learning Theory. Proceedings of Machine Learning Research, 40, 588-606.

Fatolahzadeh G. A., Maghoul P., Ojo E.R., and Shalaby A. (2024). Reliability of ERA5 and ERA5-Land reanalysis data in the Canadian Prairies. Theoretical and Applied Climatology, 155(4), 3087-3098, DOI: 10.1007/s00704-023-04771-z.

Ferguson G. A. (1954). The concept of parsimony in factor analysis. Psychometrika, 19, 281–290, DOI: 0.1007/BF02289228.

Fick S.E. and Hijmans R.J. (2017). WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37(12), 4302-4315, DOI: 10.1002/joc.5086.

Franklin J. (2009). Mapping species distributions. Spatial inference and prediction. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511810602.

Gilman S.E., Urban M.C., Tewksbury J., Gilchrist G.W., and Holt R. D. (2010). A framework for community interactions under climate change. Trends in Ecology and Evolution, 25(6), 325-331, DOI: 10.1016/j.tree.2010.03.002.

Gorsuch R.L. (2014). Factor Analysis. New York: Routledge. DOI: /10.4324/9781315735740

Harris I., Osborn T.J., Jones P., and Lister D. (2020). Version 4 of the CRU TS Monthly High-Resolution Gridded Multivariate Climate Dataset. Scientific Data, 7(109), DOI: 10.1038/s41597-020-0453-3.

Hastie T., Tibshirani R., and Friedman J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. New York: Springer. DOI: 10.1007/978-0-387-84858-7.

Hijmans R. J., Cameron S. E., Parra J. L., Jones P. G., and Jarvis A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, 25(15), 1965–1978, DOI: 10.1002/joc.1276.

Jennrich R.I. (2001). A simple general procedure for orthogonal rotation. Psychometrika, 66(2), 289-306, DOI: 10.1007/BF02294840.

Jennrich R.I. (2004). Derivative free gradient projection algorithms for rotation. Psychometrika, 69(3), 475-480, DOI: 10.1007/BF02295647.

Kaiser H.F. (1958). The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23(3), 187-200, DOI: 10.1007/BF02289233.

Malzer C. and Baum M. (2020). A Hybrid Approach To Hierarchical Density-based Cluster Selection. 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Karlsruhe, Germany, 2020, 223-228, DOI: 10.1109/MFI49285.2020.9235263.

McCarty J.P. (2001). Ecological Consequences of Recent Climate Change. Conservation Biology, 15(2), 320–331, DOI: 10.1046/j.1523-1739.2001.015002320.x.

McInnes L. and Healy J. (2017). Accelerated Hierarchical Density Based Clustering. IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 2017, 33-42, DOI: 10.1109/ICDMW.2017.12.

Mulaik S.A. (2009). Foundations of Factor Analysis. 2nd ed. New York: Chapman and Hall/CRC, DOI: 10.1201/b15851.

Nix H.A. (1986). A biogeographic analysis of Australian elapid snakes. In: R. Longmore, ed., Atlas of Elapid Snakes of Australia. Australian Flora and Fauna Series No. 7. Canberra: Australian Government Publishing Service, 4-15.

Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martínez-Meyer E., Nakamura M., and Araújo M.B. (2011). Ecological niches and geographic distributions. Princeton and Oxford: Princeton University Press, DOI: 10.1515/9781400840670.

Petrosyan V., Osipov F., Feniova I., Dergunova N., Warshavsky A., Khlyap L., and Dzialowski A. (2023). The TOP-100 most dangerous invasive alien species in Northern Eurasia: invasion trends and species distribution modelling. NeoBiota, 82, 23–56, DOI: 10.3897/neobiota.82.96282.

Phillips S., Dudík M., and Schapire R.E. (2004). A Maximum Entropy Approach to Species Distribution Modeling. In R. Greiner and D. Schuurmans, eds., Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 655-662, DOI: 10.1145/1015330.1015412.

Phillips S.J., Anderson R.P., and Schapire R.E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3), 231-259, DOI: 10.1016/j.ecolmodel.2005.03.026.

Popova E.N. and Popov I.O. (2013). Climatic factors determining ranges of agricultural pests and agents of plant diseases and model methodology for assessment of change in ranges. Problems of Ecological Monitoring and Ecosystem Modelling, 25, 177–206 (in Russian with English summary).

Popova E.N. and Popov I.O. (2019). Modeling of potential climatic ranges of biological species and their climate-driven changes. Fundamental and Applied Climatology, 1, 58-75, DOI: 10.21513/2410-8758-2019-1-58-75 (In Russian with English summary).

Post E. (2013). Ecology of Climate Change. The Importance of Biotic Interactions. Princenton and Oxford: Princenton Unversity Press. DOI: 10.2307/j.ctt2jc8jj.

Purnadurga G., Kumar T., Kundeti K., Barbosa H., and Mall R. (2019). Evaluation of evapotranspiration estimates from observed and reanalysis data sets over Indian region. International Journal of Climatology, 39(15), DOI: 10.1002/joc.6189.

Reyment R.A. and Jöreskog K. G. (1996). Applied Factor Analysis in the Natural Sciences. Cambridge: Cambridge University Press.

Rousseeuw P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65, DOI: 10.1016/0377-0427(87)90125-7.

Roweis S.T. and Saul L.K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science, 290(5500), 2323-2326, DOI: 10.1126/science.290.5500.2323.

Schimel D. (2013). Climate and ecosystems. Princenton and Oxford: Princenton Unversity Press.

Srivastava V., Lafond V., and Griess V.C. (2019). Species distribution models (SDM): Applications, benefits and challenges in invasive species management. CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources, 14(20), 1-13, DOI: 10.1079/PAVSNNR201914020.

Wierzchoń S. and Kłopotek M. (2018). Modern Algorithms of Cluster Analysis. Studies in Big Data, 34. Cham: Springer, DOI: 10.1007/978-3-319-69308-8.

Zhang H., Zheng S., Huang T., Liu J., and Yue J. (2023). Estimation of potential suitable habitats for the relict plant Euptelea pleiosperma in China via comparison of three niche models. Sustainability, 15(14), 1-23, DOI: 10.3390/su151411035.