Research of the Effectiveness of Methods for Missing Data Imputation for Assessing the Impact of Intellectual and Personal Components on the Academic Performance of Students

Anastasiia Timofeeva [a], Tatiana Avdeenko [a], Olga Razumnikova [a] and Yulia Andrusenko [b]

[a] *Novosibirsk State Technical University, 20, Karla Marksa ave., Novosibirsk, 630073, Russia*
[b] *North-Caucasus Federal University, 2, Kulakov prosp., Stavropol, 355029, Russia*

Abstract

The article explores the problem of missing data to determine the contribution of different components of intelligence and personality factors to student performance. The specificity of the available data is that missed data cannot be considered random. This leads to the instability of the results of filling in the gaps using multiple imputation. Therefore, the task of comparing their performance and choosing the best method for a particular set of data arises. It is suggested to use average dispersion of regression parameter estimates and average statistics reflecting the significance of regression parameters as indicators of method effectiveness. Two methods of multiple imputation for source data the missForest and Amelia were investigated, as well as after selecting the principal components and applying a special procedure of forming limited sets of principal components. The missForest method using the principal components shows the best results and allows identifying informative personal and intellectual predictors of students' academic performance: the level of general, emotional and social intelligence and indicators of introversion and social conformality.

Keywords [1]

Missing data imputation, principal component analysis, psychometric testing, intelligence, personality, academic performance

1. Introduction

In complex psychological studies of large samples there are problems with data analysis due to missing values of certain indicators in the collected arrays. Such missingness may be caused by both random factors and peculiarities of the study organization. For example, some characteristics are not examined on the entire sample of students, but only on a single group. As a result, a number of variables have missing values for the same objects. This complicates the analysis and filling of the gaps in the data.

Many algorithms for filling in missed data are built on the assumption that the missing values occur at random. Therefore, presence of groups of variables with missing values for the same objects may distort data processing results and reduce efficiency of methods of filling in the missed data. In this regard, it is of interest to compare the effectiveness of different methods of missing data imputation.

As a rule, efficiency of methods of filling of missing values is investigated on model examples for which true values of variables which have been lost in original data are known. For such studies, [1] a

YRID-2020: International Workshop on Data Mining and Knowledge Engineering, October 15-16, 2020, Stavropol, Russia
EMAIL: a.timofeeva@corp.nstu.ru (Anastasiia Timofeeva); tavdeenko@mail.ru (Tatiana Avdeenko); razumnikova@corp.nstu.ru (Olga Razumnikova); yulihka85@mail.ru (Yulia Andrusenko)
ORCID: 0000-0001-9900-026X (Anastasiia Timofeeva); 0000-0002-8614-5934 (Tatiana Avdeenko); 0000-0002-7831-9404 (Olga Razumnikova); 0000-0002-3392-7270 (Yulia Andrusenko)

number of measures are proposed in the work to assess the quality of filling in the missing data. However, the model examples only partially reflect the observed reality. And the advantages of one method revealed with their help compared to others can be valid only within the framework of these examples and do not apply to the existing specific data set. In the practice of researching student cognitive status, there is a problem with choosing an appropriate method for filling gaps, which would be more effective for a given

data set. Hence, there is a need to establish some measure of effectiveness that can only be calculated from observable data.

This article deals with the practical task of determining the contribution of a number of variables: different components of intelligence and personality traits in the performance of 473 students of Novosibirsk State Technical University. The data set included psychometric indicators of creative abilities, general, emotional and practical intellect, generalized personality factors (extroversion, neurotic and psychoticism) and gender-role stereotypes, but part of the sample lacked the latter indicators.

In this article it is proposed to evaluate the effectiveness of methods to fill gaps based on the stability of regression estimates. The less regression analysis results vary with different options for filling gaps, the more reliable the results the researcher gets because they remain replicable in repeat studies.

2. Overview of missing data imputation methods

The simplest methods of filling in the gaps assume that the gaps are filled separately for each variable, for example, based on the mean. They do not take into account interrelationships between variables and give a significant bias if there are many missing values.

Multiple imputations are more popular [2], but have some disadvantages. They are based on the assumption that data are missing at random (MAR). The latter means that the underlying mechanism of missing data, given the observed data, does not depend on unobservable data.

A sensitivity analysis has been proposed to assess the stability of the results of the multiple imputation with respect to model assumptions (MAR) [3]. It also allows comparing the effectiveness of different multiple imputation methods and quantifying the degree of systematic bias caused by the absence of randomness in the missed data.

Article [4] describes the pitfalls that arise when applying multiple imputation methods:

- exclusion of a response variable from the imputation procedure,
- processing non-normally distributed variables,
- plausibility and violation of the assumption of missed data randomness,
- computational problems.

In this article, multiple imputation methods that take into account the relationships between all variables in the dataset have been chosen as methods for filling the missed data, for which effectivity comparison has been performed. More specifically, the algorithms missForest [5] and Amelia [6], implemented in R.

The missForest algorithm uses a random forest trained on observable data matrix values to predict missed values. This is a non-parametric method of filling in the gaps, applicable to different types of variables. The non-parametric method makes no explicit assumptions about the functional form. Instead, it tries to estimate it so that the result appears as close to the data points as possible, but does not seem impractical. It builds a random forest model for each variable. It then uses the model to predict the missing values of the variable with the observed values.

The approach described above gives an estimate of the OOB (out of bag) imputation error and also provides a high level of control over the imputation process. Moreover, it has options to return the OOB separately (for each variable) instead of aggregation across the entire data matrix. This allows for a closer look at how accurately the model fills in the gaps for each variable.

The Amelia algorithm is based on the bootstrap EM algorithm for incomplete data. It allows getting a given number of imputed datasets where the observed sample values are the same, and the unobserved values are drawn from their posterior distribution. The correct results of this method are obtained with the following assumptions:

- all variables in the dataset have a multivariate normal distribution. Averages and covariances are used to summarize the data.
- The missing data are random.

The algorithm works best when the data have a multivariate normal distribution. Otherwise, you need to perform a transformation to get the data close to normal.

If the variables are strongly correlated and have gaps for the same objects, the Amelia method returns a warning message because the results of filling the gaps may be incorrect.

In order to exclude correlation within a group of interrelated cognitive characteristics, it is suggested to use the Principal Components (PC) analysis. Principal component extraction allows you to switch to uncorrected input characteristics. It also allows to increase stability of regression estimates, because it weakens the multicollinearity problem.

3. Description of original data

The array of analyses included normalized measures of general intellect (which was determined using the Amthower Structure of Intelligence test) (IQa), emotional intellect (Barchard test) (IQe), social intelligence (Guilford-Sullivan test) (IQs), creativity (IQc), and practical intelligence (IQp). Personal traits: extroversion (E), neuroticism (N), psychoticism (P) and social conformance (L) were determined according to the EPQ questionnaire, femininity (F), masculinity (M), androgynous (A) - according to the Bem questionnaire (for more information on methods of determining intellectual and personal traits, see Bem's questionnaire). [7, 8]).

The total sample size was 473 observations. However, the number of missing values is quite large. The number of missed data for different variables is shown in Fig. 1.

Although there are more than 100 observations for each variable, deletion of rows containing at least one skip to estimate the regression results in only 13 valid rows remaining. This explains the need to fill in the gaps to evaluate the contribution of each variable to the progress.

Replacing any missing value with the mean of each variable does not yield meaningful results. The regression model built on such data is insignificant at 10% significance level, which does not allow estimating the contribution of variables to the progress. Therefore, to obtain qualitative results, it is necessary to apply multiple imputation methods.

Groups of variables containing missing values for identical objects can be selected:

- personal characteristics: conformality (L), neuroticism (N), extroversion (E), psychoticism (P);
- gender stereotypes: masculine (M), feminine (F), anthropogenic (A) and the ratio of masculinity to femininity (K_MF).

For variables of emotional (IQe) and social (IQs) intellectual abilities, about half of the missing values are for the same objects. The same is true for IQa and IQs. For IQp the number of gaps was the highest (almost 200 students), so it was not included in the IQ system and was considered separately.

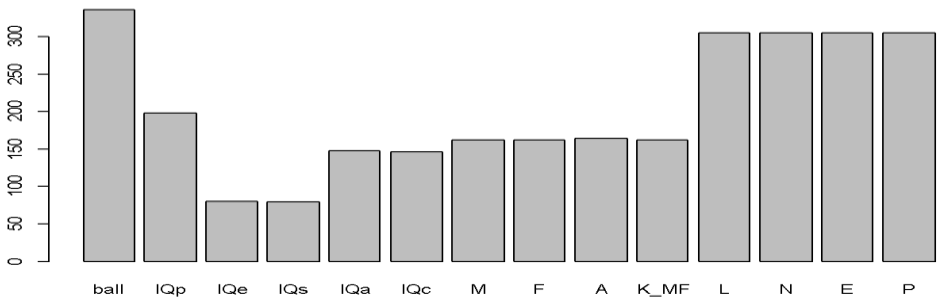

Figure 1: Number of missed observations

It was found that most indicators have an average degree of variation (relative standard deviation in the range from 10% to 30%). The exceptions are such variables as K_MF with a coefficient of variation almost equal to 5, and personal characteristics whose coefficient of variation is in the range

from 30% to 60%. However, a normality test using the Shapiro-Wilk criterion shows that for variables K_MF, L, as well as M and F, the empirical distribution does not deviate from normal at 5% significance level. The empirical distribution of the other indicators is very different from normal. For this reason, we should expect that the Amelia method will give worse results than the non-parametric missForest method.

4. Results of applying the principal component analysis

In order to pass to the independent input variables, for each selected group of cognitive indicators (IQ, gender-role stereotypes, personality characteristics), a principal component analysis was carried out. Rows with at least one missing value were excluded from the analysis. Rotation was used by the varimax method. The resulting loading matrices for the principal components, as well as the proportion of variance explained, are presented in Tables 1-3.

In most cases, the correlation between features is weak; therefore, the share of the explained variance is evenly distributed. Nevertheless, for the group of indicators of gender-role stereotypes (Table 2), it is possible to single out an insignificant component MF4, which explains only 0.5% of the variance of features. Further, it is excluded from the analysis. For the remaining groups of indicators, 4 principal components were used. For them, it is easy to establish a correspondence between the initial characteristics and the principal components based on the maximum absolute values of loadings. They are bold in Tables 1-3.

The values of the principal components were extracted from the constructed loading matrices. For those objects for which there was at least one missing value of the original features, the values of the principal components were also considered missing.

The use of two or more components for one group of features led to the fact that the problem of the presence of variables with gaps for the same objects remained unsolved. To weaken it, it is proposed to form a matrix of input factors based on a limited set of variables included in different groups of cognitive characteristics. The uncorrelatedness of the principal components allows each of them to be included in the regression model separately; this should not lead to a strong change in the parameter estimates. Next, the regression model is estimated for a limited set of variables in which the gaps are filled.

Table 1

Principal component analysis results for a group of IQ indicators

	IQ1	IQ2	IQ3	IQ4
IQe			**0.994**	
IQs	0.265			**-0.964**
IQa	**0.948**	0.123		-0.279
IQc	0.108	**0.992**		
Proportion of explained variance	0.246	0.251	0.251	0.252

Table 2

Results of the principal component analysis for a group of indicators of gender-role stereotypes

	MF1	MF2	MF3	MF4
M		**0.983**	0.182	
F	**0.947**		0.32	
A	0.269	0.125	**0.955**	
K_MF	**-0.685**	**0.705**	-0.112	0.143
Proportion of explained variance	0.36	0.37	0.265	0.005

Table 3
Results of the principal component analysis for a group of indicators of personal characteristics

	LH1	LH2	LH3	LH4
L	**0.974**		-0.133	-0.18
N	-0.177			**0.98**
E		**-0.995**		
P	-0.127		**0.991**	
Proportion of explained variance	0.25	0.25	0.25	0.25

The algorithm for generating sets of variables and estimating regression is described below.

Let there be a constant group of factors. In our case, this is the IQp variable for which the principal component analysis was not applied. In addition, there are I groups of characteristics, to each of which the principal component method was applied. We have $I = 3$. Let in each group $K_1,\ldots, K_I$ of the principal components are selected, which are included in the regression model. We have

$$K_1 = 4, K_2 = 3, K_3 = 4. \quad (1)$$

Step 1. Form a matrix G of all combinations of numbers from sets,

$$\{1,...,K_1\},\{1,...,K_2\},...,\{1,...,K_I\} \quad (2)$$

one from each set. Each row of the matrix corresponds to a specific combination. The number of columns in the matrix equal I.

Step 2. Put t: = 1.

Step 3. We include in the set of input factors variables of the constant group of factors, the g_{ts}-th variable from the s-th group of characteristics, to which the principal component method was applied, for all s = 1, ..., I, where g_{ts} is an element of the matrix G.

Step 4. Estimate a linear regression model of the form:

$$y = \beta_0 + \sum_{i=1}^{r} \beta_i x_i + \varepsilon, \quad (3)$$

where y is academic performance, $x_1,...,x_r$ - is the set of variables selected in step 3.

Step 5. While

$$t < \prod_{s=1}^{I} K_s , \quad (4)$$

t:=t+1, go to step 3, otherwise the end of the algorithm.

The estimation results are generalized to the initial set of k features (in our case, k = 12) by averaging the estimates for each variable obtained for all possible sets of input factors. For our case, there will be 48 such sets. As a result, we get

$$\text{mean}_j\left(\hat{\beta}_i\right), \quad (5)$$

the average value of the assessment of the contribution of the i-th feature for a given j-th variant of filling in the gaps. The variance is calculated at the same time

$$\text{var}_j\left(\hat{\beta}_i\right), \quad (6)$$

estimates of the contribution of the i-th feature, obtained for a given j-th variant of filling in the gaps.

5. Performance indicators

Since the considered multiple imputation algorithms produce random results, it is of interest, first of all, the stability of the regression estimates constructed from data with filled gaps. Let the algorithms be applied to the same data set m times, then we get m sets of regression estimates. They are used to calculate the variance of the regression estimates. The Mean Variance for all estimates will be an indicator characterizing the stability of the estimates.

$$MV = \frac{1}{k}\sum_{i=1}^{k} \mathrm{var}\left(\hat{\beta}_i\right), \tag{7}$$

where $\mathrm{var}\left(\hat{\beta}_i\right)$ is the variance of the estimate $\hat{\beta}_i$ over all random imputations, *k* is the number of elements of vector

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k\right) \tag{8}$$

except the intersection. When using the algorithm for generating sets of independent variables

$$\mathrm{var}\left(\hat{\beta}_i\right) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{var}_j\left(\hat{\beta}_i\right). \tag{9}$$

However, the magnitude of the variance depends on the scale; therefore, comparison of the variation in estimates constructed from different datasets may not always be meaningful. Therefore, it is additionally proposed to use the indicator proposed in the work [9]:

$$S = \frac{1}{k}\sum_{i=1}^{k} t\left(\hat{\beta}_i\right), \tag{10}$$

Where

$$t(\hat{\beta}_i) = \frac{\left|\mathrm{mean}\left(\hat{\beta}_i\right)\right|}{\mathrm{sd}\left(\hat{\beta}_i\right)} \tag{11}$$

is the analogue of *t*-statistics, $\mathrm{mean}\left(\hat{\beta}_i\right)$ is the estimate $\hat{\beta}_i$ averaged over all random imputations,

$$\mathrm{sd}\left(\hat{\beta}_i\right) = \sqrt{\mathrm{var}\left(\hat{\beta}_i\right)} \tag{12}$$

is the standard deviation of the estimate $\hat{\beta}_i$ over all random imputations. When using the algorithm for generating sets of independent variables

$$\mathrm{mean}\left(\hat{\beta}_i\right) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{mean}_j\left(\hat{\beta}_i\right). \tag{13}$$

This statistic characterizes the quality of estimates. It grows with increasing absolute values of estimates and decreasing their deviation. Thus, the S-statistic is an averaged t-statistic, and can be interpreted in a similar way.

Table 4 presents indicators of the quality of the regression estimation, calculated for 500 random imputations. It is worth noting that in terms of computational performance, missForest is significantly slower than Amelia. However, missForest significantly outperforms Amelia in terms of stability.

Table 4

Performance indicators

Index	Method	Raw scaled data	Principal components	Subsets of PCs
MV	Amelia	0.3983	0.0878	0.0515
	missForest	0.0038	0.0014	0.0206
S	Amelia	0.5230	0.6567	**1.1542**
	missForest	3.5869	**4.5104**	2.9137

The use of the method of principal components can significantly reduce the average variance of estimates, this decrease is especially significant for the Amelia method (4.5 times). However, the S-statistic does not increase significantly as a result. The use of limited sets of principal components from different groups of characteristics improved S-statistics for regression estimates constructed after filling in the gaps with the Amelia method.

The best result was achieved using the missForest method using principal components. At the same time, the application of the proposed procedure for the formation of limited sets of principal components made it possible to improve the results of the Amelia method.

6. Results of assessing the contribution of cognitive and personal characteristics to student performance

Let us compare the results of applying the principal component analysis and of using raw data in terms of the obtained regression estimates and their significance (Tables 5-6).

In general, for parameters with high values $t(\hat{\beta}_i)$ (in particular, $t(\hat{\beta}_i) > 2$) in Table 6, there are no significant discrepancies in terms of the direction of influence (negative or positive) of psychological characteristics on student performance.

Table 5

Results of assessing the contribution of cognitive and personal characteristics to student performance

Input variables	MissForest		Amelia	
	$\text{mean}\left(\hat{\beta}_i\right)$	$t\left(\hat{\beta}_i\right)$	$\text{mean}\left(\hat{\beta}_i\right)$	$t\left(\hat{\beta}_i\right)$
IQp	-0.024	1.497	-0.022	0.100
IQe	0.073	**4.759**	-0.019	0.139
IQs	0.026	1.826	0.131	0.583
IQa	-0.044	**2.586**	-0.062	0.212
IQc	0.031	1.708	0.093	0.451
M	-0.257	**2.773**	-0.060	0.054
F	-0.004	0.042	0.236	0.228
A	-0.341	**17.429**	-0.611	**2.553**
K_MF	0.049	0.434	-0.106	0.073
L	-0.269	**3.544**	-0.123	0.318
N	0.134	**2.536**	0.294	0.934
E	-0.295	**3.279**	-0.234	0.925
P	-0.145	**4.218**	-0.090	0.229

Table 6

Results of Regression Estimation Using Principal Components

PCs	Input variables	MissForest Principal components		Amelia Subsets of PCs	
		$\text{mean}\left(\hat{\beta}_i\right)$	$t\left(\hat{\beta}_i\right)$	$\text{mean}\left(\hat{\beta}_i\right)$	$t\left(\hat{\beta}_i\right)$
IQp	IQp	-0.025	1.232	0.008	0.142
IQ1	IQa	-0.166	**5.664**	-0.167	1.264
IQ2	IQc	0.032	1.368	-0.038	0.398
IQ3	IQe	0.057	**3.114**	0.081	0.659
IQ4	-IQs	-0.047	**2.307**	-0.130	1.060
MF1	F, -K_MF	-0.043	1.986	0.042	0.278
MF2	M, K_MF	-0.268	**7.151**	-0.281	**2.597**
MF3	A	-0.280	**15.301**	-0.378	**3.525**
LH1	L	-0.407	**9.274**	-0.174	0.878
LH2	-E	0.233	**3.065**	0.278	1.710
LH3	P	-0.076	**2.194**	-0.061	0.309
LH4	N	0.084	1.471	0.235	1.029

Noteworthy is the more significant influence on the performance of indicators of personal characteristics (LH and MF, Table 6) than intellectual abilities (IQ). Apparently, the use of multiple imputation for filling the gaps leads to the fact that the effectiveness of the implementation of a behavioral response (passing exams), first of all, according to the psychobiological cognitive-adaptive model of personality [10], is determined by the most generalized personality traits: gender-role stereotypes and personality superfactors (L, E , P) and further - cognitive systems of information and memory selection, which underlie the organization of personality traits, intelligence and adaptive behavior.

The opposite contribution to the performance of general and social intelligence (respectively, IQa and IQs in Table 6) obtained according to the MissForest method, with the noted positive role of its emotional component (IQe), may reflect the dominant influence of personality characteristics as "personal intelligence" [11] on effective delivery exams with high marks. This interpretation is supported by the significant contribution of other personality indicators: E and L, with a higher academic performance corresponding to a tendency to introversion and low social conformity. It is noteworthy that the conclusion about the predominant influence of personality characteristics in the obtained regression model of academic performance follows from the use of the principal component method for both the MissForest and Amelia algorithms.

7. Conclusion

In this paper, some performance indicators are proposed for missing value imputation. They allow you to choose an appropriate method for dealing with missing data that provides more stable results of estimating the impact of intellectual and personal properties on the academic performance.

The close relationship of intellectual and personal properties in the analysis of the effectiveness of educational activities is shown in other studies (for example, [12, 13, 14]). It should be noted that there is a complex indirect relationship between general, social and emotional intelligence associated with the type of educational practice and examination success [15, 16, 17], and the relationship between psychological properties and the results of educational activity is not always linear and unidirectional. So earlier, using cluster analysis with the use of the discrete optimization method, we have shown the multidirectional contribution of creativity to the success of learning, depending on the level of general intelligence [18].

Consequently, despite some value of the developed procedures for the formation of sets of input factors based on the principal component analysis and the proposed approach to determining the effectiveness of methods for filling in missing psychometric data in the particular problem we have considered, further studies of the effectiveness of solving this problem are required depending on the dimension and nature of the collected arrays.

8. Acknowledgements

The research is supported by the Ministry of Science and Higher Education of Russian Federation (project No. FSUN-2020-0009).

9. References

[1] Bergold S., Steinmayr R. Personality and intelligence interact in the prediction of academic achievement // J. Intell. 2018. V. 6. N 27. e6020027.

[2] Crameri A. et al. Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy //Frontiers in psychology. – 2015. – T. 6. – Art. 1042.

[3] Fonteyne L., Duyck W., De Fruyt F. Program specific prediction of academic achievement on the basis of cognitive and non-cognitive factors // Learn. Individ. Differ. 2017. V. 56. P. 34–48.

[4] Gil-Olarte Márquez P, Palomera Martín R, Brackett MA. Relating emotional intelligence to social competence and academic achievement in high school students. Psicothema. 2006;18 Suppl:118-123.

[5] Héraud-Bousquet V. et al. Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data //BMC medical research methodology. – 2012. – Т. 12. – №. 1. – С. 73.
[6] Honaker, J., King, G., Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1–47. URL http://www.jstatsoft.org/v45/i07/
[7] Kiss M., Kotsis Б., Kun A.I. The relationship between intelligence, emotional intelligence, personality styles and academic success // Business Education & Accreditation. 2014. V. 6. N 2. P. 23–34.
[8] Matthews, G. (2015). Personality, Cognitive Models of. International Encyclopedia of the Social & Behavioral Sciences, 870–875. doi:10.1016/b978-0-08-097086-8.25075-7
[9] Mayer, J. D. (2015). The personality systems framework: Current theory and development. Journal of Research in Personality, 56, 4–14. doi:10.1016/j.jrp.2015.01.001
[10] Razumnikova OM Differential psychology. Novosibirsk, Publishing house of NSTU, 2019.160 p.
[11] Razumnikova OM What is intelligence? Novosibirsk, Publishing house of NSTU, 2018.78 p.
[12] Razumnikova OM, Mezentsev Yu.A. Features of the ratio of creativity and emotional intelligence with the academic performance of university students depending on the level of general intelligence // Questions of psychology. - 2020. - No. 2. - P. 1-10.
[13] Sánchez-Álvarez N, Berrios Martos MP, Extremera N. A Meta-Analysis of the Relationship Between Emotional Intelligence and Academic Performance in Secondary Education: A Multi-Stream Comparison. Front Psychol. 2020;11:1517. Published 2020 Jul 21. doi:10.3389/fpsyg.2020.01517
[14] Stavseth M. R., Clausen T., Røislien J. How handling missing data may impact conclusions: a comparison of six different imputation methods for categorical questionnaire data //SAGE Open Medicine. – 2019. – Т. 7. – С. 2050312118822912.
[15] Stekhoven, D.J. and Buehlmann, P. (2012), 'MissForest - nonparametric missing value imputation for mixed-type data', Bioinformatics, 28(1) 2012, 112-118, doi: 10.1093/bioinformatics/btr597
[16] Sterne J. A. C. et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls //Bmj. – 2009. – Т. 338.
[17] Timofeeva A. Y. Robust principal component regression on compositional covariates with application to educational monitoring / A. Y. Timofeeva // Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis, AMSA'2017 : proc. of the intern. workshop, Krasnoyarsk, 18–22 Sept. 2017. – Novosibirsk : NSTU publ., 2017. – P. 241-248.
[18] Truninger M, Fernández-I-Marín X, Batista-Foguet JM, Boyatzis RE, Serlavós R. The Power of EI Competencies Over Intelligence and Individual Performance: A Task-Dependent Model. Front Psychol. 2018;9:1532