Journal of Physics: Conference Series **1694** (2020) 012012 doi:10.1088/1742-6596/1694/1/012012

Development of cluster models for municipal educational institutions of Tomsk region

I K Kvasnikova[1], A V Lepustin[2], Yu Ya Katsman[2] and E V Lepustina[3]

[1] Education Quality Estimation Center, Tomsk Regional Institute of Advanced Training and Retraining of Educators, 10, Pirogova st., 634034, Tomsk, Russia

[2] Division for Information Technology, Tomsk Polytechnic University, 30, Lenina ave., Tomsk, 634050, Russia

[3] Department of Radio Engineering Systems, Tomsk State University of Control Systems and Radioelectronics, 40, Lenina ave., 634050, Tomsk, Russia

foxkik@tpu.ru

Abstract. This paper describes the cluster models construction for schools in the Tomsk region. K-means method is used to distribute objects between clusters using the multidimensional vector of variables (socio-economic status of students, qualification and age of teaching staff, students' participation in academic competitions, students in difficult living conditions, etc.). The cluster models were constructed for urban schools (regional center), country and ungraded schools. The computer simulation using STATISTICA system was performed due to large amount of source data and complexity of the developed models. The analysis showed the relationship between the values of cluster variables and the results of Unified State Exams (USE).

1. Introduction

Assessment of school (municipal educational institutions) performance is the subject of close and regular attention aimed both for the improvement of the quality of educational achievements and for the planning (prediction) of these achievements using the available data. In a number of papers [1, 2] was discussed that assessment of school performance using the results of USE and the State Final Exams for 9th grade (SFE-9) imprecisely depends on territorial location of the school (town, country), socio-economic status of students (parents), qualification and age of teaching staff, etc. Noticed that the results of USE and SFE-9 of different schools depend on the number of participants and winners of academic competitions (on different levels), school functioning in difficult social conditions. Classification of the variety of schools, even if they are located on the same territory is the non-trivial problem due to incorrect rating for different types of schools (gymnasium, lyceum, public school). Indeed, the number of lyceums and gymnasiums with most quality education have only 10th and 11th grades for pupils with best results [3, 4]. This type of students' selection makes the school comparison meaningless and incorrect even on same territory.

2. Results and discussion

2.1. Problem statement and preliminary research

In this paper we used the characteristics (factors) of financial, social and other indicators of schools and students, which were got during forming the social passport of school (as the element of the regional system of the educational quality assessment) [5]. The results of this monitoring were used as the dataset including 188 parameters for each of 229 schools of Tomsk region.

At the first stage of this research, we estimated the students' results of USE in Russian language and Mathematics using statistical criteria (Student and Fisher). The results were estimated depending on the type of each school (regional center, small town, country, small (ungraded) school) in order to get the assessment of the regional school performance. Using the methods of linear correlation analysis allowed us determining the most significant factors, which can affect the students' educational quality. Based on these results, the multifactor linear regression models were developed to estimate the educational achievements in Russian Language and Mathematics for the students from different types of schools [6]. During the investigation, we estimated the quality of these models and determined optimal model dimensions taking into account the uncertain impact of various factors on performance of different types of schools. Therefore, for schools of regional center and results of USE (Russian Language), optimal model dimension is 6 and corrected coefficient of determination does not exceed 60%. However, a statistical model is considered good when it has coefficient of determination more than 75%. [7]. As for results of USE (Mathematics) and for country schools, the model characteristics are even lower, although obtained results and conclusions are well-corresponded to independent research [8].

At the second stage, we developed the factor models of educational quality assessment using the most significant characteristics, affecting the students' educational achievements [9]. Models were constructed with the principal component method. This and using the statistical criteria (Kaizer and Cattell) allowed reducing the number of factors to three, four [10, 11]. For visual explanation of the factor load values, we implemented the rotation of factors (*Varimaxrow* method). The results analyses showed that the simpler three-factor model describes about 67% of the total variance. For the four-factor model, the similar results are 75%, which we considered as satisfactory, despite the complexity of the model.

Within constructed stochastic models of the educational quality assessment, we have taken into account the following independent factors: socio-economic status of students, qualification and age of teaching staff, the number of participants and winners of academic competitions (on different levels), the number of students with police records, etc. Moreover, the models take into account the territorial factor: regional center, small town, country and village, ungraded schools. Probably, before constructing the models of educational achievements of schools/students and counting the ratings it is worth to combine schools into similar groups (clusters), where the fluctuation (variance) of independent factors could be significantly lower than the same fluctuation/variance between groups. Herewith, the dividing into the clusters can be realized with logical, intuitive assumptions [12], or based on strict mathematical algorithms [13].

2.2. Constructing the cluster model of Tomsk region schools

In this paper, we divided schools into the clusters with the K-means method [14]. We classified the schools using multi-dimensional vector of 12 variables (not including results of USE of Russian Language and Mathematics). All objects are divided into three clusters. Based on preliminary analysis of impact of different factors on educational achievements of students, we selected these 12 of 188 variables (see table 1).

Table 1. The variables, used in cluster analysis.

No.	Full name of variable
1	The part of pedagogues - psychologists + speech and language pathologists
2	The part of pedagogues of additional education

3	The part of complete families
4	The part of families with both parents working
5	The part of families with both parents having higher education
6	The part of families with one of the parents having higher education
7	The part of families living in socially dangerous conditions
8	The part of students having a police records (or other services)
9	Total amount of students studying profile programs at the 10-11th grades
10	Total amount of classes having profile programs at the 10-11th grades
11	Total amount of students having "Good" and "Excellent" marks at basic school
12	Total amount of students, having " Good" and "Excellent" marks at the secondary school

For all cluster algorithms we have to realize the estimation of the distance between clusters or/and objects. Due to different types of scales, used in different measurements, we have to standardize the source data. Likewise, for similar types of scales, but large range of values. In this paper was applied the Z-Scores standardization, which transforms all variables to the -3...+3 range. All statistic investigations and results analysis were performed with licensed system STATISTICA [15].

2.3. Analysis of the obtained results

Thus, all schools of Tomsk region were grouped into three clusters: the 1st cluster contained 85 schools, the 2nd cluster – 100 schools, the 3rd – 37 schools. Schools with at least one missing variable value were excluded from the analysis. Finally, 222 of 229 schools were analyzed. The Table 2 shows the results for cluster 1.

Table 2. Characteristics of the variables for 1 cluster.

No.	Variable	Descriptive Statistics for Cluster 1 (CLASTER_standart) Cluster contains 85 cases		
		Mean	Standard	Variance
1	Part_psychol+language path.	0,718942	0,870955	0,758564
2	Part_ped_additional_education	-0,065719	0,001442	0,000002
3	Part_complete_families	0,361506	0,701680	0,492355
4	Part_both_parents_working	0,453319	0,699908	0,489871
5	Part_2_higher_education	0,156827	0,647466	0,419212
6	Part_1_higher_education	0,341110	0,782707	0,612630
7	Part_dangerous_conditions	-0,214154	0,324629	0,105384
8	Part_records	-0,182010	0,412189	0,169899
9	Profile_prog_12	-0,050176	0,623384	0,388608
10	Profile_class_12	-0,160689	0,460577	0,212131
11	Good_basic_12	0,205782	0,573367	0,328750
12	Good_second_12	-0,017615	0,370397	0,137194

Table 2 contains standardized data: mean values of variables – column 3, standard deviation – column 4, dispersion – column 5. Similar data was obtained for clusters 2 and 3. Graphically we show mean values for three clusters in the Figure 1.

We can see, that most of mean values differ significantly for three clusters.

During the objects clustering, the distance between groups of objects is usually calculated. In this paper, we applied Euclidean distance, evaluated with the method of "Nearest neighbor". The obtained distances between clusters in multidimensional vector (12 variables) are shown in Table 3.

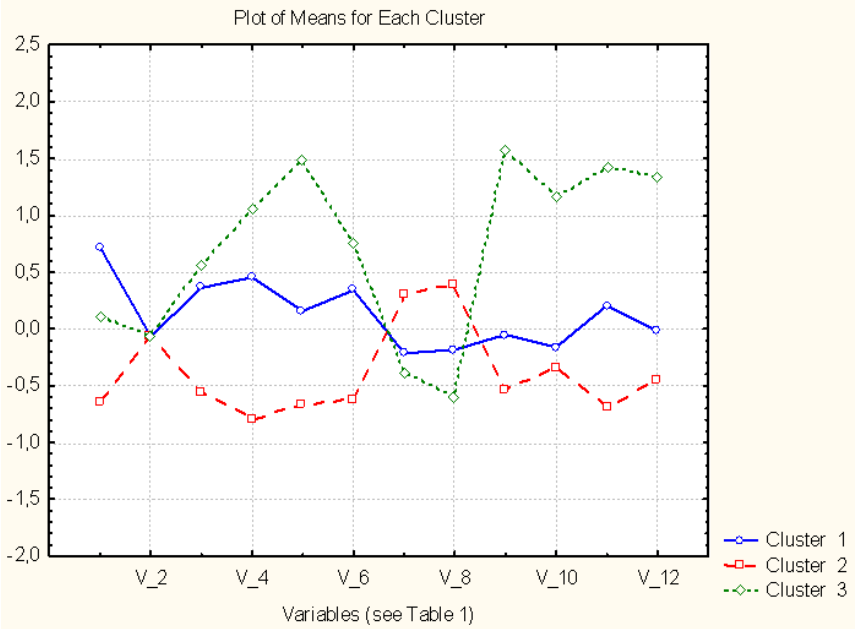

Figure 1. Mean values of variables for three clusters (Tomsk region).

Table 3. Euclidean distance between clusters.

Cluster No.	Euclidean Distances between Clusters (CLASTER_standart). Distances below diagonal. Squared distances above diagonal		
	No. 1	No. 2	No. 3
1	0,000000	0,638171	0,883301
2	0,798856	0,000000	2,285442
3	0,939841	1,511768	0,000000

The obtained results demonstrate that clusters 2 and 3 are maximally different, while clusters 1 and 2 have minimal distance.

In order to estimate the quality of regional schools clustering we obtained the variance within and between clusters. These values are presented in Table 4. Note: the higher value of variance between clusters and the lower they within clusters simultaneously, the better clustering quality.

Table 4. Values of variance between and within clusters.

Variable	Analysis of Variance (CLASTER_standart)					
	Between SS	df	Within SS	df	F	signif. p
Part_psychol+language path.	85,1710	2	134,6403	219	69,2677	0,000000
Part_ped_additional_education	0,0001	2	0,0004	219	8,2353	0,000356
Part_complete_families	53,9155	2	165,4414	219	35,6848	0,000000
Part_both_parents_working	120,1256	2	93,2658	219	141,0352	0,000000
Part_2_higher_education	129,3946	2	90,7735	219	156,0886	0,000000
Part_1_higher_education	68,6462	2	127,5388	219	58,9370	0,000000
Part_dangerous_conditions	19,0599	2	208,1572	219	10,0264	0,000068
Part_records	31,0797	2	195,9622	219	17,3668	0,000000
Profile_prog_12	118,8549	2	103,5953	219	125,6294	0,000000
Profile_class_12	64,1073	2	121,5759	219	57,7397	0,000000

Good_basic_12	123,5569	2	100,0709	219	135,1989	0,000000
Good_second_12	86,5515	2	138,9364	219	68,2138	0,000000

While analyzing the results in Table 4, we have to consider that columns 2, 4 contain sum of squares of the difference, not variances (STATISTICA). Variances can be obtained by dividing values from columns 2, 4 to values from column 3, 5 (the number of degree of freedom) accordingly. The quality of school clustering we can estimate with values F (Fisher–Snedecor statistics) and with the probability p (columns 6 and 7 accordingly). F and p parameters characterize the contribution of the variable to clustering. The best clustering is achieved with maximal values of Fisher–Snedecor criteria and minimal values of probability (less than 0.05). In result of classification we can see, that all 12 variables have great influence on regional school clustering.

2.4. Analysis of school performance within different clusters for Tomsk region

During clustering, we considered socio-economic status of students, qualification and age of teaching staff and other parameters, but we did not take into account the educational efficiency. Thus, a question of accordance of educational achievements within different classes became very important.

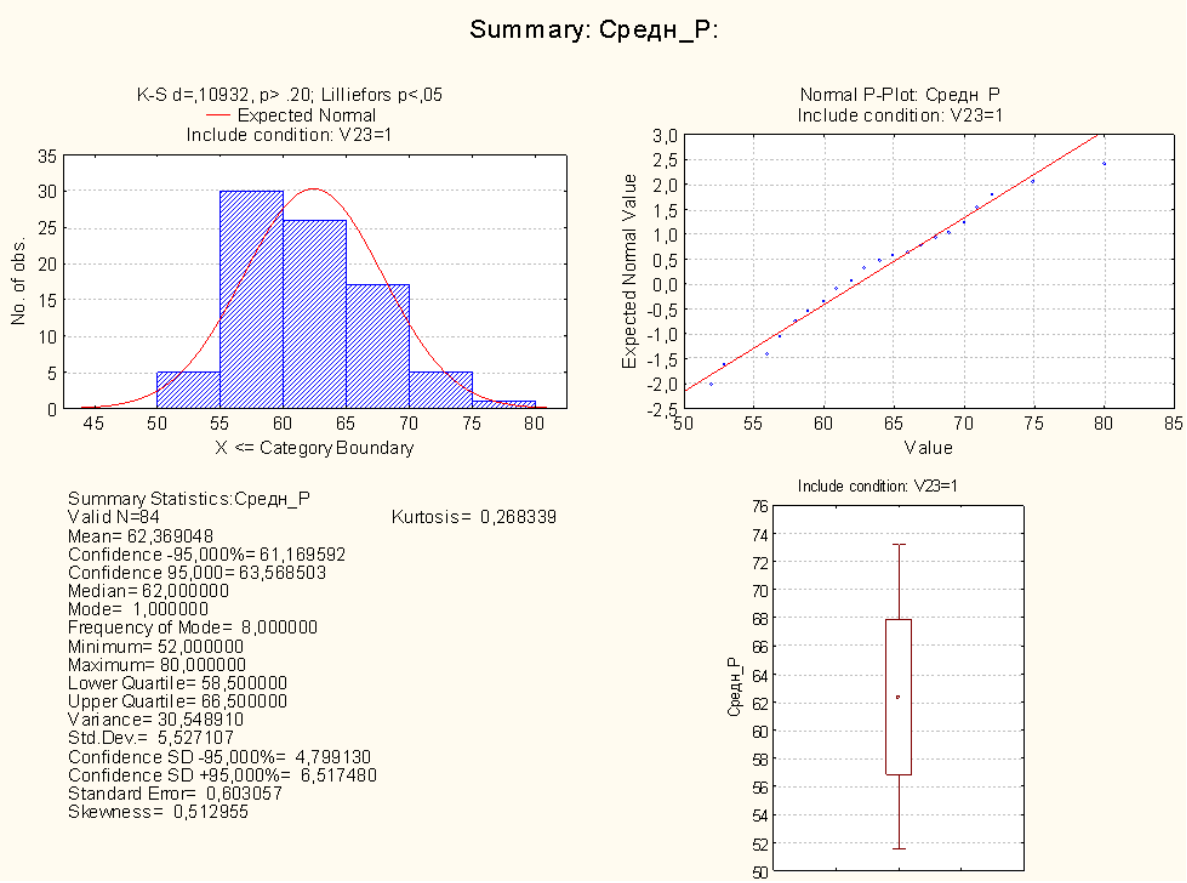

Figure 2. The average score estimation of USE (Russian Language) for 1 cluster.

Education quality of students was estimated using the results of USE in Russian Language and Mathematics. We calculated base and average scores for both disciplines. As a result of investigation we considered that base and average scores have similar presentation, therefore we will discuss only the average scores of USE.

Since for this purpose we applied point and interval estimations, the normal distribution was verified. The example of this investigation [16] is shown in the Figure 2.

The criteria of Kolmogorov–Smirnov and Lillefors [17] confirmed the theory of normal distribution of the scores, and we can see this on the normal probability plot - empirical data have a clear match with theoretical line with insignificant deviation.

Analysis of the results of USE in Mathematics within the first cluster, as the same investigations within second and third clusters also confirmed the validity of the hypothesis of normal distribution.

We collected the scores for different clusters in the Table 5, compared and analyzed them.

Table 5. Results of USE within 3 clusters (Tomsk region).

Variable	CLASTER_1				CLASTER_2				CLASTER_3			
	N	Mean	Conf.-	Conf.+	N	Mean	Conf.-	Conf.+	N	Mean	Conf.-	Conf.+
Average_R	84	62,37	61,17	63,57	86	58,34	56,90	59,77	37	69,19	67,34	71,04
Average_M	84	40,74	39,31	42,16	86	38,48	36,72	40,23	37	48,86	46,17	51,55

Let's explain the content of the table for the first cluster (other clusters are similar). Column 2 – the sample size (*N*), column 3 – average rating value (*Mean*), columns 4, 5 - the boundaries of the confidence interval (95%) for mean value (left boundary – *Conf.*- and right boundary – *Conf.*+, accordingly). The sample size may be different from the number of schools in each cluster due to exception from the sample those schools where one of four values was absent.

The analysis of results obtained shows that average scores of USE in Russian Language and Mathematics significantly differs between clusters. Thus, students of 3rd cluster have the best scores, students from 1st cluster – lower scores, and students from 2nd cluster of schools have the worst scores.

Further investigations were performed for schools with different territorial location: regional center, small town, country and village, ungraded schools.

The cluster models based on 12 variables were constructed for different types of school using the method of K-means. We calculated point and interval estimations of the results of USE in Russian and Mathematics for each type of school and each cluster.

3. Conclusion

Main results of provided investigations:

- Preliminary research allowed determining 12 of 188 variables, which correlate most significantly with educational quality.
- Constructing cluster models should take into account the territorial locations of schools.
- Having the standardized variables for particular school, we can predict the results of USE with high probability (about 90%).
- The conclusions mostly apply for regional schools and Tomsk schools.
- As for urban and ungraded schools, the probability of the prediction is very small due to short distance between clusters (insignificant difference), so the confidence intervals for average scores of USE in Russian and Mathematics partly overlap.
- Between different clusters and types of schools, the scores of USE in Russian Language vary more widely than the scores of USE in Mathematics.

References

[1] Bochenkov S A and Valdman I A 2013 Interpretation and presentation of USE results: problems and possible solutions *Voprosy obrazovaniya / Educational Studies*. *Moscow* No. 3, 2013 (Moscow: HSE) pp 6–27

[2] Prakhov I A and Yudkevich M M 2012 Effect of family income on USE performance and the choice of university *Voprosy obrazovaniya / Educational Studies*. *Moscow* No. 1, 2012

(Moscow: HSE) pp 126–147
[3] Bessudnov A R and Malik V M 2016 Socio-economic and gender inequalities in educational trajectories upon completion of lower secondary education in Russia *Voprosy obrazovaniya / Educational studies. Moscow* No. 1, 2016 (Moscow: HSE) pp 135–167
[4] Katsman Yu Ya and Temirbaev S K 2016 Statistical analysis of education quality in schools of Tomsk Oblast by assessing grades of graduates from the 9th and 11th classes *Advances in Computer Science Research* Vol 51 (Amsterdam: A Press) pp 45–48
[5] Tomsk region education quality estimation center. InfoCollector Software (School Passport) 2.0. Tomsk. 2014. http://coko.tomsk.ru/files/infomonitoring/InfoCollector.pdf
[6] Katsman Yu Ya, Lepustin A V and Ilyukhin B V 2014 The influence of contextual factors on the assessement of the effectiveness of work of schools in the Tomsk region *Modern Problems of Science and Education* vol 6 (Penza: publishing house Academy of Natural History) pp 1–11
[7] Draper N R and Smith G 1998 *Applied Regression Analysis* (New York: Wiley-Interscience)
[8] Yastrebov G, Bessudnov A, Pinskaya M and Kosaretsky S 2013 Problem of educational results' contextualization: schools, social characteristics of pupils, level of territory deprivation. *Voprosy obrazovaniya / Educational studies. Moscow* No. 4, 2013 (Moscow: HSE) pp 188–246
[9] Katsman Yu Ya, Lepustin A V, Ilyukhin B V, Lepustina E V and Zenkova Z N 2016 The stochastic model of the impact of context factors to educational results of Tomsk school graduates *IEEE Global Engineering Education Conf. (EDUCON)* (Abu Dhabi: UAEproceedings, IEEE). pp 767–771
[10] Jolliffe I T 2002 *Principal Component Analysis* 2nd ed (NY: Springer)
[11] Yeomans K A and Golder P A 1982 The Guttman-Kaiser criterion as a predictor of the number of common factors *Journal of the Royal Statistical Society* Series D (The Statistician) vol 31 No. 3, 1982 pp 221–229
[12] Firsova A V 2014 Cluster model as the instrument of educational system status assessement in Moscow region *Ratings in Education: From One-time Practices to Cultural Solutions* (Moscow: HSE Publishing House) pp 120–124
[13] Kim Jae-On, Mueller Charles W and Klecka William R 1989 *Factor, Discriminant and Cluster Analysis* Translated to russian ed Enyukov I S (Moscow: Finance and Statistics Publ.)
[14] Hartigan J A, Wong M A 1979 A K-means clustering algorithm. *Jornal of the Royal Statistical Society* Series C (Applied Statistics) vol 28 No. 1, 1979 pp 100-108
[15] Sá J 2007 *Applied Statistics Using SPSS, STATISTICA, Matlab and R* (Berlin: Springe)
[16] Borovikov V 2003 *STATISTICA. The Art of Analyzing Data on a Computer: For Professionals* 2nd ed (St. Petersburg: Piter)
[17] Kobzar A I 2006 *Applied Mathematical Statistics. For Engineers and Scientists* (Moscow: Fizmatlit)