STATISTICS IN TRANSITION new series and *STATISTICS OF UKRAINE, February 2023*
A New Role for Statistics: Joint Special Issue
Vol. 24, No. 1, pp. 199–212, DOI 10.59170/stattrans-2023-011
Received – 01.07.2022; accepted – 05.10.2022

A comparative analysis of the principal component method and parallel analysis in working with official statistical data

Halyna Holubova[1]

ABSTRACT

The dynamic development of the digitized society generates large-scale information data flows. Therefore, data need to be compressed in a way allowing its content to remain complete and informative. In order for the above to be achieved, it is advisable to use the principal component method whose main task is to reduce the dimension of multidimensional space with a minimal loss of information.

The article describes the basic conceptual approaches to the definition of principle components. Moreover, the methodological principles of selecting the main components are presented. Among the many ways to select principle components, the easiest way is selecting the first k-number of components with the largest eigenvalues or to determine the percentage of the total variance explained by each component. Many statistical data packages often use the Kaiser method for this purpose. However, this method fails to take into account the fact that when dealing with random data (noise), it is possible to identify components with eigenvalues greater than one, or in other words, to select redundant components. We conclude that when selecting the main components, the classical mechanisms should be used with caution.

The Parallel analysis method uses multiple data simulations to overcome the problem of random errors. This method assumes that the components of real data must have greater eigenvalues than the parallel components derived from simulated data which have the same sample size and design, variance and number of variables.

A comparative analysis of the eigenvalues was performed by means of two methods: the Kaiser criterion and the parallel Horn analysis on the example of several data sets. The study shows that the method of parallel analysis produces more valid results with actual data sets. We believe that the main advantage of Parallel analysis is its ability to model the process of selecting the required number of main components by determining the point at which they cannot be distinguished from those generated by simulated noise.

Key words: principal components, principal component analysis, factor analysis, Kaiser criterion, parallel analysis, simulation

[1]National Academy of Statistics, Accounting and Audit. Kyiv, Ukraine. E-mail: g_kondrya@ukr.net.
ORCID: https://orcid.org/0000-0003-4847-5235

1. Introduction

With the powerful development of the digital economy and the information society as a whole, large amounts of data are produced on a daily basis. Big data is a large data set generated by people using information and communication technologies. Currently, there is no single methodology for generating and summarizing big data that could be used as a universal information base (Osaulenko et al., 2021).

Since socio-economic phenomena and processes are characterized by multidimensionality, which generates large databases, there is a need to summarize, group and concisely identify this information. For the convenience of statistical analysis, it is necessary to determine the main factors or components that form and, accordingly, characterize the phenomenon under study.

For example, the Human Development Index covers a system of statistical indicators that can be summarized in several components: health indicators, education indicators, indicators of material well-being of the population. The assessment of the level of development of information and communication technologies (ICT), which is calculated and published annually by the International Telecommunication Union, is based on 11 indicators, which can be summarized in three sub-indices: availability of infrastructure and access to ICT; intensity of ICT use; ability to use ICT effectively (Korepanov, 2018).

The introduction of experimental statistics, the use of applied statistics methods, the transformation of alternative data sources (for example, departmental statistics) for the production of official statistical information will allow the harmonization of official statistics and achieve comparability of various statistical indicators with international comparisons, classifications, etc.

The most common method of information optimization is the principal components analysis (PCA), which allows one to organize a large array of data. In order to study the internal structure of the object, the dimension of the initial feature set should be compressed, replacing it with a minimum number of components (Ierina, 2014). The main components store all the information about the object of study, Figure 1.

Figure 1 clearly illustrates the transformation of the raw data of multidimensional space into principal components. Currently, there are various methods of implementing PCA. The question arises, which principal components extraction algorithm will work best with the official statistical data? It should be noted that statistical data are usually heterogeneous (there are atypical population units (outliers), a high value of standard deviation, skewness, etc.); are not always subject to the law of normal distribution; some statistical indicators may be incomparable or multidirectional, sometimes incomplete, etc. That is why the choice of the PC method is extremely important, since the formation of the PC serves as the ultimate goal –

for grouping, typology or clustering of data, and an intermediate goal – for advanced statistical research.

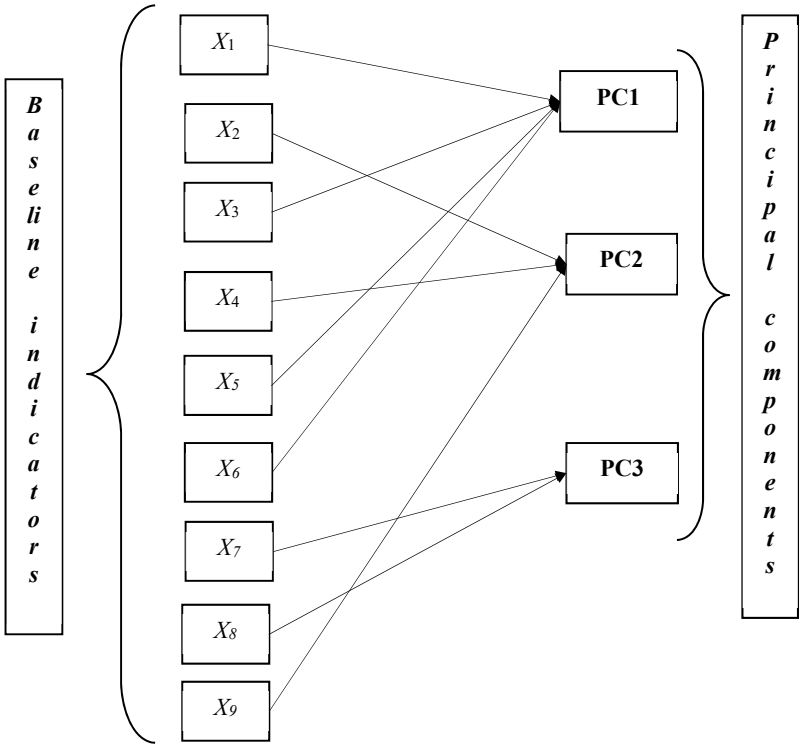

Figure 1. The relationship of the original features (X) and the principal components (PC)

2. Literature review

Applied aspects in the study of methodological principles in separation of the main components are revealed in the works of foreign scholars, including J. L. Horn (1965), H. F. Kaiser (1970), L. W. Glorfeld (1995), R. L. Gorsuch (1983), Ö. Çokluk & D. Koçak (2016), A. V. Silverstein (1987), W. R. Zwick & W. F. Velicer (1986) and many other specialists who studied the mechanism and algorithm in the formation of principal components using various mathematical methods. Domestic scientists use the method of the main components for grouping statistical data, to identify factors influencing the object under study (Holubova, 2013), (Lepeyko & Shcherbak, 2018), (Rosen et al., 2018) or as a tool of public administration (Chinkulyak & Pogrebnyak, 2015), etc.

3. Problem statement

The method of Principal component analysis (PCA) is a powerful research model with the main task to reduce the dimension of multidimensional space with minimal information loss.

This method is useful when working with official statistics, as it is able to group statistics both in statics and in dynamics. Based on the main components it is available to:

- rank and classify objects, countries, regions, enterprises, etc.;
- measure the relationship between primary indicators and key components;
- perform regression analysis, etc.

One of the advantages of using the method of principal components is the ability to get rid of multicollinearity between the original features and perform regression analysis on the principal components.

Assessing the socio-economic development of the country and its regions on a number of statistical indicators, it is possible to identify the economic, social, demographic, political component and so on. Based on the state of the environment according to official statistics, certain types of environmental risks are distinguished according to the degree of risk. By studying demographic statistics, it is possible to classify the population, for example, on type of aging, or to identify factors that affect the decline in birth rates in Ukraine and so on. Analysis of medical statistics data allows, for example, to identify classes of morbidity by age groups or to identify factors that shape the medical system in the country as a whole, and others. Therefore, this method is universal and can be used in various studies using data from both official and administrative statistics.

The PCA method was first proposed in 1901 by K. Pearson, who studied the problem of the best approximation of a set of points by lines and planes.

PCA should not be confused with factor analysis (FA). The latter is a popular method of detecting interpreted linear relationships between variables called factors. In factor analysis, different types of vector rotation are usually used to redistribute variation between factors, while maintaining the total variance of the selected factors. Determining the number of factors is more important than the type of rotation, because the power of factor analysis depends on the ability to distinguish important factors from others. Therefore, it is very important to determine the exact balance between correlations. Determining the number of factors requires close attention, because if the number of isolated factors is greater or less than necessary, it can lead to serious errors that affect the results of the study.

Instead, PCA is a useful method to reduce the number of observed variables to a smaller set of independent components. Therefore, the main goals of PCA are (Holubova, 2020):

1. Data visualization for research analysis, which allows to reveal the latent characteristics of data and interpret the components.
2. Decrease in the number of predictors for future analyzes, such as regression of major components.

4. Methodology

PCA uses elements of linear algebra to determine the basic linear structure inherent in a data matrix. The basis of mathematics in PCA is the decomposition of singular values, which is a generalization of the decomposition of eigenvalues (lambda numbers, λ). The intrinsic value of the principal component is the amount of deviation in the original data, so maximizing the deviation is important because it provides the most information about the actual data. Understanding how these mathematical combinations work is not necessary to understand PCA, but understanding the basic principles of the principal components method is essential when interpreting PCA results.

The study revealed the methodological principles of several methods for selecting the main components (PC). One of the simplest methods for selecting a subset of the PC is to select the first k-number of components with the largest eigenvalues λ. As a result, the main components that best explain the deviations from the data are selected.

The Scree plot stony decline involves the construction of a graph where the abscissa is plotted against the ordinal number of the eigenvalue, and the ordinate – its value. According to R. Cettel (1966), it is necessary to find the point of the greatest slowdown in the decline of eigenvalues and take into account only the factors that correspond to eigenvalues to the left of this point. This criterion is not statistically sound and often leaves not all significant factors in the model.

The criterion of Bartlett's xi-square tests the hypothesis that other eigenvalues are equal, that is, each eigenvalue λ is evaluated sequentially until the null hypothesis is rejected.

One of the classic methods of the PC selection is to study the percentage of total variance, which is explained by each component. Having set a predetermined threshold (usually 75% of the total variance explained), the first k-principal components that collectively explain this variance fraction can be selected as a subset of the components. However, this method of selection, like other methods described above, cannot fully take into account the variance of the data.

Many statistical data packets often use a method that preserves all PC with eigenvalues $\lambda > 1$. It is also called the Kaiser rule, the Kaiser test, or the Kaiser-Gutmann criterion. The basic idea is that with standardized data, the variance of each of the

source variables is 1. Therefore, principal components with an eigenvalue of more than one explain more variance than one variable in the source data. This method is popular and practical, but does not take into account the fact that even with random data (noise) you can identify components with eigenvalues greater than one. In these situations, the variance explained by the components is not really useful, as it is due to accidental error or noise.

Parallel analysis (PA) uses multiple data simulations to overcome the problem of random error. The essence of this method is that non-trivial components from real data should have greater eigenvalues than parallel components derived from simulated data that have the same sample size, variance and number of variables. PA is also called the parallel analysis of Horn in honour of its creator J. Horn. The process of performing parallel analysis is based on the Monte Carlo method, namely, it is a simulation of a large number of data sets. Horn argued that the number of iterations should be sufficient, that is, to obtain the most objective results (for example, 1000 or more repetitions), although there are no strict limits. Experiments were recorded when the results did not show a significant difference between one simulation and one hundred iterations. Each simulated data set contains the same number of variables and observations as the original data. For each simulated variable, data are generated by constructing a sample from a multidimensional normal distribution, with the standard deviation equal to the standard deviation of the corresponding actual data variable.

Repeating the steps of n-times gives n-sets of eigenvalues with the calculation of average eigenvalues by sets. This leads to a single set of average eigenvalues $\bar{\lambda}$, with which the eigenvalues obtained from the actual data set are compared. During the development of the PA method, researchers made an assumption that the use of average eigenvalues is similar to setting the error rate of the first type I (α) at 0.50 (instead of the more acceptable level $\alpha = 0.05$), and this may lead to the existence of factors (extra components). With this in mind, L. W. Glorfeld (1995) and R. A. Harshman & J. R. Reddon (1983) proposed the use of a 95 percent threshold for eigenvalues generated from random data. This is also similar to setting α to 0.05, which is a more common standard for type I error. Eigenvalues from actual data are compared with the values of the 95th percentile of generated data, and not with the average eigenvalue $\bar{\lambda}$ (Hayton & Alllen, 2004).

Therefore, the obtained eigenvalue of the PC from the original data should be compared with the upper 95th percentile, calculated from the simulated data sets. If the eigenvalue from the source data is greater than the upper percentile of the simulated data, the component is selected, otherwise it is discarded. The idea is that due to a random error in the data (caused by sample size, sample design, etc.), the PCA generates some components with eigenvalues greater than one. In general, the first eigenvalues generated by noise data will grow with an increasing number of variables and fall with a decreasing number of observations. Preserving only those PC with

eigenvalues that exceed the 95th percentile of the simulated eigenvalues ensures that the discrepancies explained by these PC are likely to represent real variance rather than noise variance. That is, parallel analysis is considered more useful in practice than the method of selection of principal components by the Kaiser criterion or other methods of selection.

A.V. Silverstein (1987) compared the Kaiser method and the method of parallel analysis on the example of 24 data sets, and it was found that parallel analysis gives better results. W. R. Zwick & W. F. Velicer (1986) conducted a study comparing five methods used to determine the factors (parallel analysis, the method of minimum mean partial correlation, the graph of Scree plot stony decline, the criterion of Bartlett's xi-square, Kaiser's test) taking into account different conditions (sample size, the number of variables and components and their factor loads, etc.). The researchers concluded that the parallel analysis is consistent with the actual data set used to determine the number of factors, with an accuracy of 92%.

5. The application examples

During the study, the author developed several different data sets that are publicly available on the Internet.

When implementing the principal components method, certain statistical preconditions should usually be followed. All variables should be quantitative (categorical variables are excluded from the analysis) and homogeneous, distribution is symmetrical, and the number of observations should prevail over the number of variables. However, depending on the type of study, there may be exceptions, for example, in medicine, chemistry, biostatistics and other sciences, including working with the real statistical data. Or, for example, in the conditions of laboratory tests or in expensive sample observations, when it is not possible to involve a sufficient number of respondents in the experiment, and so on.

Data set 1 (Glorfeld, 1995)

We have information from 500 Facebook (2016) users on 14 indicators: the number of daily posts; the number of posts per hour; the number of posts about personal life; free time; the number of users who subscribed to your page, the number of people who liked your page, the number of users who liked your photo; the number of comments, likes, distributions, etc. Among all these indicators that can be quantified, it is actually difficult to single out a priori the main components that clearly visualize or typify public activity on Facebook, or in some way can describe the principles of interaction with this social media. A comparative analysis of eigenvalues λ was performed by two methods: the Kaiser criterion and the parallel Horn analysis. The obtained results of eigenvalues are given in Table 1.

Table 1. Eigenvalues of the principal components

Principal components	Eigenvalues ƛ (initial data)	Eigenvalues ƛ (Parallel analysis)		
		Average	Upper limit	Lower limit
PC1	5.920	1.289	1.356	1.235
PC2	1.740	1.222	1.270	1.180
PC3	1.658	1.171	1.212	1.136
PC4	1.124	1.126	1.163	1.094
PC5	1.002	1.086	1.117	1.057
PC6	0.835	1.047	1.076	1.016
PC7	0.621	1.009	1.038	0.981
PC8	0.454	0.975	1.002	0.944
PC9	0.304	0.939	0.968	0.913
PC10	0.143	0.905	0.932	0.876
PC11	0.116	0.869	0.899	0.836
PC12	0.069	0.831	0.862	0.798
PC13	0.014	0.790	0.823	0.754
PC14	0.001	0.739	0.779	0.694

As we can see, according to the Kaiser criterion, five main components are selected with the values of more than one (Figure 2), which explains 81.2% of the variation. According to the method of Parallel analysis, only three main components are identified, as evidenced by Figure 3, which visualizes the clipping of three components. The intrinsic value of PC4 is 1.124, which is less than the upper limit of the 95 percent interval (1.163), which gives grounds to exclude this component from further analysis, because its variance is caused by sampling noise, not the real process.

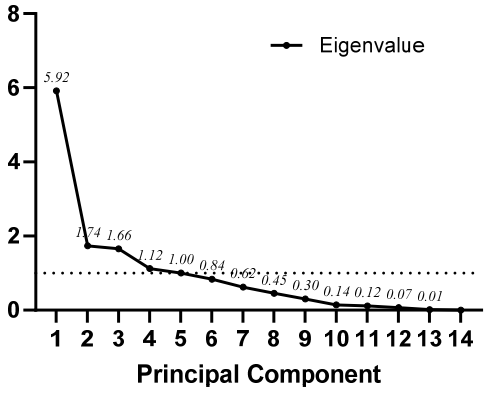

Figure 2. The principal components of the Kaiser criterion

Source: built by the author in the GraphPadPrism packet.

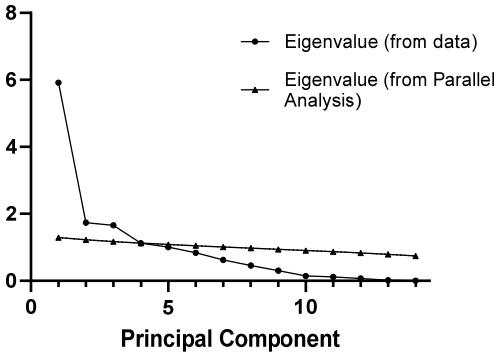

Figure 3. The principal components of the PA method

Source: built by the author in the GraphPadPrism packet.

Data set 2 (Gene Dataset)

For purposes of illustration, a sample of the human gene pool (20 patients and their characteristics of 100 genes, i.e. a matrix of 20 per 100) is considered. Since the variables *(m)* are greater than the observations *(n)*, the maximum number of components that can be selected is *n*-1, in our example - 19. According to the Kaiser criterion, it is established that there are 19 main components and one main component according to the method of Parallel analysis (Table 2, Figures 4, 5). In addition, the number of iterations (10, 100, 1000 and 5000 simulations were used) did not affect the result, i.e. the isolation of only one component is confirmed.

Table 2. Eigenvalues of the principal components

Principal components	Eigenvalues $\hat{\lambda}$ (initial data)	Eigenvalues $\hat{\lambda}$ (Parallel analysis)		
		Average	Upper limit	Lower limit
PC1	17.291	9.749	10.606	9.013
PC2	8.457	8.808	9.449	8.282
PC3	7.587	8.131	8.650	7.678
PC4	6.924	7.575	8.011	7.156
PC5	6.425	7.058	7.477	6.647
PC6	6.157	6.589	6.986	6.229
PC7	5.983	6.151	6.489	5.807
PC8	4.816	5.730	6.050	5.400
PC9	4.524	5.344	5.653	5.019
PC10	4.472	4.969	5.287	4.666
PC11	4.140	4.621	4.924	4.334
PC12	3.983	4.278	4.567	3.984
PC13	3.723	3.954	4.239	3.661
PC14	3.282	3.640	3.920	3.360

Table 2. Eigenvalues of the principal components (cont.)

Principal components	Eigenvalues λ̂ (initial data)	Eigenvalues λ̂ (Parallel analysis)		
		Average	Upper limit	Lower limit
PC15	3.117	3.318	3.600	3.037
PC16	2.745	2.989	3.284	2.712
PC17	2.452	2.686	2.964	2.389
PC18	2.201	2.365	2.655	2.046
PC19	1.720	1.990	2.324	1.651

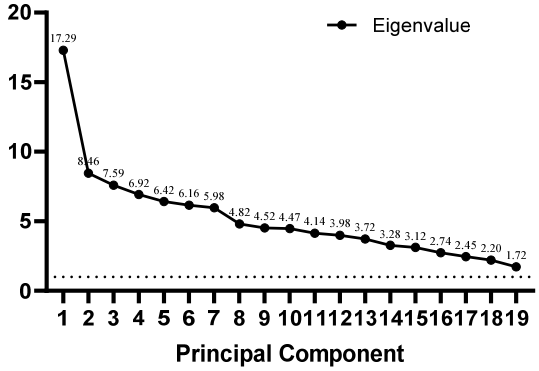

Figure 4. The principal components of the Kaiser criterion

Source: built by the author in the GraphPadPrism packet.

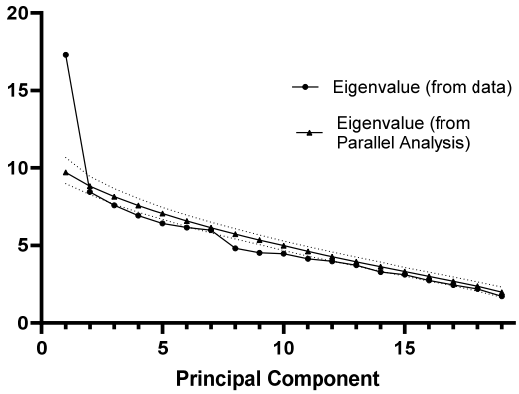

Figure 5. The principal components by the PA method

Source: built by the author in the GraphPadPrism packet.

It is claimed that the number of observations should exceed the number of signs at least twice.

Data set 3 (Decathlon Dataset, 2004).

The author considers a set of decathlon data. These are the results of the 41st athlete in 10 sports at the Olympic Games (2004). The initial data are symmetric, the coefficients of variation for each distribution (ten indicators) in the range of 3-8%, which indicates the homogeneity of the population and the reliability of the average value.

According to the Kaiser criterion, four principal components, which characterize 75% of the variation, are identified. PA allocates only one principal component with the number of simulations 100 and 1000, Table 3. If we use 10 iterations (i.e. only ten correlation matrices are modelled), then two principal components are distinguished. We can assume that this is exactly the case when the number of simulations matters (the more iterations, the more valid the results).

Table 3. Eigenvalues of the principal components

Principal components	Eigenvalues λ̂ (initial data)	Eigenvalues λ̂ (Parallel analysis)		
		Average	Upper limit	Lower limit
PC1	3.272	1.853	2.128	1.642
PC2	1.737	1.559	1.748	1.402
PC3	1.405	1.350	1.493	1.214
PC4	1.057	1.168	1.298	1.061
PC5	0.685	1.016	1.126	0.908
PC6	0.599	0.874	0.972	0.755
PC7	0.451	0.729	0.842	0.623
PC8	0.397	0.605	0.713	0.505
PC9	0.215	0.477	0.581	0.383
PC10	0.182	0.344	0.454	0.240

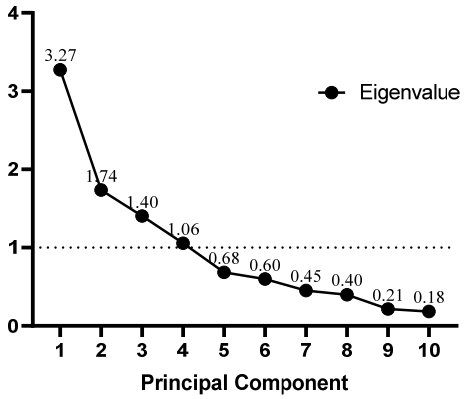

Figure 6. The principal components of the Kaiser criterion

Source: built by the author in the GraphPadPrism packet.

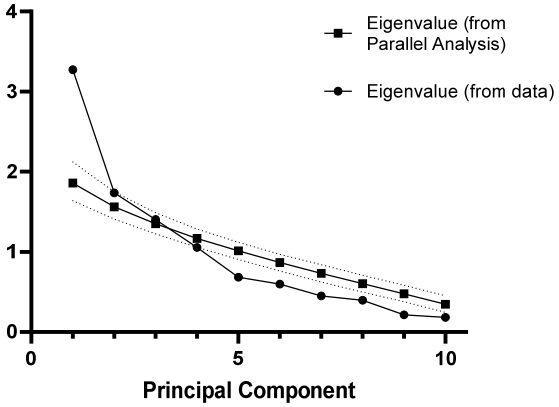

Figure 7. The principal components of the PA method
Source: built by the author in the GraphPadPrism packet.

The results of the study showed that the method of parallel analysis gives consistent results with the actual data sets, taking into account the sample size, symmetry of its distribution, variance and so on. We believe that the method under study gives more objective results in determining the exact number of factors.

6. Conclusions

The data sets used by the author for analysis are not official statistical data. However, these dates clearly characterize the peculiarities of different samples: Facebook Dataset is heterogeneous and has outliers; Gene Dataset is unbalanced in terms of the number of indicators and observations; Decathlon Dataset is indicative, at first glance (homogeneous and symmetrical). On the basis of the Kaiser method, which belongs to classical methods, redundant factors were selected in each of the data sets. According to the results of the Parallel analysis, which is based on multiple simulations, the real number of the main components was determined. Therefore, in our opinion, classical methods of selection of major components should be used with caution. Especially, the statistical data have certain features (heterogeneity, asymmetry, imbalance, etc.), that is why the author considers it appropriate to use Parallel analysis in the context of working with the official statistical data.

The main advantage of Parallel analysis should be the ability to model the process of selecting the number of PCs by determining the point at which the principal components cannot be distinguished from those generated by simulated noise. In our opinion, multiple simulations can protect against erroneous results.

References

Cattell, R. B., (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, pp. 245–276.

Chinkulyak, N. M., Pogrebnyak, L. O., (2015). Statystychnyi analiz yak instrument derzhavnoho upravlinnia [Statistical analysis as a tool of public administration] *Derzhavne upravlinnia – Governance*, 1, pp. 82–88, [in Ukrainian].

Çokluk, Ö., Koçak, D., (2016). Using Horn's parallel analysis method in exploratory factor analysis for determining the number of factors. Educational Sciences: *Theory & Practice*, 16, pp. 537–551.

Decathlon Dataset, (2004). Retrieved from: https://malouche.github.io/data_in_class/decathlon_data.html.

Facebook Dataset, (2016). Machine learning repository. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Facebook+metrics.

Gene Dataset. Machine learning repository. Retrieved from: https://archive.ics.uci.edu/ml/datasets.php?format=&task=&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table.

Glorfeld, L. W., (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, pp. 377–393.

Gorsuch, R. L., (1983). Factor analysis (2nd ed.). Hillsdale, NJ: LawrenceErlbaum.

Harshman, R. A., Reddon, J. R., (1983). Determining the number of factors by comparing real with random data: A serious flaw and some possible corrections. *Proceedings of the Classification Society of North America at Philadelphia*, pp. 14–15.

Hayton, J., Alllen, D., (2004). Factor Retention. Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. Retrieved from: https://www.researchgate.net/publication/235726204_Factor_Retention_Decisions_in_Exploratory_Factor_Analysis_A_Tutorial_on_Parallel_Analysis/link/5582a85008ae6cf036c1a886/download.

Holubova, H. V., (2013). Statystychnyi analiz osnovnykh faktoriv vplyvu na tranzyt vantazhiv v Ukraini [Statistical analysis of basis factor influence on in-transit freight in Ukraine by regression model] Visnyk Kyivskoho natsionalnoho universytetu im. T. Shevchenka. Ekonomika – Bulletin of Taras Shevchenko National University of Kyiv. *Economics*, 134, pp. 12–16, [in Ukrainian].

Holubova, H. V., (2020). Pryntsypy vyboru holovnykh komponent: osoblyvosti prykladnoho modeliuvannia [Principles of choice of main components: features of applied modeling] Novi dzherela ta metody poshyrennia danykh u statystytsi: materialy XVIII Mizhnarodnoi naukovo-praktychnoi konferentsii z nahody Dnia pratsivnykiv statystyky]. New sources and methods of data dissemination in statistics: materials of the XVIII International scientific-practical conference on the occasion of the Day of Statistics. Kyiv, Informatsiino-analitychneahenstvo. *Information and Analytical Agency*, pp. 155–160, [in Ukrainian].

Horn, J. L., (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32, pp. 179-185.

Ierina, A. M., (2014). Komponentnyi analiz [Component analysis] Statystychne modeliuvannia ta prohnozuvannia [*Statistical modeling and forecasting*], 348, pp 287–313.

Kaiser, H. F., (1970). A second-generation Little Jiffy. *Psychometrika*, 35, pp. 401-415.

Korepanov, O. S., (2018). Metodolohiia indeksnoho analizu rivnia rozvytku informatsiinoho suspilstva [Methodology of index analysis of the level of development of the information society]. *Statystyka Ukrainy – Statistics of Ukraine*, 1, pp. 6–15, [in Ukrainian].

Lepeyko, T., Shcherbak, A., (2018). Determining factors to ensure the effective formation of the information process in the industrial enterprise management. *Development Management*, 16(4), pp. 88–97.

Osaulenko, O., Holubova, H. and Horobets, O., (2021). Implementing Bid Data in the Public Administration. Stratehiia rozvytku Ukrainy: finansovo-ekonomichnyi ta humanitarnyi aspekty: materialy VIII Mizhnarodnoi naukovo-praktychnoi konferentsii – Strategy of development of Ukraine: financial and economic and humanitarian aspects: materials of the VIII International scientific-practical conference. Kyiv, Informatsiino-analitychne ahenstvo, pp. 219–222, [in English].

Rosen, V. P., Reutsky, M. O. and Demchik, Ya. M., (2018). Zastosuvannia metoda holovnykh component dlia identyfikatsii holovnykh faktoriv vplyvu na velychynu elektrospozhyvannia [Application of the principal components method to identify the main factors influencing the amount of electricity consumption]. *Enerhetyka: ekonomika, tekhnolohii, ekolohiia: naukovyi zhurnal – Energy: economics, technology, ecology: a scientific journal*, No. 3 (53), pp. 81–87, [in Ukrainian].

Silverstein, A. B., (1987). Note on the parallel analysis criterion for determining the number of common factor or principal components. *Psychological Reports*, 61, pp. 351–354.

Zwick, W. R., Velicer, W. F., (1986). Comparison of five rules for determining the numberof components to retain. *Psychological Bulletin*, 99(3), pp. 432–442.