



Application of k-means clustering in psychological studies

Kyrylo Zakharov^a

^aInstitute of Social and Political Psychology, National Academy of Educational Sciences, Ukraine

Abstract ■ This article analyzes the scientific literature on k-means clustering and considers how this method is used in psychological studies, based on a sample of articles. It is shown that the majority of publications contains insufficient description of k-means clustering procedures, which can lead to non-reproducibility of research results. This article describes in detail the general steps to follow to perform a k-mean clustering analysis and the decisions to be taken in accordance with its use in the research practice. Recommendations for further improving the quality of scientific papers using k-means clustering are given.

Keywords ■ cluster analysis, k-means clustering, algorithms, data processing, validation of clustering solutions, meta-analysis.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

kirillzakharov13@gmail.com

KZ: 0000-0001-8396-4187

[10.20982/tqmp.12.2.p087](https://doi.org/10.20982/tqmp.12.2.p087)

Introduction

In recent years, k-means clustering has been increasingly used in psychological research. The main scope of this application is the construction of different classifications used to describe differences between groups of people. Usually, theoretical typologies are based on understanding of the object as a system and building representations of its structural levels (Нагибина, 2011). The main feature of such typologies is that they are created on some theoretical assumptions. Despite this, modern statistical methods allow to build typologies based on data. K-means clustering is one of these methods. It is commonly used in studies of emotional intelligence (Аршава & Носенко, 2011; Дерев'янку, 2009; Турецька, 2014), coping skills (Боснюк, 2010; Єсип, 2014) risky behavior (Бунас, 2012), in Medical Psychology (Бородавко, 2011; Завгородня & Арсен'єв, 2014; Зотова, 2012; Кліманська, 2013), Educational Psychology (Носенко & Труляев, 2013; Сотнікова-Мелешкіна, 2012), Psychology of Communication (Виноградов, 2013; Петренко, 2012a, 2012b) and other fields.

Problem formulation. The k-means clustering method allows scientists to build empirical classifications, which are based on experimental data. Despite the relative simplicity of the method, its practical application in studies and its description in publications has many defects, which subsequently can lead to low stability of ob-

tained classifications and non-reproducible results. Often, the approach used in such publications is to select 2 to 5 tests for an investigation of a psychological phenomenon and split participants into groups according to their quantitative results using cluster analysis. Differences between the obtained clusters are evaluated with statistical tests and statistically significant differences get meaningful interpretation. Sadly, publications using k-means clustering lack a number of important points, such as justification of the chosen number of clusters and sample size, selection of variables and data preprocessing, the choice of initial cluster centers, identification of software or algorithms used, a verification of the validity of the clustering solution. Since the cluster analysis can produce different solutions based on the same data (Ким, Мюллер & Клекка, 1989), each step of the method should be carefully described.

The present study aims are to review existing descriptions of k-means clustering application in a sample of articles and to develop recommendations for improvement of the quality of such studies in future.

General Information on K-means Clustering

Application of cluster analysis in psychology began in the end of the 30-ies of the XXth century. According to Ким et al. (1989), it is considered that the first publications were made by K. Trion. The increased use of cluster analysis was triggered by the development of computing power and ability to process large databases.



The goal of cluster analysis is to find a structure in a given data set - clusters. Objects within a cluster should be similar to one another and be different from the objects of other clusters (Буреева, 2007; Зорина & Слонимская, 2008; Ким et al., 1989). The final partition of data must satisfy a certain optimality criterion, expressing the desirability of different partitions (Буреева, 2007).

Generally, k-means clustering algorithm examines objects as a collection of points in Euclidean space, although there are variations of algorithm, which use non-Euclidean metrics (Cao, Wang, Ma, & Ding, 2012). In this article we assume that Euclidean distances are used to calculate similarity between objects. According to this, k-means clustering's objective is to minimize the variance of the data within cluster (Дубровская & Князев, 2011). An optimal partition is achieved by minimizing the amount of squared distances from the points to the cluster centers. To this end, an iterative procedure is used (Xu & Wunsch, 2008). The method's name was suggested by MacQueen in 1967 (MacQueen, 1967), though k-means algorithms were already published by Forgy (Forgy, 1965) and developed by Lloyd (Lloyd, 1982). Lloyd proposed his algorithm in 1957, but the article was published outside of Bell Laboratories only in 1982. In 1967 Friedman and Rubin made another publication on k-means (Friedman & Rubin, 1967).

Morissette and Chartier (2013) stated that k-means clustering is most often used:

1. In exploratory data analysis and creation of classification in applied researches and data mining applications.
2. For data reduction (reduction of data complexity).
3. As the initial step of more complex and computationally intensive algorithms, which gives an approximate division of data as a new starting point - reducing the noise in the dataset.

General Principle of the Algorithm

In general, k-means clustering is implemented with the following steps:

1. K initial cluster centers are chosen.
2. Each object is assigned to the nearest cluster center.
3. Cluster centers are recalculated using the current distribution of points in clusters.
4. If the convergence criterion is not satisfied, the second step is returned. A typical convergence criterion is the satisfying a minimum threshold for changes in the location of new cluster centers or finding a minimal decrease in the mean square error (Abonyi & Feil, 2007).

The distance between objects is most often calculated with some form of Euclidean distance (Daniel & Goldstein, 2013).

We have already mentioned that k-means clustering al-

gorithm has several different variations. Despite the fact that the goal of clustering is to find a structure, in fact, each clustering method adds a certain structure in the data (Ким et al., 1989). Therefore, it is important to know the differences between the various algorithms of k-means clustering and their features. The three most popular algorithms are by Forgy and Lloyd, MacQueen and Hartigan and Wong (Morissette & Chartier, 2013). One should note that there are also newer specific algorithms such as genetic k-means, spherical k-means, kernel k-means, etc. (Basel, Rui, David & Asoke, 2015).

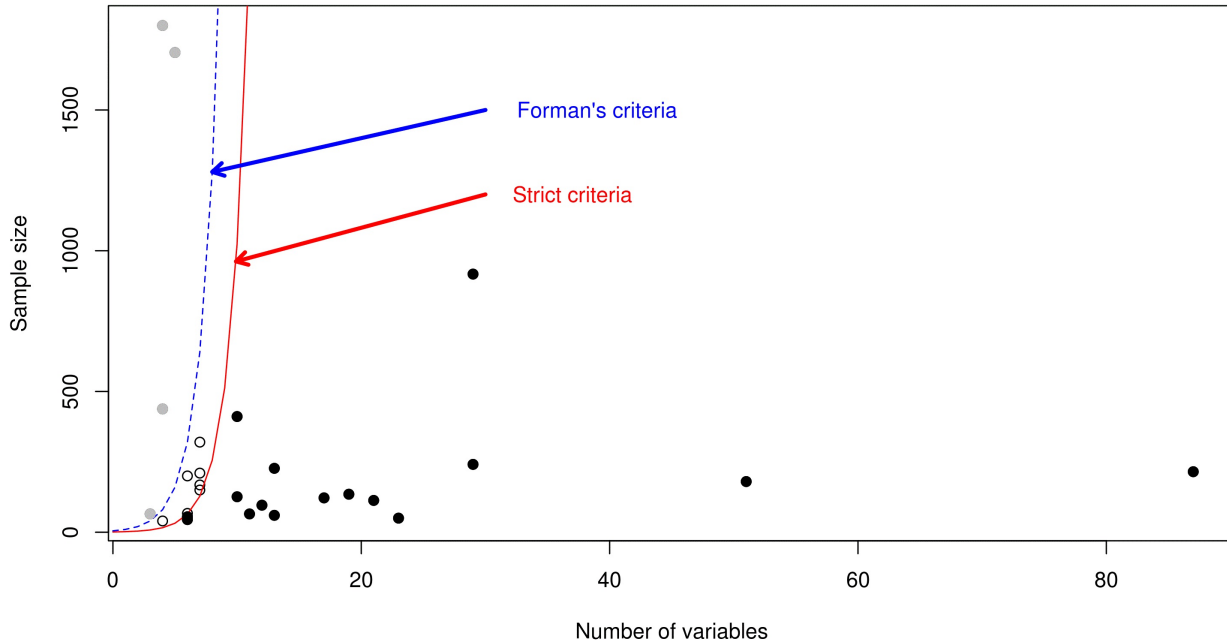
The Forgy and Lloyd algorithm uses a serial model of centroids, where each centroid is understood as the geometric center of an objects set. The algorithm uses all the objects at the same time at each step for cluster transformation. It has different variants regarding clusters initialization. The Forgy and Lloyd algorithm involves a random selection of N initial clusters centroids from an existing set of objects. On the other hand, the random division method randomly generates a set of clusters, then these sets are used to determine the initial centroids. The MacQueen algorithm is the most popular, and the main difference with previous ones is that centroid locations are recalculated after each new assignment of an object into the cluster. The algorithm created by Hartigan and Wong tries to find a separation of objects as to ensure a minimum variance value within cluster. Note that a single object in this case may be assigned to another subspace, even if it belongs to the subspace of the nearest centroid. Thus the Forgy and Lloyd and McQueen algorithms try to optimize the total sum of squares, while the Hartigan and Wong algorithm - the within-cluster sum of squares (Morissette & Chartier, 2013).

k-means clustering is one of the most famous and popular algorithms, partly due to the simplicity of its implementation (Xu & Wunsch, 2008). In addition, the method does not require the calculation and storage of the matrix of distances or similarities between objects (Буреева, 2007). In the context of big data mining, the algorithm takes advantage of its parallelization, such that computing time and memory requirements increases linearly. The disadvantages of k-means clustering are the following:

- The algorithms always converge, but can result in a local minimum and a sub-optimal separation of data;
- All three methods tend to create clusters of equal size, even if it is not the best reflection of "natural" groups;
- the Forgy and Lloyd algorithm can create empty clusters, and the algorithms by McQueen and Hartigan and Wong produce solutions which are sensitive to the order in which points are presented (Morissette & Chartier, 2013);
- the choice of different initial centers leads to different



Figure 1 ■ Scatterplot of relationship between the number of variables used for k-means clustering and sample size in the sample of articles. Lines reflect Forman’s and the strict criteria. Gray dots represent articles which meet both criteria, white dots are articles which meet Forman’s criterion, but don’t meet the strict criterion and black dots are articles which do not meet both criteria



decisions;

- the algorithms are sensitive to outliers and noisy data (Xu & Wunsch, 2008).

Materials and Methods

This study was conducted as part of the author’s thesis “The construction of social representations of Ukraine’s political future”. It was decided to use k-means clustering as data processing method. So the author reviewed current state of k-means application in psychological researches. We selected only scientific publications on psychology written by Ukrainian scientists, who used k-means clustering. This involved a search for the relevant publications on the official site of the Vernadsky National Library of Ukraine and the search results of the Google Scholar search engine. The selection of the publications was made by requesting «psychology» with «К середніх», «К-средних» and «k-means» to find all papers in Ukrainian and Russian languages. We checked all authors to eliminate Russian articles. 36 scientific publications were found. The complete list of works is presented in Appendix A. The material was gathered in March 2015. The earliest article publica-

tion date is 2009. That fact that no work prior to 2009 was found is explained by the fact that it is the year when Scientific Publishing Infrastructure "Scientific Periodicals of Ukraine" was introduced. So some earlier publications can be overlooked.

Table 1 presents the number of articles in the sample by years they were published.

We note an increase in the number of publications where k-means clustering was used. We conducted a content analysis to find justification of the number of groups for clustering, sample sizes, selection of variables and presence or absence of data pre-processing, the choice of initial cluster centers, the software or the algorithm description and the validation of cluster analysis solutions in the publications.

This sample containing articles only from one country and published in Ukrainian and Russian languages poses a problem for generalizability. So the following empirical findings should be considered as examples which illustrate issues related to k-means application.

**Table 1** ■ The number of articles in the sample by years

Year	2009	2010	2011	2012	2013	2014	Total
Number of articles	1	3	3	11	9	11	36

Results

We found that the description of how cluster analysis was performed was mostly limited to a brief mentioning of the method, comparison of obtained groups with statistical tests and the interpretation of statistically significant differences between the clusters. Most of the time, many of the key steps of k-means application were not given to the reader.

Number of Variables and Sample Size

One of the tasks in designing and planning a research is the choice of method and the determination of the sample size.

With regard to k-means clustering, it should be reminded that a researcher will receive a clustering solution regardless of the sample size and the number of variables (Xu & Wunsch, 2008). Mooi and Sarstedt also indicate that there are no generally accepted rules regarding the minimum required sample size, relationship between the objects and the number of variables for clustering (Mooi & Sarstedt, 2011). However, it does not mean that the cluster analysis is a universal method for any data. Forman recommends a sample size of at least 2^k , where k is the number of variables used for clustering. There is also a stricter recommendation that the preferred sample size should be at least $5 \cdot 2^k$ (Dolnicar & Grün, 2008).

The Ukrainian publications in the studied sample, who used k-means clustering, analyzed from 1 to 9 psychological tests and quizzes and the number of variables used for clustering ranged from 3 to 87 variables (median - 10.5), whereas sample sizes ranged from 40 to 1800 respondents (median value - 136.5). Studies with a small number of variables typically included clustering on results of 1-2 psychological tests or on results of the preliminary factor analysis. Studies with a large number of variables, generally used all the raw data gathered in the survey.

A scatterplot of association of the sample sizes and numbers of variables for clustering in the articles in the sample is shown in Figure 1. The plot shows all the studies which were reported with the number of variables and sample size. Some publications report several different studies (e. g. pilot and complementary studies) using k-means clustering and some publications have no description of sample sizes, so only 27 of 39 of studies are represented in the Figure.

The plot shows that most of the studies do not meet

even the minimum criterion. The Forman's criterion was satisfied only in 11 of the observed studies (28.2%) and the strict criterion - in 4 studies (10.3%). It should be noted that we may underestimate the number of variables, since researchers do not always clearly indicate whether they used only the total score or the results of sub-scales for clustering.

Determining the Number of Clusters

The algorithms of k-means all assume that the researcher has prior knowledge and can set the number of clusters. But the actual number of groups in a set of objects is often unknown (Xu & Wunsch, 2008). Researchers have several different ways to determine the number of clusters:

- using preliminary information (Зорина & Слонимская, 2008);
- determining the number of clusters empirically (Mooi & Sarstedt, 2011; Xu & Wunsch, 2008; Зорина & Слонимская, 2008);
- visually determining the number of clusters (Everitt, Landau, Leese & Stahl, 2011).

The first way implies that the researcher has theoretical knowledge about the number of groups that should be received as a result of the cluster analysis, or has some assumptions about the future use of the resulting partition. So, the following comparison group analysis may involve additional requirements on the number of objects per clusters.

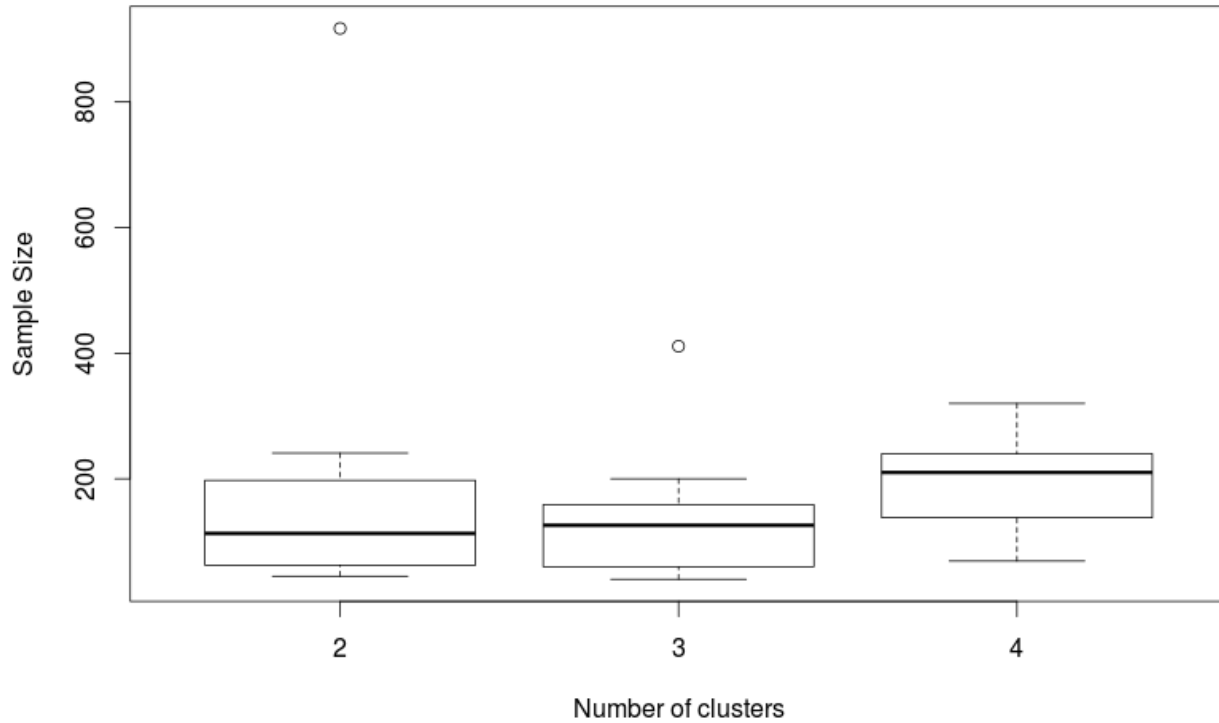
Popular strategies in the empirical determining of the number of clusters are to find it on the basis of:

- the size of resulting clusters, i. e. number of objects in clusters;
- the «Scree»-type diagram (a plot which reveals the relationship between the number of clusters and the sums of within-cluster variance; Зорина & Слонимская, 2008);
- the results of hierarchical cluster analysis;
- the statistical comparison of different clustering solutions with the number of clusters from K_{min} to K_{max} (minimal and maximum number of clusters respectively; Daniel & Goldstein, 2013).

Visually determining the number of clusters can be made with a set of histograms of variables distributions, as well as two-dimensional scatter plots. In case of very large number of data, a projection of data in spaces with smaller dimension can be built using principal component analysis, factor analysis or multidimensional scaling (Everitt et



Figure 2 ■ Boxplot of sample sizes against different number of clusters



al., 2011).

There are also other formal techniques to reduce the degree of subjectivity in taking a decision on the number of clusters. Milligan and Cooper (1985) describe 30 of these techniques.

The Ukrainian scientists from the sample often decide to partition data in 2 to 4 clusters. The distribution of the number of clusters is shown in Table 2.

Of the 36 publications, only 9 include the reasoning behind the chosen number of clusters. Two researchers proceeded from the interpretability of the obtained clustering solutions and 7 used empirical methods.

Among articles with empirical methods, three authors reported the application of preliminary hierarchical clustering to determine the number of groups for k-means (Боснюк, 2010; Завгородня & Арсеньев, 2014; Шамлян, 2014). This is not a universal solution as well, because it raises further questions about the kind of metrics used to estimate the distance between objects, the method of clustering («the nearest neighbor», «Ward’s method», etc.), and which distance between cluster should be used to allocate groups. However, only in the paper by Bosniuk, where a dendrogram is provided, we can find that the author used

the Euclidean distance and the Ward’s method (Боснюк, 2010). Two other studies just mentioned preliminary hierarchical and further k-means clusterings.

In the study where the largest number of clusters was selected (18 groups) the author proposed a measure of where data can be recreated from the clusters as a criterion for selecting the number of groups, namely, matching correlations between variables of raw data and correlations between the estimated centroids of clusters (Виноградов, 2013).

One of the authors tried to find three clusters in data, but one group was too small, so he identified only two clusters instead (Рафа, 2012). Another article reported that the number of clusters was identified according to minimal number of k-means iterations (Зотова, 2012). Grys reported that she experimented with the data to identify the number of clusters, but did not specified the content of such experiments (Грись, 2014).

The general trend is that with increasing sample size, researchers tend to choose a larger number of clusters (Spearman’s $\rho(37) = .343, p \leq .05$). Figure 2 shows boxplots of the sample sizes distribution depending on the number of clusters. The four studies with 5 or more clus-

**Table 2** ■ The number of clusters in the works in the sample

Number of clusters	2	3	4	5	7	18
Number of studies	12	13	10	2	1	1
Percent of studies	30.8%	33.3%	25.6%	5.1%	2.6%	2.6%

ters are not presented on the plot.

Selection of Variables and Data Preprocessing

Usually, Euclidean distance is used as a measure of the distance between objects. It is important to take into account the following consequent restrictions. The use of this metric is useful when:

- observations are taken from the population, with multivariate normal distribution, the variables are mutually independent and have the same variance;
- variables are homogeneous in their physical meaning and are equally important for classification;
- all variables have the same units of measurement (Дубров, Мхитарян & Трошин, 2003).

In addition, most algorithms do not give successful results in case of «bridges» between clusters or in case clusters are not spherical (Дубровская & Князев, 2011), that is, when there are many objects that are located between two groups.

Because of these problems, a researcher has to provide exploratory data analysis and make the following decisions:

- deleting outliers if any (including multidimensional outliers);
- standardizing or weighting the data to provide «identical» units of measurement;
- dealing with correlated variables – change the variables included in the analyses or use data reduction.

Direct application of k-means clustering on raw data can lead to the situation when the variables with higher range or standard deviation may be decisive for classification (Завгородня & Арсеньев, 2014). However, the decision about outlier removal must be carefully considered. In some situations, outliers can belong to subgroups, which merely are not represented enough in the sample (Ким et al., 1989).

None of the considered publications in the sample reported presence or absence of outliers.

Since in many applications of k-means clustering, and especially in psychology, variables describing the object are measured in different units, a common practice is to standardize the variables. Often, calculation of the Z-score is used, more rarely variants of the transformation by the deviation from the median or by the range of data are presented (Everitt et al., 2011; Буреева, 2007). Standardization of variables assumes that “the importance of a vari-

able decreases with increasing variability” (Everitt et al., 2011, p. 67) and the effects of each variable on the clustering can vary from one data set to another (Ким et al., 1989).

There is no unambiguous position on the variable standardization. In 1996, Schaffer and Green conducted a study on ten real data sets and showed that data should not be standardized in any way. In 2004, Steinley criticized the study, showing that the results are unreliable because the real structure of the data sets was unknown to researchers, and indicated that the optimal scale is standardization on values range (as cited in Steinley, 2006, p. 11).

Among the studies in the sample, we found only three which clearly reported the use of data standardization, but no justification was provided (Боснюк, 2010; Бунас, 2012, 2014).

The last question with regard to the preliminary data processing is the absence of highly correlated variables. Such variables will excessively affect the clustering solution (Mooi & Sarstedt, 2011). During the calculation of the objects similarity, this leads to reweighting of the variables, and their influence will be equivalent to a single variable having been enlarged several times (Ким et al., 1989). A Common solution is preliminary data reduction using principal component analysis (PCA) or the exploratory factor analysis (EFA).

Nevertheless, Dolnicar and Grün (2008) indicate that such data transformation leads to other problems:

- the obtained clusters are based on the transformed variables, and there is no guarantee that the next set of data reflect the same principal component structure;
- the transformation implies the loss of the initial data variance, which may lead to distorted solutions;
- elimination of groups of variables with small factor or component loads can lead to losing of important information about the niche segments of the data;
- interpretation of the clustering solutions can be difficult.

Three papers in the sample reported preliminary transformation of data (Белікова, 2012; Меднікова, 2013; Нахабіч, 2013). Such practice is observed in studies with large samples (1704 and 1800 respondents – in the papers, which reported sample sizes).



Selection of Initial Cluster Centers

The result of k-means clustering can also be determined by the choice of initial cluster centers. The untoward choice of them can lead to a suboptimal clustering solution (Михалевич & Примина, 2006).

There are different possible solutions to that problem:

- to appoint initial cluster centers on the basis of existing knowledge and theories;
- to use centers of hierarchical clustering partitions as initial centers (Daniel & Goldstein, 2013). In case of a large amount of data, this method is not optimal computationally;
- to limit the choice of initial centers to areas with high density of data (Steinley, 2006);
- to reassign initial cluster centers using the bootstrap procedure (Everitt et al., 2011);
- to randomly assign K cluster centers or make a large number of random partitions (over 5000 as recommended by Steinley, 2003) into K groups, then on the basis of such partitions define initial cluster centers and select the partition, which will provide the minimum value of a specific criterion (Abonyi & Feil, 2007; Everitt et al., 2011).

None of the papers in the sample consider the question of the choice of initial cluster centers.

Differences in Algorithms and the Use of Various Software

After data preprocessing, a researcher must select an algorithm for k-means clustering. The Publication Manual of the American Psychological Association (2009) recommends to give the proper name of the software used along with the version number in text. And this is important given that different software applies various algorithms of k-means clustering by default:

- IBM SPSS package uses the Lloyd algorithm by default and chooses the first k elements as the initial cluster centers (Technote (troubleshooting), n.d.);
- Statistica uses the Lloyd algorithm by default and selects the first k elements of the data set as the initial centers (STATISTICA Help, n.d.);
- SAS has a function FASTCLUS (the implementation of the McQueen algorithm) and PROC FASTCLUS (the implementation of the Hartigan and Wong algorithm); the initial centers are selected with decreasing density of the data (SAS Institute Inc., 2009);
- Mathematica has a function FindClusters, which implements an alternative clustering algorithm, called K-medoids. This algorithm is equivalent to the Forgy and Lloyd algorithm, but it uses objects from a set of data as the cluster centers rather than the arithmetic mean of

cluster objects (Morissette & Chartier, 2013);

- the function «k means» in Matlab uses a serial algorithm in the first phase, and then applies an iterative algorithm in the next phase (Morissette & Chartier, 2013);
- the library Scikit-learn for the Python programming language applies the Lloyd algorithm and the initial cluster centers are determined by a special algorithm «k-means++», which speeds up the convergence (Pedregosa et al., 2011);
- the statistical programming language R has the function «kmeans», which uses the Hartigan-Wong algorithm by default, the initial cluster centers are selected randomly.

None of articles in the sample directly mentioned specific k-means algorithm. Among the observed papers of the sample there are only two (5.6%) which reported the use of the statistical programming language R, three (8.3%) – use of Statistica and ten cases (27.8%) – use of SPSS. 58.3% of the publications did not report any software or specific algorithm used to implement k-means clustering.

Analysis of the Results of K-means Clustering

After application of k-means clustering, the Ukrainian psychologists investigated the peculiarities of the obtained groups. Usually they formulated a number of hypotheses about the equality of the estimated parameters in groups and applied relevant statistical tests. Table 3 shows the distribution of commonly applied methods.

As you can see, most researchers have published the values of the obtained centers of clusters or present the plots of centroids. In most cases, the study ends with reporting significant differences between the centers of the obtained clusters. All variables including those which were used for clustering are compared and the differences found are reported. Frequently the scientists used parametric statistics (the Student's t-test, ANOVA), but did not provide information whether the data met the requirements of those tests or correction for multiple comparisons, which leads to an underestimation of the likelihood of the type I statistical error.

Also, not all the papers raise the issue of validation of clustering solutions.

It is important to realize that there are situations when data have no natural groups (structure), and then output of k-means algorithm can be a meaningless artifact (Xu & Wunsch, 2008). Since a researcher usually has no prior knowledge about the real structure of the data, there is a danger of accepting any clustering solution, even if there are no «real» clusters in the data (Everitt et al., 2011). In order to avoid subjective assessment of a «reality» of clusters, statistical tests can be used (Daniel & Goldstein, 2013). Evaluating clustering solutions can be viewed in terms of

**Table 3** ■ The number of papers with different statistical methods for the analysis of clustering solutions

Method	The number of papers	% of papers
Descriptive statistics (mean, standard deviation, etc.)	30	83.3%
Student's t-test for validation of a clustering solution	18	50.0%
Analysis of variance (ANOVA) for validation of a clustering solution	12	33.3%
Phi coefficient for validation of a clustering solution	3	8.3%
Mann–Whitney U test for validation of a clustering solution	2	5.6%
Kruskal-Wallis H Test for validation of a clustering solution	1	2.8%
The application of factor analysis to the results of clustering	1	2.8%

reliability and validity of the clustering solution, where the first term refers to the stability and reproducibility of the solution, and the second term refers to the assessment of whether there is a real structure in the observed data or not.

Reliability is closely related to the stability of the solution over time, and it can be assessed by carrying out a similar study, repeated after a period of time (Mooi & Sarstedt, 2011).

Nevertheless, there are recommendations about assessment of the current result. For example, If the data set is sufficiently large, it can be randomly divided into two parts and each subsample must be clustered separately with the same parameters of clustering (Mooi & Sarstedt, 2011).

Some papers recommend to use various clustering algorithms on the data or other distance measures and select the algorithm that shows the most stable solutions (including the use of k-means as well as hierarchical cluster analysis; Mooi & Sarstedt, 2011; Дубровская & Князев, 2011). This approach is possible, but it can be misleading since each of the cluster analysis algorithms not only reveals the structure of the data, but also adjusts the data under certain predefined model (Вэн, 1980).

We did not find any authors checking the reliability of clustered solutions in the sample of papers.

With regard to assessing the validity of a clustering solution, it is possible to use external and internal criteria. Their meaning is disclosed in relation to the data used for validation - in the first case, the data that were not used in clustering is used for validation of clusters, in the second case – the data which were used in the partition are used again for validation.

Given an existing set of data and the obtained cluster structure, the external criterion compares the clustering solution with a predetermined structure based on the a priori information (Xu & Wunsch, 2008). For such information, you can use the structures obtained from expert opinions. Another path it to use an external variable that has not been used in the cluster analysis, but which is theoretically linked to variables used in clustering. There must

be significant differences between the groups concerning this variable (Mooi & Sarstedt, 2011).

We found the application of external criteria in various forms in 16 scientific publications of the sample. Most of them estimated the statistical differences between the clusters based on the results of psychological tests that were not used in the partition of the groups into clusters.

The internal criterion evaluates the cluster structure solely on the raw data used for clustering, without any external information. However, without a prior information on the presence of the structure in the data, the concept of the «lack of structure» in the dataset is not clear (such «lack of structure» can be one of the possible null hypotheses) and it does not define what should be used as a test to determine the presence of this structure (Ким et al., 1989).

Jane and Dubes proposed to apply statistical hypothesis testing, in which the null hypothesis, which states that the data has no structure or that the structure in the data is random. They identified the following three null hypotheses:

1. The hypothesis of a random position. H_0 : All accommodation of N points of the data in a particular region of the n -dimensional space are equally possible.
2. The hypothesis of a random graph. H_0 : All $N \times N$ matrix of similarity are equally possible.
3. The hypothesis of randomness of a cluster assignment. H_0 : All rearrangement of clusters on N objects are equally possible (cited by Xu & Wunsch, 2008).

The same authors noted that reasonable alternative hypotheses have not yet been developed, and mathematically useful definition of a «cluster structure» does not exist (Ким et al., 1989).

In practice, basic data distribution for the null hypothesis is created using statistical sampling techniques like Monte Carlo simulation and bootstrap. Such a basic allocation allows to determine a critical value of a statistical criterion which will be compared with some index of validity calculated on the real data (Xu & Wunsch, 2008).

For example, SAS uses an approach called Cubic Clustering Criterion (CCC) which compares the explained variance of the obtained clustering solution with the sum of ex-



plained variance of clustering solutions obtained by clustering uniformly distributed data in the space with the same number of measurements (Daniel & Goldstein, 2013).

Another option is the application of discriminant analysis on the results of cluster analysis. This approach provides visualization of solutions in the canonical independent axes, allows to see the number of objects that have been properly reclassified into groups based on the discriminant analysis algorithm, and also provides an alternative way to classify new objects (Daniel & Goldstein, 2013).

In the sample of articles, we found a widely used practice of applying an analysis of variance (ANOVA, MANOVA) for assessing cluster validity. We are to remind that the use of the Student's t-test was found in half of the papers and the analysis of variance in one-third of the papers.

Also such validity checks are proposed in a number of textbooks (Моої & Sarstedt, 2011; Буреєва, 2007; Пістунюв, Антонюк & Турчанінова, 2008). Furthermore, it is proposed to exclude the variables among which significant differences are not revealed from clustering (Пістунюв et al., 2008). ANOVA can be applied to any solutions given k-means clustering. The procedure is becoming popular because its results are always highly significant (Ким et al., 1989). Also we assume that the spread of such a practice was due to typical reports produced as the result of clustering by modern statistical software.

In particular, the output of k-means clustering in Statistica displays the results of ANOVA which tests differences between clusters (Буреєва, 2007). SPSS also provides an opportunity to request the analysis of variance (IBM SPSS Statistics Base 22, n.d.). Here we note that the documentation for both programs clearly indicates the limitations of this method. The documentation of Statistica describes k-means clustering as «a procedure reverse to the analysis of variance (ANOVA)» (STATISTICA Help, n.d.) and the User Guide of IBM SPSS clearly states that the provided F-statistic is subsidiary and its relative size only informs about the role of each variable in the process of the data partition (IBM SPSS Statistics Base 22, n.d.). In other words, the goal of k-means clustering is to create groups which differ from each other and have the minimal in-group variance, so it is not surprising that k-means clustering produces groups with small standard errors for most variables and statistical tests show high levels of significance.

In 1989, Kim et al. proposed to use the Monte Carlo method to justify clusters reliability. A researcher needs to generate a random dataset with the basic features corresponding to the characteristics of the actual data, but which does not contain clusters (in particular, generated data should inherit means, standard deviations and correlation matrix between the original variables). Then, the

actual and the generated data sets are used with the same clustering algorithms and both solutions are compared by appropriate methods (Ким et al., 1989). On the sets of generated data which do not contain clusters, the critical value of F-test may be calculated, i. e. there must be a value when the probability of obtaining the same or higher values of the criterion in the absence of clusters in the data would not exceed 5%. With this approach, the critical value of the F-test for the final clustering solution will be quite different from the critical value obtained only with the number of degrees of freedom.

Discussion

In the study, we carried out an overview of a sample of scientific publications in the field of psychology which used k-means clustering. It shows that most papers do not contain an adequate description of the method application that would allow to clearly reproduce these studies. Also, many studies do not estimate the validity of the obtained clustering solutions. We provided an analysis of the literature and an overview of the modern software products which implement popular algorithms of k-means clustering. On the basis of this work we recommend to improve k-means application in scientific psychological studies with the following steps:

1. Advantages and disadvantages of k-means clustering should be reconsidered regarding research design and obtained data.
2. A required sample size based on the number of variables that will be used in clustering must be planned in advance.
3. The procedures for selecting the number of clusters must be clearly specified.
4. The results of the exploratory data analysis and the description of raw data should be included in the paper. All the data preprocessing or lack of it must be mentioned.
5. It is necessary to exactly specify what software and which version of it was used in calculating the results of clustering. Also, default settings or all additional settings of the algorithm must be reported. If a researcher implemented an algorithm on his/her own, he/she must clearly indicate what specific algorithm was used and what the procedure for selecting the initial cluster centers was.
6. It is necessary to not only carry out a meaningful interpretation of the differences in the clusters, but also to assess the reliability and validity of the clustering solutions.

Given the fact that k-means clustering is becoming more widely used in modern psychological studies, these recommendations may be useful in future works.



References

- Abonyi, J. & Feil, B. (2007). *Cluster analysis for data mining and system identification*. Blackwell.
- Association, A. P. (2009). *Publication manual of the American psychological association, 6th edition*. DC: American Psychological Association: Washington.
- Basel, A.-J., Rui, F., & Asoke, K. N. (2015). *Integrative cluster analysis in bioinformatics*. Wiley.
- Cao, H., Wang, P., Ma, R., & Ding, J. (2012). On non-euclidean metrics based clustering. In *Intelligent science and intelligent data engineering* (pp. 655–663). Springer.
- Daniel, N. A. & Goldstein, G. (2013). *Cluster analysis in neuropsychological research*. New York: Springer Publishing Company.
- Dolnicar, S. & Grün, B. (2008). Challenging “factor–cluster segmentation”. *Journal of Travel Research*, 47(1), 63–71. doi:10.1177/0047287508318910
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley. doi:10.1002/9780470977811
- Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768–769. Retrieved from <http://www.jstor.org/stable/2528559>
- Friedman, J. H. & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62, 1159–1178. doi:10.1080/01621459.1967.10500923
- IBM SPSS statistics base 22. (n.d.). Retrieved November 5, 2015, from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/22.0/en/client/Manuals/IBM_SPSS_Statistics_Base.pdf
- Inc, S. I. (2009). *Sas/stat(r) 9.2 user's guide, second edition*. Cary: NC: SAS Institute Inc.
- Lloyd, S. P. (1982). Least square quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. doi:10.1109/TIT.1982.1056489
- MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th berkeley symposium on mathematical statistics and probability 1* (pp. 281–297). University of California Press.
- Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159–179. doi:10.1007/BF02294245
- Mooi, E. & Sarstedt, M. (2011). *A concise guide to market research. the process, data, and methods using ibm spss statistics*. Berlin: Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-12541-6
- Morissette, L. & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1), 15–24. Retrieved November 5, 2015, from <http://www.tqmp.org/RegularArticles/vol09-1/p015/p015.pdf>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Statistica help. (n.d.). Retrieved November 5, 2015, from <http://documentation.statsoft.com>
- Steinley, D. (2003). Local optima in k-means clustering: what you don't know may hurt you. *Psychological methods*, 8(3), 294–304. doi:10.1037/1082-989X.8.3.294
- Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1), 1–34. Retrieved November 5, 2015, from <http://onlinelibrary.wiley.com/doi/10.1348/000711005X48266/abstract>
- Technote (troubleshooting). (n.d.). K-means cluster (quick cluster) results sensitive to case order. Retrieved from <http://www-01.ibm.com/support/docview.wss?uid=swg21476878>
- Xu, R. & Wunsch, D. (2008). *Clustering*. John & Sons: Wiley. doi:10.1002/9780470382776
- Аршава, І. Ф. & Носенко, Е. Л. (2011). Суб'єктивне благополуччя і його індивідуально-психологічні та особистісні кореляти. *Вісник Дніпропетровського університету. Сер. : Педагогіка і психологія*. 20(18), 3–9. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/vdupp_2012_20_18_3.pdf
- Белікова, Ю. В. (2012). Моделі досягнення успіху у професійній діяльності в залежності від рівня емоційного інтелекту (EQ). *Сучасні суспільні проблеми у вимірі соціології управління : збірник наукових праць ДонДУУ. Серія "Соціологія"*, XIII(257), 259–266. Retrieved May 11, 2015, from <http://repository.hneu.edu.ua/jspui/bitstream/123456789/4512/1/%D0%91%D0%B5%D0%BB%D0%B8%D0%BA%D0%BE%D0%B2%D0%B0%20%D0%AE.%D0%92.%20%D0%9C%D0%BE%D0%B4%D0%B5%D0%BB%D1%96%20%D0%B4%D0%BE%D1%81%D1%8F%D0%B3%D0%BD%D0%B5%D0%BD%D0%BD%D1%8F%20%D1%83%D1%81%D0%BF%D1%96%D1%85%D1%83%20%D0%B2%20%D0%BF%D1%80%D0%BE%D1%84%D0%B5%D1%81%D1%96%D0%B9%D0%BD%D1%96%D0%B9%20%D0%B4%D1%96%D1%8F%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%96%20%D0%B7%D0%B0%D0%BB%D0%B5%D0%B6%D0%BD%D0%BE%20%D0%B2%D1%96%D0%B4%20%D1%80%D1%96%D0%B2%D0%BD%D1%8F%20%D0%B5%D0%BC%D0%BE%D1%86%D1%96%D0%B9%D0%BD%D0%BE%D0%>



[B3%D0%BE%20%D1%96%D0%BD%D1%82%D0%B5%D0%BB%D0%B5%D0%BA%D1%82%D1%83%20\(EQ\).pdf](#)

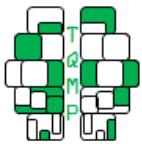
- Бородавко, О. О. (2011). Дослідження психометричних характеристик методики діагностики міжособистісних відносин лікаря та пацієнта в процесі лікування. *Український вісник психоневрології*, 19(4), 115–119. Retrieved May 22, 2015, from http://nbuv.gov.ua/j-pdf/Uvp_2011_19_4_31.pdf
- Боснюк, В. Ф. (2010). Відмінності у використанні захисно-копінгової поведінки в залежності від емоційного ресурсу особистості. *Проблеми екстремальної та кризової психології. Збірник наукових праць*, 11, 23–30. Retrieved May 11, 2015, from http://edu-mns.org.ua/nmc/123/EKP_11.pdf
- Бунас, А. А. (2012). Схильність до ризику в системі особистісних детермінант. *Збірник наукових праць Інституту психології імені Г. С. Костюка Національної АПН України. Проблеми загальної та педагогічної психології*, 24(6), 51–59. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/pzpp_2012_24_6_8.pdf
- Бунас, А. А. (2014). Прогностична компетентність як соціально значуща властивість особистості при прояві схильності до ризикованої поведінки. *Вісник Одеського національного університету. Психологія*, 19(2), 52–61. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Vonu_psi_2014_19_2_8.pdf
- Буреєва, Н. Н. (2007). *Многомерный статистический анализ с использованием pnp "statistica". учебно-методический материал по программе повышения квалификации «применение программных средств в научных исследованиях и преподавании математики и механики»*. Нижний Новгород: Нижегородский государственный университет им. Н.И. Лобачевского.
- Виноградов, О. Г. (2013). Номенклатура сигнатур емоційного реагування на ситуації публічного виступу. *Актуальні проблеми соціології, психології, педагогіки : зб. наук. пр.* 18, 119–126. Retrieved May 11, 2015, from <http://www.socd.univ.kiev.ua/sites/default/files/library/elopen/aktprob.18.119.pdf>
- Вэн, Р. Д. (1980). *Классификация и кластер*. Москва: Мир.
- Грись, А. М. (2014). Психологічні бар'єри студентів-психологів у процесі їх підготовки дороботи з соціально дезадаптованими неповнолітніми. *Проблеми загальної та педагогічної психології : [зб. наук. праць Інституту психології імені Г. С. Костюка НАПН України / за ред. С.Д.Максименка]*. XVI(2), 289–298. Retrieved May 11, 2015, from <http://lib.iitta.gov.ua/1692/1/%25344.Pdf>
- Дерев'янка, С. П. (2009). *Емоційний інтелект як чинник соціально-психологічної адаптації особистості до студентського середовища* (Автореф. дис. канд. психол. наук: 19.00.05, Ін-т соц. та політ. психології АПН України, Київ). Retrieved May 11, 2015, from <http://mydisser.com/en/catalog/view/252/257/18207.html>
- Дубров, А. М., Мхитарян, В. С., & Трошин, Л. И. (2003). *Многомерные статистические методы*. Москва: Финансы и статистика.
- Дубровская, Л. И. & Князев, Г. Б. (2011). *Компьютерная обработка естественно-научных данных методами многомерной прикладной статистики: учебное пособие*. Томск: ТМЛ-Пресс.
- Єсип, М. З. (2014). Прагматична оцінка релігійних способів долаючої поведінки. *Прагматична оцінка релігійних способів долаючої поведінки*, 24, 314–327. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Pspl_2014_24_28.pdf
- Завгородня, Н. І. & Арсен'єв, Н. І. (2014). Результати систематизації чинників формування розладів адаптації у жінок, що народили недоношену дитину. *Ліки України плюс*, 3, 57–60. Retrieved May 11, 2015, from http://repo.knmu.edu.ua/bitstream/123456789/7331/1/lu_PLUS_3_2014_.pdf
- Зорина, Т. Г. & Слонимская, М. А. (2008). *Маркетинговые исследования: учеб. пособие*. Минск.
- Зотова, Л. Н. (2012a). Профиль невротических склонностей как личностный фактор, определяющий внутреннюю картину здоровья у студентов. Retrieved May 11, 2015, from <http://journals.hnpu.edu.ua/ojs/psycho/article/view/1167/1173>
- Зотова, Л. Н. (2012b). Профиль невротических склонностей как личностный фактор, определяющий внутреннюю картину здоровья у студентов. *Вісник Харківського національного педагогічного університету імені Г.С. Сковороди. Психологія*, 44(1), 79–92. Retrieved May 11, 2015, from <http://journals.hnpu.edu.ua/ojs/psycho/article/view/1167/1173>
- Ким, Д. О., Мюллер, Ч. У., & Клекка, У. Р. (1989). *Факторный, дискриминантный и кластерный анализ: пер. с англ.* Москва: Финансы и статистика.
- Кліманська, М. Б. (2013). Психологічні особливості використання гумору медичними працівниками. *Проблеми сучасної психології*, 21, 225–236. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Pspl_2013_21_22.pdf



- Меднікова, Г. І. (2013). *Характеристика самоактуалізації студентів з різними варіантами структури ставлення до себе*. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Nvmdups_2013_2_10_41.pdf
- Михалевич, І. М. & Примина, С. П. (2006). *Применение математических методов при анализе геологической информации (с использованием компьютерных технологий): учеб. пособие*. Иркутск: Иркут. гос. ун-т.
- Нагибина, Н. Л. (2011). Психологические классификации: история, методология, проблемы. психологические и психоаналитические исследования. *Психологические и психоаналитические исследования 2010-2011*, 10–24. Retrieved May 11, 2015, from http://www.iidp.ru/inpsycho/paper/files/collection_2010_2011.pdf
- Нахабич, М. А. (2013). "ціннісна" мапа України в системі аксіологічних досліджень. *Вісник Одеського національного університету. Соціологія і політичні науки*, 18(2(1)), 118–130.
- Носенко, Е. Л. & Труляев, Р. О. (2013). Психологічний механізм впливу позитивних цінностей педагога на суб'єктивне благополуччя учнів. *Психологічний механізм впливу позитивних цінностей педагога на суб'єктивне благополуччя учнів*, 21(19), 102–109. Retrieved May 11, 2015, from http://irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?C21COM=2%5C&I21DBN=UJRN%5C&P21DBN=UJRN%5C&IMAGE_FILE_DOWNLOAD=1%5C&Image_file_name=PDF/vdupp_2013_21_19_16.pdf
- Петренко, Т. П. (2012а). Діагностичний потенціал невербальних засобів спілкування у дослідженні "позитивних цінностей" та темних рис особистості. *Вісник Дніпропетровського університету. Сер. : Педагогіка і психологія*, 20(18), 129–134. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/vdupp_2012_20_18_22.pdf
- Петренко, Т. П. (2012b). Невербальні засоби спілкування як аспект імпліцитної діагностики особистісних характеристик. *Вісник Національної академії Державної прикордонної служби України*, 5. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Vnadps_2012_5_46.pdf
- Пістунов, І. М., Антонюк, О. П., & Турчанінова, І. Ю. (2008). *Кластерний аналіз в економіці: навч. посібник*. Дніпропетровськ: Національний гірничий університет.
- Рафа, М. Б. (2012). Оптимальні особливості суб'єктивної реконструкції особистого досвіду як детермінанти життєвої успішності в юнацькому віці. *Вісник Харківського національного педагогічного університету імені Г. С. Сковороди. Психологія*, 42(2), 134–146. Retrieved May 11, 2015, from [http://nbuv.gov.ua/j-pdf/VKhnpu_psykhol_2012_42\(2\)_18.pdf](http://nbuv.gov.ua/j-pdf/VKhnpu_psykhol_2012_42(2)_18.pdf)
- Сотнікова-Мелешкіна, Ж. В. (2012). "адаптаційний профіль" старшокласників із різним характером професійної готовності. *Вісник Харківського національного університету імені В. Н. Каразіна. Сер. : Медицина*, 1024(24), 42–51. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/VKhM_2012_1024_24_9.pdf
- Турецька, Х. І. (2014). Особливості емоційного інтелекту працівників сфери ІТ, схильних до емоційного вигорання. *Наук. пер. вид. "Український психолого-педагогічний науковий збірник"*. *Наук. журнал*, 2, 141–146.
- Шамлян, К. М. (2014). Особливості вольової організації особистості студентів. *Проблеми сучасної психології*, 23, 707–719. Retrieved May 11, 2015, from http://nbuv.gov.ua/j-pdf/Pspl_2014_23_64.pdf

Appendix A: The sample of articles considered in this study

- Аршава, І. Ф., & Бунас, А. А. (2013). Вивчення компонентів Я-концепції та самовідношення у осіб з ризикованою поведінкою. *Вісник Харківського національного університету імені В. Н. Каразіна. Серія: Психологія*, 1065(52), 7–11. Retrieved from http://nbuv.gov.ua/j-pdf/VKHiPC_2013_1065_52_3.pdf
- Аршава, І. Ф., & Кутувий, К. П. (n.d.). Комп'ютерна діагностика емоційної стійкості людини, яка працює в екстремальних умовах професійної діяльності. *Вісник Національної академії Державної прикордонної служби України*, 5. Retrieved from http://nbuv.gov.ua/j-pdf/Vnadps_2012_5_29.pdf
- Аршава, І. Ф., & Носенко, Е. Л. (2011а). Обґрунтування вибору параметрів імпліцитної діагностики емоційної стійкості при її комп'ютерному моделюванні. *Вісник Дніпропетровського національного університету. Серія "Педагогіка і психологія"* 17, 3–16. Retrieved from <http://psyresearch.dp.ua/index.php/psy/article/view/2/4>
- Аршава, І. Ф., & Носенко, Е. Л. (2011b). Суб'єктивне благополуччя і його індивідуально-психологічні та особистісні кореляти. *Вісник Дніпропетровського університету. Серія "Педагогіка і психологія"*, 20(18), 3–9. Retrieved from http://nbuv.gov.ua/j-pdf/vdupp_2012_20_18_3.pdf
- Белікова, Ю. В. (2012). Моделі досягнення успіху у професійній діяльності в залежності від рівня емоційного інтелекту (EQ). *Сучасні суспільні проблеми у вимірі соціології управління: збірник наукових праць ДонДУУ*.



Серія “Соціологія,” XIII(257), 259–266.

- Бородавко, О. О. (2011). Дослідження психометричних характеристик методики діагностики міжособистісних відносин лікаря та пацієнта в процесі лікування. *Український вісник психоневрології*, 19(4), 115–119. Retrieved from http://nbuv.gov.ua/j-pdf/Uvpr_2011_19_4_31.pdf
- Боснюк, В. Ф. (2010). Відмінності у використанні захисно-копінгової поведінки в залежності від емоційного ресурсу особистості. *Проблеми екстремальної та кризової психології*. Збірник наукових праць, 11, 23–30. Retrieved from http://edu-mns.org.ua/nmc/123/EKP_11.pdf
- Бунас, А. А. (2012). Схильність до ризику в системі особистісних детермінант. Збірник наукових праць Інституту психології імені Г. С. Костюка Національної АПН України. *Проблеми загальної та педагогічної психології*, 24(6), 51–59. Retrieved from http://nbuv.gov.ua/j-pdf/pzpp_2012_24_6_8.pdf
- Бунас, А. А. (2014). Прогностична компетентність як соціально значуща властивість особистості при прояві схильності до ризикованої поведінки. *Вісник Одеського національного університету*. Психологія, 19(2), 52–61. Retrieved from http://nbuv.gov.ua/j-pdf/Vonu_psi_2014_19_2_8.pdf
- Виноградов, О. Г. (2013). Номенклатура сигнатур емоційного реагування на ситуації публічного виступу. *Актуальні проблеми соціології, психології, педагогіки: збірник наукових праць*, 18, 119–126. Retrieved from <http://www.socd.univ.kiev.ua/sites/default/files/library/elopen/aktprob.18.119.pdf>
- Грись, А. М. (2014). Психологічні бар'єри студентів-психологів у процесі їх підготовки до роботи з соціально дезадаптованими неповнолітніми. *Проблеми загальної та педагогічної психології: [збірник наукових праць Інституту психології імені Г. С. Костюка НАПН України / за ред. С.Д.Максименка]*, XVI(2), 289–298. Retrieved from <http://lib.iitta.gov.ua/1692/1/>
- Гупаловська, В. А. (2010). Емпіричне дослідження гендерних особливостей задоволення життям жінок. *Соціогуманітарні проблеми людини*, 5, 185–200. Retrieved from <http://dspace.nbuv.gov.ua/bitstream/handle/123456789/27360/18-Gupalovska.pdf?sequence=1>
- Дерев'яно, С. П. (2009). Емоційний інтелект як чинник соціально-психологічної адаптації особистості до студентського середовища (Автореф. дис. канд. психол. наук: 19.00.05). Ін-т соц. та політ. психології АПН України, Київ. Retrieved from <http://mydisser.com/en/catalog/view/252/257/18207.html>
- Доцевич, Т. І. (2014). Темпоральні особливості метакогнітивної компетентності викладачів вищої школи. *Naukovij oglas*, 5(6), 176–184. Retrieved from <http://naukajournal.org/index.php/naukajournal/article/view/228/363>
- Єсип, М. З. (2014). Прагматична оцінка релігійних способів долаючої поведінки. *Прагматична оцінка релігійних способів долаючої поведінки*, 24, 314–327. Retrieved from http://nbuv.gov.ua/j-pdf/Pspl_2014_24_28.pdf
- Завгородня, Н. І., & Арсенєв, Н. І. (2014). Результати систематизації чинників формування розладів адаптації у жінок, що народили недоношену дитину. *Ліки України плюс*, 3, 57–60. Retrieved from http://repo.knmu.edu.ua/bitstream/123456789/7331/1/lu_PLUS_3_2014_.pdf
- Захарко, О. О. (2011). Соціально-психологічні особливості детермінації дозвіллевих практик студентської молоді. *Проблеми сучасної психології: зб. наук. пр. Кам'янець-Поділ. нац. ун-ту ім. Івана Огієнка, Ін-ту психології ім. Г. С. Костюка НАПН України*, 11, 278–291. Retrieved from <http://www.twirpx.com/file/742646/>
- Зотова, Л. Н. (2012). Профиль невротических наклонностей как личностный фактор, определяющий внутреннюю картину здоровья у студентов. *Вісник Харківського національного педагогічного університету імені Г.С. Сковороди*. Психологія, 44(1), 79–92. Retrieved from <http://journals.hnpu.edu.ua/ojs/psycho/article/view/1167/1173>
- Кліманська, М. Б. (2013). Психологічні особливості використання гумору медичними працівниками. *Проблеми сучасної психології*, 21, 225–236.
- Лоза, О. О. (2013). Диференціація типів прояву перфекціонізму. *Сборник научных трудов SWorld*, 3(23). Retrieved from <http://www.sworld.com.ua/konfer32/1027.pdf>
- Ляковська, І. Л. (2014). Психологічна корекція негативних проявів кризи середнього віку (автореферат канд. психол. наук, спец.: 19.00.07 - педагогічна та вікова психологія). Національний педагогічний університет імені М.П. Драгоманова, Київ. Retrieved from <http://enpuir.npu.edu.ua/bitstream/123456789/5342/1/Liaskovskaya.pdf>
- Меднікова, Г. І. (2013). Характеристика самоактуалізації студентів з різними варіантами структури ставлення до себе. *Науковий вісник Миколаївського державного університету імені В. О. Сухомлинського*. Серія: Психологічні науки, 2(10), 200–205. Retrieved from http://nbuv.gov.ua/j-pdf/Nvmdups_2013_2_10_41.pdf
- Меднікова, Г. І. (2014). Структура ставлення до себе студентів як компоненту системи детермінант становлення особистісної зрілості. *Naukovij oglas*, 6(5). Retrieved from <http://naukajournal.org/index.php/naukajournal/article/>



view/242/392

- Мельничук, М. М. (2014). Дослідження детермінант толерантності як системної характеристики особистості. Вісник ХДУ Серія Психологічні науки, 2(1), 94–98. Retrieved from http://pj.kherson.ua/file/psychology_02/ukr/part_1/19.pdf
- Нахабич, М. А. (2013). “Ціннісна” мапа України в системі аксіологічних досліджень. Вісник Одеського національного університету. Соціологія і політичні науки, 18(2(1)), 118–130.
- Носенко, Е. Л., & Новицька, І. В. (2013). Особистісні фактори успішності розпізнавання усних повідомлень з прихованими мисленневими образами. Вісник Дніпропетровського університету. Серія: Педагогіка і психологія, 21(19), 93–102.
- Носенко, Е. Л., & Труляев, Р. О. (2013). Психологічний механізм впливу позитивних цінностей педагога на суб’єктивне благополуччя учнів. Психологічний механізм впливу позитивних цінностей педагога на суб’єктивне благополуччя учнів, 21(19), 102–109. Retrieved from http://irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?C21COM=2&I21DBN=UJRN&P21DBN=UJRN&IMAGE_FILE_DOWNLOAD=1&Image_file_name=PDF/vdupp_2013_21_19_16.pdf
- Петренко, Т. П. (2012a). Діагностичний потенціал невербальних засобів спілкування у дослідженні “позитивних цінностей” та темних рис особистості. Вісник Дніпропетровського університету. Серія: Педагогіка і психологія, 20(18), 129–134. Retrieved from http://nbuv.gov.ua/j-pdf/vdupp_2012_20_18_22.pdf
- Петренко, Т. П. (2012b). Невербальні засоби спілкування як аспект імпліцитної діагностики особистісних характеристик. Вісник Національної академії Державної прикордонної служби України, 5. Retrieved from http://nbuv.gov.ua/j-pdf/Vnadps_2012_5_46.pdf
- Рафа, М. Б. (2012). Оптимальні особливості суб’єктивної реконструкції особистого досвіду як детермінанти життєвої успішності в юнацькому віці. Вісник Харківського національного педагогічного університету імені Г. С. Сковороди. Психологія, 42(2), 134–146. Retrieved from [http://nbuv.gov.ua/j-pdf/VKhnpu_psykhol_2012_42\(2\)_18.pdf](http://nbuv.gov.ua/j-pdf/VKhnpu_psykhol_2012_42(2)_18.pdf)
- Сотнікова-Мелешкіна, Ж. В. (2012). “Адаптаційний профіль” старшокласників із різним характером професійної готовності. Вісник Харківського національного університету імені В. Н. Каразіна. Серія: Медицина, 1024(24), 42–51. Retrieved from http://nbuv.gov.ua/j-pdf/VKhM_2012_1024_24_9.pdf
- Труляев, Р. О. (2012). Позитивні цінності педагога як фактор його професійної ефективності (досвід використання наукових здобутків позитивної психології). Вісник Дніпропетровського університету. Сер.: Педагогіка і психологія, 20(18), 140–150. Retrieved from http://nbuv.gov.ua/j-pdf/vdupp_2012_20_18_24.pdf
- Турецька, Х. І. (2014). Особливості емоційного інтелекту працівників сфери ІТ, схильних до емоційного вигорання. Наук. пер. вид. Український психолого-педагогічний науковий збірник. Наук. журнал, 2, 141–146.
- Ушанова, А. М. (2014). Особливості життєстійкості осіб з різною антиципацією задоволеності особистісним вибором (pp. 105–112). Presented at the Сучасні підходи та технології соціально-психологічної та корекційної роботи з різними віковими групами населення: збірник матеріалів міжнародної науково-практичної конференції, 20-22 червня 2014 року, Одеса: Одеський нац. ун-т. Retrieved from http://www.irbis-nbuv.gov.ua/cgi-bin/irbis_nbuv/cgiirbis_64.exe?C21COM=2&I21DBN=UJRN&P21DBN=UJRN&IMAGE_FILE_DOWNLOAD=1&Image_file_name=PDF/Vonu_psi_2014_19_2_39.pdf
- Четверик-Бурчак, А. Г. (2013). Емоційний інтелект як чинник успішності вирішення головних життєвих завдань. Психолінгвістика, 14, 163–184. Retrieved from http://nbuv.gov.ua/j-pdf/psling_2013_14_15.pdf
- Шамлян К. М. (2014). Особливості вольової організації особистості студентів. Проблеми сучасної психології, 23, 707–719. Retrieved from http://nbuv.gov.ua/j-pdf/Pspl_2014_23_64.pdf

Citation

Zakharov, K. (2016). Application of k-means clustering in psychological studies. *The Quantitative Methods for Psychology*, 12(2), 87–100. doi:10.20982/tqmp.12.2.p087

Copyright © 2016, Zakharov . This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 22/02/2016 ~ Accepted: 24/03/2016