Quantitative large-scale study of school student's academic performance peculiarities during distance education caused by COVID-19*

V.A. Yunusov, A.F. Gilemzyanov, F.M. Gafarov, P.N. Ustin, A.R. Khalfieva

Kazan Federal University, Kazan, Russia

Abstract. The paper presents the large-scale analysis results of the distance learning impact caused by COVID-19 and its influence on school student's academic performance. This multidisciplinary study is based on the large amount of the raw data containing school student's grades from 2015 till 2021 academic years taken from "Electronic education in Tatarstan Republic" system. The analysis is based on application of BigData and mathematical statistics methods, realized by using Python programming language. Dask framework for parallel cluster-based computation, Pandas library for data manipulation and large-scale analysis data is used. One of the main priorities of this paper is to identify the impact of different educational system's factors on school student's academic performance. For that purpose, the quantile regression method was used. This method is widely used for processing a large-scale data of various experiments in modern data science. Quantile regression models are designed to determine conditional quantile functions. Therefore, this method is especially suitable to exam conditional effects at various locations of the outcome distribution: e.g., lower and upper tails. The study-related conditional factors include such factors as student's marks from previous academic years, types of lessons in which grades were obtained, and various teacher's parameters such as age, gender and qualification category.
Keywords: *Data Science, Big Data, Python, Dask, Quantile Regression, Conditional Quantile Functions, COVID-19.*
DOI: 10.14357/20790279230113

Introduction

The future of people and their future living standards are closely related to the education they receive. A higher education level is essential for achieving higher living standards. The education system should provide equal opportunities for equal success for everyone, regardless of individual and socio-cultural characteristics, socio-economic status, health status, pandemics and other factors. Failure to achieve educational goals will negatively affect a person throughout his life. Therefore, it is necessary to take a comprehensive approach to the problem of inequality in education. The COVID-19 pandemic is fundamentally changed the society, often exacerbating social and economic inequalities [2]. It is necessary to develop quantitative models based on modern methods of mathematical statistics in combination with BigData methods [12] to quantify the impact of the COVID-19 pandemic on educational systems.

Empirical quantitative analysis in education, psychology, and the social sciences is typically based on linear statistical models such as least squares regression (OLS), analysis of variance or covariance, or weighted linear models (e.g., at multiple levels) to calculate either mean scores of associations between dependent and independent variable or group differences in the dependent variable that controls the other independent variables included to the model. The regression coefficients obtained with such linear modeling approaches are averages, usually corrected for estimates of the effects of covariates present in the model, especially when analyzing observational or quasi-experimental data. In OLS regression-based modeling approaches, the regression coefficients show the "influence" of the independent variable x on the mean of the dependent variable y, taking into account the influence of the remaining independent variables [7,11].

* The study (all theoretical and empirical tasks of the research presented in this paper) was supported by a grant from the Russian Science Foundation, project № 22-28-00923, "Digital model for predicting the academic performance of school-children during school closings based on big data and neural networks".

Quantile regression was introduced nearly 30 years ago as an extension of the typical regression model (OLS) and addresses the shortcomings of the typical regression model by allowing to conditionally estimate different points (called quantiles) in a score distribution [10]. This method has become a comprehensive approach in linear and non-linear response models for conditional quantile functions. The quantile regression method, based on minimizing the residual of the "testing function", allows to evaluate all conditional quantile functions, just as the classical linear regression methods, based on least squares estimation, offer a mechanism for estimating conditional means of functions. In this sense, the regression median is a special case of the quantile regression model because the median is the 0.50 quantile (or 50th percentile). Thus, the quantile regression method is gradually becoming a unified statistical methodology and is widely used in education, economics, biology, ecology, finance, econometrics, statistics and applied mathematics [18].

In addition to assessing the impact of variables on different parts of distribution, quantile regression method has a number of other advantages over OLS. Firstly, it gives less weight to outliers of dependent variable compared to OLS. Secondly, it is a more reliable method because it allows to distinguish marginal effects of independent variables in quantiles of the dependent variable. Thirdly, when the errors are not normal, quantile regression estimates can be more efficient than OLS estimates. Finally, the semi-parametric nature of the approach weakens the restrictions on the constancy of parameters throughout the distribution of the dependent variable [6,7].

In this study, we used quantile regression method to study the impact of COVID-19 pandemic on school student's achievements by using a large data set covering data from all schools of Tatarstan Republic [19]. The data includes student's marks for main subjects for grades from 1 to 11, as well as the results of homework, tests, laboratory work, practical works, essays, tests, answers during the lesson, essays, presentations, dictations, etc. The whole database (records from 2015 till 2021) contains more than two billion information units, including information on the progress of more than a million students and the professional activities of more than 120,000 teachers.

The main purpose of this paper is to expand the quantile regression methodology by using BigData methods for using in educational analytics field. We've used quantile regression methodology to build a models of academic performance dependence for the entire academic year on many different factors: on grades for the last academic year and previous mouth, and on various teacher characteristics. This work is aimed to assess the quantitative factors of influence (regression parameters), caused by introduction of distance learning on the school student's academic performance. To solve this problem, we evaluated quantile regressions in the 10th, 25th, 50th, 75th, and 90th quantiles. The distance learning format caused by the COVID-19 epidemic was carried out in schools of the Republic of Tatarstan in April and May of the 2019-2020 academic year, therefore we studied this period in detail. In the 2020-2021 academic year, the schools of the Republic Tatarstan no longer switched to the distance learning and studied full-time as usual.

1. Review of Literature

Recently, a huge amount of various data has been accumulated in various in-formational systems and such databases exists in educational systems too. The analysis of such a large amount of data in the field of education analytics has becoming widespread in order to improve the educational and methodological process [13]. For example, in [17] authors use machine learning to perform an analysis of a large amount of student's data, including their academic performance, in order to identify those who are at risk of being expelled before the end of grade 9. Special emphasis is placed on the use of modern data analysis technologies: classification methods based on trees and support vector machines, thanks to which the forecast efficiency exceeds 90% [17].

Quantile regression is widely used in statistical analysis of educational data. This method is used to investigate the dependence of student's academic achievement in mathematics on factors like family background [18]. The analysis was based on a large sample of data from 2000-2002 years. As a result of the analysis, the authors determine the factors that significantly affect certain quantiles of the distribution of student grades, which makes possible to determine the impact of various indicators on their academic performance [18]. Quantile regression is an appropriate method for evaluating effects in different quantiles, including points in the upper and lower tails of the achievement distribution [14]. Therefore, quantile scores for multiple predictors can be obtained separately in the upper tail of the distribution at the 75th, 80th, 90th, or 95th percentiles [11].

A distinctive feature of quantile regression in the educational field is that it can be used to assess the degree of difference in the influence of factors on weak and strong students [20]. In work [1], by using quantile regression, the analyzed influence of grades received by students in the final year of the school on their per-

formance in the university. As a result of the analysis of a large amount of data, the authors concluded that the average grade for the first half of the graduating class of the school and the grade in the certificate on academic performance at the university has a significant impact. Moreover, for students in the highest quantiles (based on the weighted average score for university courses), this effect is stronger than for students with poor performance. The authors also compare the results obtained with the results of a linear regression model and concluded that quantile regression allows to consider some important aspects of the relationship under study in more detail [1].

Also, quantile regression can be used to assess the influence of students on each other. The work [15] reports the effect of peers on PISA scores by analyzing a large sample of statistical data. The average score of peers in the class and the heterogeneity of their grades is used as independent variables of the regression model, and as the dependent variable is the result of the PISA exam. The authors conclude that for underperforming students these factors have a greater influence, than for students in the highest quantiles (for whom the effect is absent or becomes negative), which suggests that diversity is needed to achieve higher average results.

In addition, the work [4] used the method of quantile regression to analyze da-ta from large-scale studies of educational data of Italian schools conducted by the INVALSI institute. The regression model investigated the dependence of academic performance in mathematics and reading on various characteristics of both the students themselves and on geographical factors. As a result, in this work authors established characteristics that significantly affect student performance: gender, immigrant status and the distribution of performance across all regions of Italy. The authors also concluded that quantile regression is a powerful tool for building a model for multivariate statistical analysis of a large amount of data [4].

Quantile regression is also used to analyze Big Data, because big volumes of datasets make the estimation of regression parameters extremely difficult due to the vigorous computation and the limited storage space. The solution of this problem is described in paper [3]. Authors propose an approach, which simply saves the compact statistics of each data block and uses them to obtain an estimate of all the data with an asymptotically small approximation error, instead of processing all the data together. Another solution is proposed in work [21]. Authors use subsampling algorithm for the following use of composite quantile regression — an improved quantile regression method. Data is split into subsamples, and then the optimal subsampling is used for computing the resulting estimators. To deal with high-dimensional data in work [8] authors developed a new approach by using distributed computing. In this case, only the master machine computes penalized quantile regression estimations, while the other machines only compute subgradient of the local data. The efficiency of the proposed meth-od was confirmed on both the numerical simulation and prerecorded Big Data analysis [8].

2. Methods

2.1. Quantile regression

Quantile regression [10] can be considered as an extension of the least squares method for estimating conditional mean models to estimate an ensemble of models for multiple conditional quantile functions, by taking into account the effect of a set of covariates on the response variable [4] While the classical linear regression model detects the change in the conditional mean of the dependent variable associated with the change in covariates, the quantile regression model detects changes in the conditional quantiles. There-fore, since multiple quantiles can be modeled, a better understanding of how response distributions are affected by predictors can be gained by gaining information about changes in location, distribution, and shape. By analogy with the classical linear regression structure, the linear regression model for the θ-th conditional quantile y_i can be expressed as

$$Q_{y_i(\theta)|x_i} = x_i^T \beta_\theta \quad (1)$$

where y – is a scalar dependent variable, x_i^t - vector of k × 1 independent variables, β – coefficients vector, θ – the conditional quantile of interest, and it is assumed that

$$Q_\theta(u_{i,\theta} \mid x_{i,\theta}) = 0 \quad (2)$$

$u_{i,\theta}$ – residual term of the regression model in θ-*th* quantile.

From Equation 1, it turns out that in comparison to classical linear regression methods based on minimizing sums of squared residuals, quantile regression methods are based on minimizing asymmetrically weighted absolute residuals:

$$\min_\beta \sum_{y_i \geq x_i^T \beta} \theta \mid y_i - x_i^T \beta \mid + \sum_{y_i < x_i^T \beta} (1-\theta) \mid y_i - x_i^T \beta \mid \quad (3)$$

Substituting θ=3, equation (3) gives the median solution, and by using any θ from 0 to 1 allows to study the structure of the dependence anywhere in the conditional distribution of the response variable [4].

The estimation of the coefficients for each quantile regression is based on the weighted data of the entire sample [7].

The $\hat{\beta}_\theta$ coefficient in linear quantile regression models has the same interpretation as in other linear models, i.e.

$$\hat{\beta}_\theta = \frac{dQ_\theta(y_i \mid x)}{dx} \quad (4)$$

means that each coefficient $\hat{\beta}_\theta$ can be interpreted as the rate of change of the θ-th quantile of the distribution of the dependent variable per unit change in the value of the corresponding regressor, keeping the rest unchanged.

However, important differences between least squares regression and quantile regression models relates to monotonic equivariance and robustness to distribution assumptions in conditional quantiles compared to these properties in the conditional mean setting [4].

2.2. Dask-based HPC computational framework

For efficient process of a large amount of unstructured data we have to use Big Data methods, based on the power of computing clusters. In this work we used computational cluster containing 4 virtual machines (each VM has 1TB HDD, 32 GB RAM, 16 CPU cores), with parallel computing framework Dask installed. Dask is a flexible parallel big data processing library, designed to provide scalability and to extend the capabilities of existing Python packages and libraries [16]. The computational framework is based on Dask framework, because it very suitable for processing large datasets and Dask is able to perform computations with data volumes that are larger than the available memory of single computer [9, 5]. Dask has a dynamic task scheduler optimized for cluster based HPC computation, and "Big Data" collections like parallel lists, dataframes and arrays, and extend common interfaces like NumPy, Pandas, or Python iterators running on top of dynamic task schedulers [16].

We obtained anonymized datasets, describing different entities (grades, les-son topics, timetable, information about teachers and students) as separate csv of xml files. The total size of raw data files is more than 120 GB. At the initial pre-processing stage, the raw datasets (csv of xml files) were loaded into data structures called Dask DataFrames. We used Dask's DataFrame. merge() method to merge by some key the data frames obtained by loading different data files, be-cause the raw data has been scattered in different files. Grouped, reduced and aggregated DataFrames obtained from raw datasets subsequently were processed by using quartile regression methods in parallel mode (for different grades, sub-jects, etc). The analysis was carried out by using the quantile regression method implemented in the statsmodels.formula.api library.

As an example, here we briefly present Python scripts and a diagram describing the process of Dask framework based parallel calculation of quantile regression coefficients for one case (quantile regression coefficients calculation for specific grade and subject) (Fig. 1).

Data processing in parallel mode is started by calling student's mean marks containing dataframe's apply method, and by specifying the corresponding method name (ProcSubjects), which must be executed in parallel mode as a parameter of dataframe's apply method. To perform parallel processing for distinct grades, the following call to the apply method ProcessGrade is used. The example of one kind of pipeline

Algoritm1

```
def ProcessSubject(df_marks_subject):
    #quantile regression calculation for spesific grade and subject
    #.............................
    #.............................
def ProcessGrade(df_marks_grade):
    return df_marks_grade.groupby(['subject_title']).apply(ProcessSubject)
dask_job=df_marks.groupby(['grade_number']).apply(ProcessGrade)
result=dask_job.compute()
```

Algorithm2

```
def ProcessGrade(df_marks_grade):
    #quantile regression calculation for spesific grade and subject
    #.............................
    #.............................
def ProcessSubject(df_marks_subject):
    return df_marks_subject.groupby(['grade_number']).apply(ProcessGrade)
dask_job= df_marks.groupby(['subject_title']).apply(ProcessSubject)
result=dask_job.compute()
```

Fig. 1. Examples of Python scripts for Algorithm1 and Algorithm2

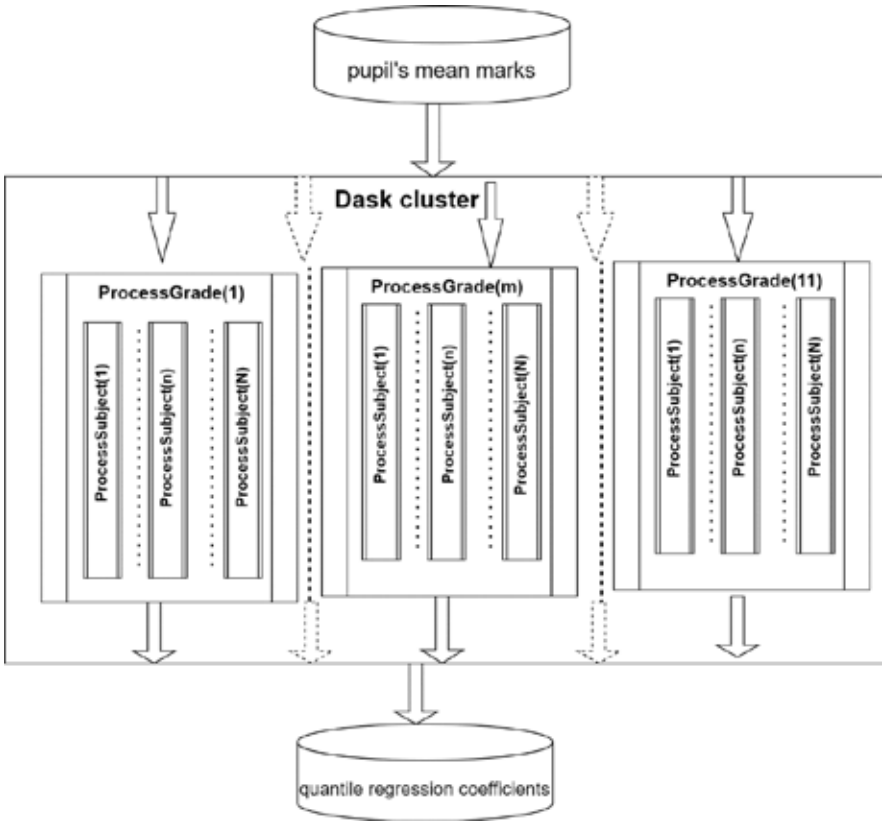

Fig. 2. Dask distributed based computational algorithm architecture for Algorithm 1

(Algorithm1) on Dask-based distributed data processing frame-work is shown schematically in Fig. 2.

The calculation process started in Dask cluster effectively executed in parallel mode (Fig. 2). The results presented in these graphs show the acceleration of the computing process when using a computing cluster in comparison with system based only 1 VM. For evaluation of the computational efficiency computations performed by cluster, we conducted a comparative analysis of the speed of performing calculations for Algorithm1 and Algorithm2, using only one node (1 VM) and full computational cluster (4 VMs). We also analyzed the influence of the npartitions parameter, which sets the number of data partitions into which Dask splits the initial dataframe at the beginning of processing (Fig. 3). By using a parallel algorithm

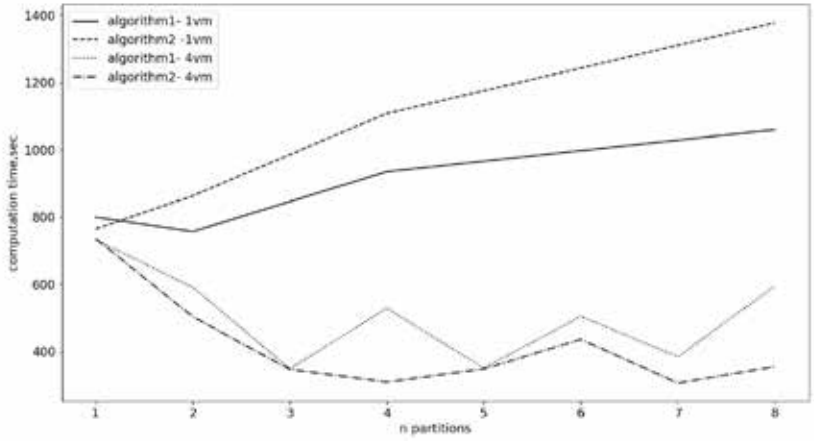

Fig. 3. Comparative analysis of computation time for Algorithm1 and Algoritm2 on a computing cluster consisting of 4 virtual machines and one virtual machine

based on the Dask cluster (with 4 VMs) speeds up the calculation process by almost 3 times (from 786 seconds to 280 seconds for Algorithm 2). The value of the npartitions parameter greatly influences the speed of calculations, with an increase this parameter value decreases the calculation time if the full cluster (4 VM) is used. Increasing this parameter value in the computational system containing only using 1 VM leads to an increase in the calculation time. The graphs also show that the use of Algorithm2 gives slightly higher performance compared to Algorithm1.

3. Results

The analysis of the data was carried out in several stages, divided according to the objects of the distinct study, and mostly performed in parallel mode by using Dask framework. Basically, we have maximally carefully analyzed 25% and 75% quantiles. The most significant results are presented in Fig. 4-7. The central line shows the values of the regression coefficients, dotted lines show 95 % confidence intervals for the regression coefficient's distribution.

At the first stage, we studied the dependence of school student's marks on the characteristics of teachers: age, qualification category, and gender for different subjects (Fig. 4). For comparison, we also provided the values of the quantile regression coefficients for the 2018-2019 academic year (full academic year without distance learning) Basically, before the of the distance learning begin in April month, the distribution of the quantile regression coefficient is the same as for all previous months, which indicates a similar distribution of marks for all students in these months.

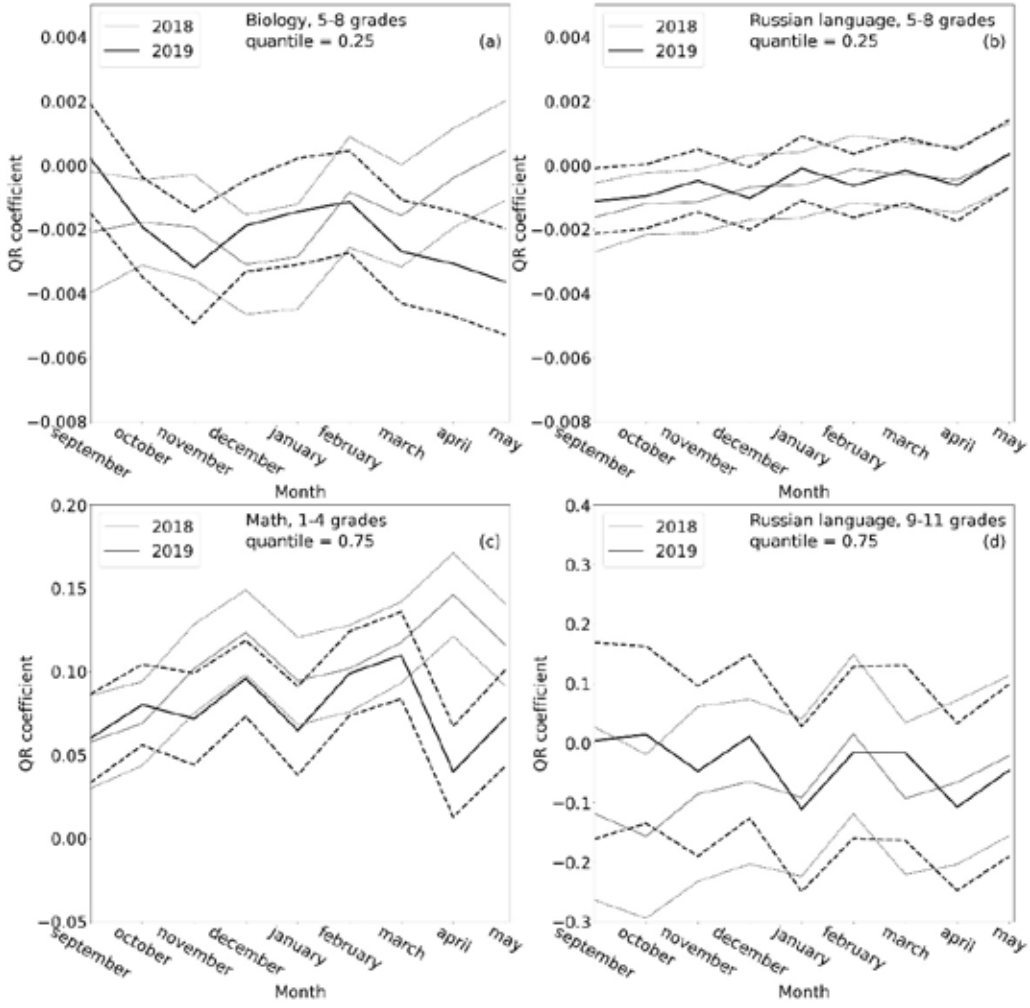

Fig. 4. Distribution of quantile regression coefficient describing the influence of the teacher's characteristics on student's grades for 2018-2019 and 2019-2020 academic years: (a) regressor variable - teacher age, 25% quantile, grades 5-8, subject Biology; (b) regressor variable - teacher age 25% quantile, grades 5-8, subject Russian language; (c) regressor variable - teacher category 75% quantile, grades 1-4, subject Mathematics; (d) regressor variable – teacher's gender, 75% quantile, grades 9-11, subject Russian language

Significant differences in quantile regression coefficient appears in April and May months for some subjects, and quantiles. In the case of a teacher's age as a regressor variable, differences appear in biology (Fig. 4(a)) and foreign language for 25% quantile (grades 5–8, not shown), i.e., for low-achieving students. But in other groups considered or in other subjects, there is no statistically significant differences (for example see Fig. 4(b)). By using qualification category as a regressor variable significant differences appeared in one of the main subjects: mathematics (Fig. 4(c)), geometry and algebra for all age groups and quantiles of 25% and 75%, this is especially noticeable for the April month. By using teacher's age as a regressor variable we discovered, that the distribution of the regression coefficient did not change significantly during distance education (for example see Fig. 4(d)).

We also noticed, that the values of the regression coefficients in the models for 2018-2019 academic year, describing the dependence of school student's marks on the characteristics of teachers, are usually higher than in the such models for the 2019-2020 academic year. This fact indicates that after the introduction of distance learning, the teacher's features began to play a smaller role in the distribution of regression coefficients for all students.

At the next stage, the dependence of marks for all subjects for the 2019–2020 and 2020–2021 academic years on marks for the previous academic year was studied (student mean marks for the previous academic year were taken as a regressor variable). In Figure 5 we present of the regression coefficients values for individual months, and for different subjects. It can be seen that immediately after beginning of the distance education (April-May 2020), the dependence of

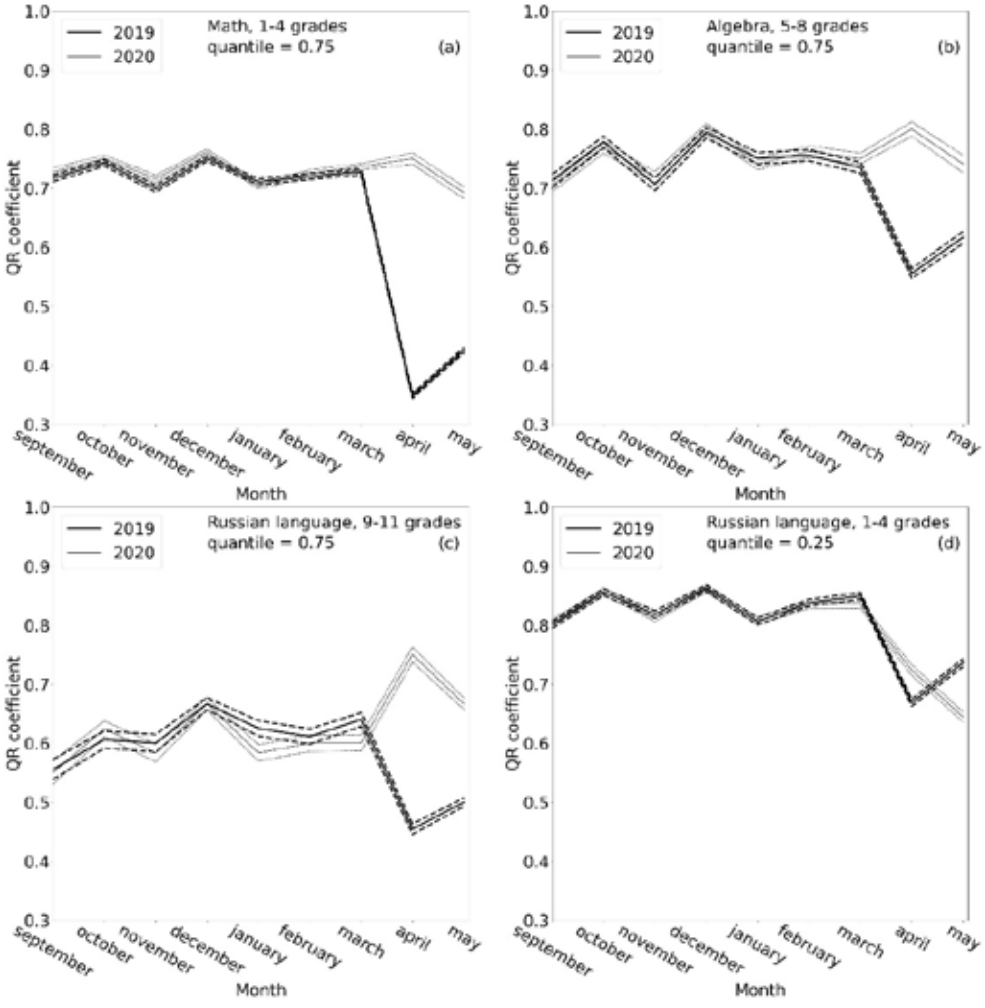

Fig. 5. The quantile regression coefficients, describing the dependence of marks on the previous academic year marks, for all subjects for the 2019–2020 and 2020–2021: (a) 75% quantiles, grades 1-4, subject Mathematics; (b) 75% quantile, grades 5-8, subject Algebra; (c) 75% quantile, grades 9-11, subject Russian language; (d) 25% quantile, grades 1-4, subject Russian language

grades on the grades of the previous year decreased significantly (the drop was up to 0.4, see Fig. 5(a)), which can be explained by the period of adaptation to the new format of education. Maximal drop of regression coefficient values is observed for 75% quantile (Fig. 5(a, b, c)), whereas for 25% quantile the lower values of this is drop observed (Fig. 5(d)). This feature is observed for all subjects and for different age groups of students. For the 2020-2021 academic year, there is no such sharp decline, what indicates the normalization of the educational process for this time.

A similar analysis of quantile regression coefficients conducted also for different types of assessment without division to subjects (Fig. 6). Here we observed a sharp decrease in the quantile regression coefficient for students of all age groups for "classwork" and "homework", 75% quantile (Fig. 3(a, c)) in April and in May of 2019-2020 academic year. At the same time, the regression coefficients for "control work" remained practically unchanged (Fig. 3(b)). Also, the regression coefficients for lover quantiles does not changes significantly (Fig. 3(d)). But just a year after the introduction of distance learning, the regression coefficient did not decrease, and in some cases even increased significantly (the increase was up to 0.35, see Fig. 3(d)).

At the final stage of the study, we analyzed the coefficients of quantile regression model describing the dependence of marks in the month of distance education start (April) on marks in the previous month (February). The detailed analysis was performed for distinct subjects and distinct types of assessment (the most characteristic results are shown in Figure 7). According these data, we concluded that during distance education for students from the 10%, 25% and 50% quantiles, nothing has changed in terms of academic subjects and types of assessment. However, for students in 75% and 90% quantiles, the transition to distant education has a critical impact, due to which the dependence of grades during distance education on grades for the pre-distance period decreased significantly (see, Fig. 7).

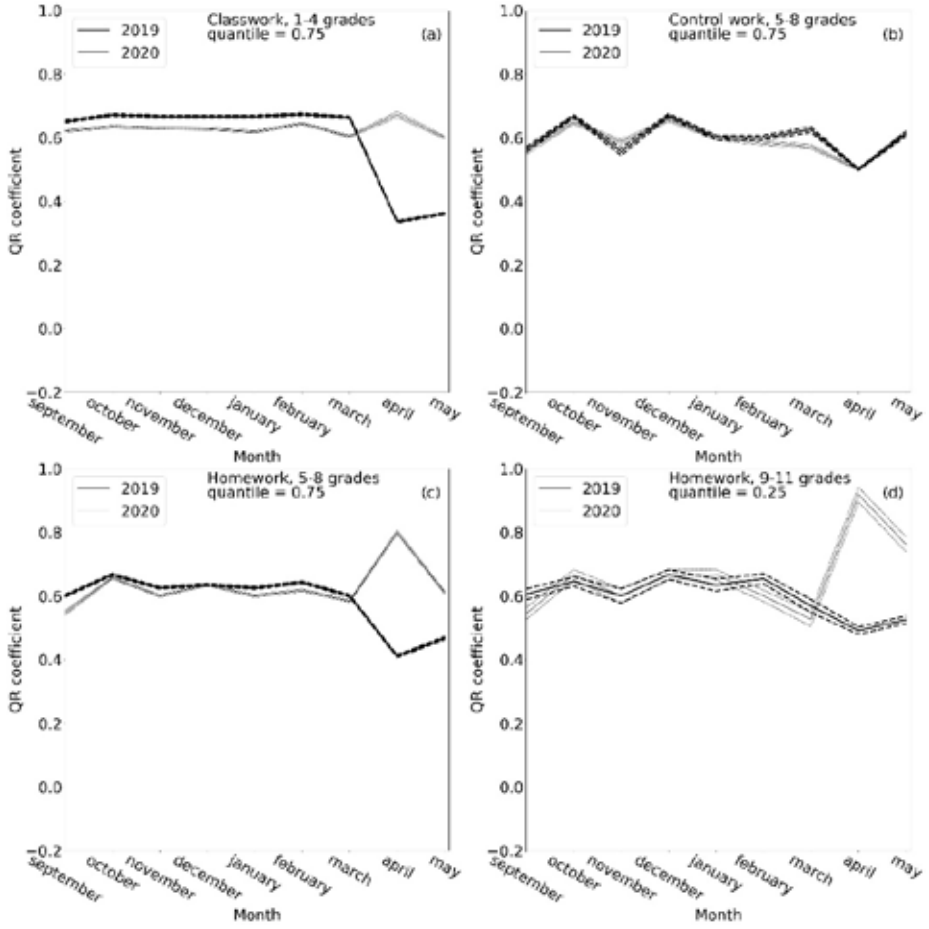

Fig. 6. The quantile regression coefficients for different types of assessment for 2019-2020 and 2020-2021 academic years: (a) 75% quantile, grades 1-4, type of assessment «class work»; (b) 75% quantile, grades 5-8, type of assessment «control work»; (c) 75% quantile, grades 5-8, type of assessment «homework»; (d) 25% quantile, grades 9-11, type of assessment «homework»

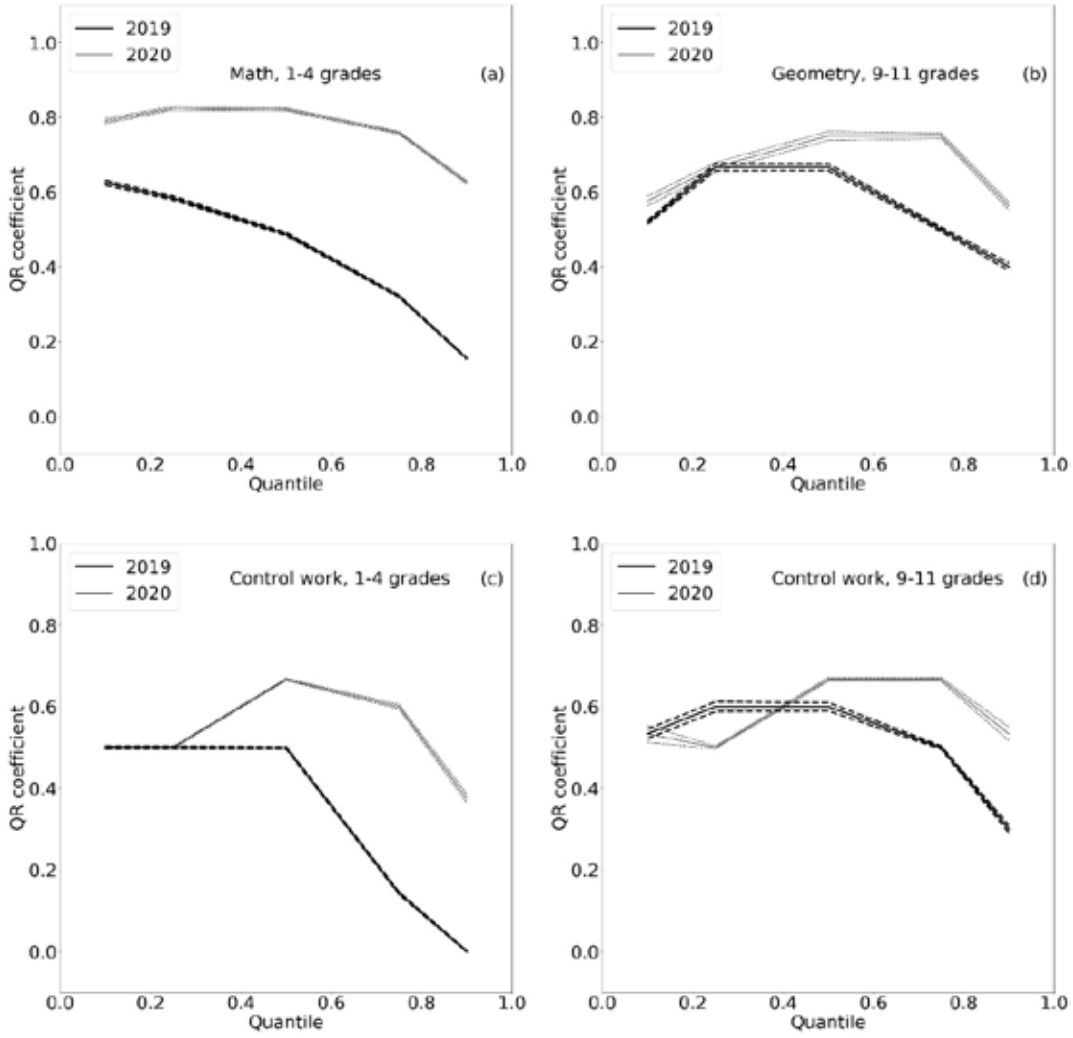

Fig. 7. Distribution of the 2018-2019 and 2019-2020 quantile regression coefficients de-scribing the dependence of April marks on the corresponding February marks: (a) grades 1-4, subject Mathematics; (b) 9-11 grades, subject Geometry; (c) grades 1-4, type of assessment «control work» ; (d) 9-11 grades, type of assessment «control work» . The values of quantile regression coefficients were plotted out against to the corresponding quantiles.

Conclusions

In this paper, we analyzed school student's academic performance in the peri-od before and during distance learning caused by COVID-19. The analysis per-formed on the basis of data obtained from "Electronic education in the Republic of Tatarstan" system. The analysis and interpretation of distance education effect is performed by using quantile regression approach. The Big Data processing framework Dask is used as a basis of data processing systems computational architecture. We developed high-performance cluster-based data processing pro-gram scripts for efficient quantile regression coefficients calculation in parallel mode, and performed analysis of the proposed algorithm's computational speed.

In the course of the study, we established that after transition to a distance learning, the first two months (April and May of 2019-2020 academic year) showed the greatest differences in the parameter values of quantile regression model, what indicate a period of adaptation to the new learning model. Statistically significant differences are existing both in the dependence of marks on the teacher's features, and for regression model coefficients for different years of study. Also, the parameters of quantile regression models significantly differ for different quantiles. Thus, it was possible to establish that during transitional moment (the period from February to April), the quantile regression coefficients for the group of students with high academic performance dropped sharply. This means that the dependence of April grades on

February grades has decreased in the 2019-2020 academic year, what means that if a student received high grades before, then new grades are less determined by old ones. These findings suggests that there are significant alterations in academic performance in different groups of students immediately after transition to distance learning format in April 2020. Moreover, there may be differential study-related factors effects at different points in the conditional distribution of school students' academic performance.

Marks in mathematics, geometry, algebra began to depend more on the teachers' qualification category for "weak" and "strong" students (changes are insignificant for "average" students). This result shows that in the face of unexpected and fundamental changes in the educational process, teachers' qualification is one of the stable factors influencing the assessment of the academic success of students in the field of technical disciplines. We also conclude that older teachers were worse adapted to the distance learning format during the period of mass digitalization of the educational process and assessed the progress of students in a simplified format.

Students did not immediately join the distance learning format, this caused them difficulties, therefore, their grades in the most difficult and important sub-jects decreased. It is easier for students to learn mathematics offline. As for the humanities and natural sciences, they are easier to perceive in a distance learning format. It was more difficult to learn such subjects as physics, algebra, geometry and the Russian language for school students in a distance learning format. Stronger students who are accustomed to express themselves in the classroom lose this opportunity in the distance learning format and therefore have lower grades compared to last year. Here we can talk about how limited the possibilities of the distance learning for-mat in the educational process are. This is also evidenced by the fact that by the new academic year the educational process has normalized.

During the distance learning format, the opportunities for the manifestation of each student are limited. Most likely, there is a distraction factor, when online students are more distracted, less focused on the educational process. While in the offline learning format, each student is in front of the teacher, everyone is included in the process of education, and strong students have more opportunities to actively manifest themselves, get involved in the process, and compete with each other. Here we can say about the importance of the influence of the educational environment on the manifestation of strong students, so on their grades. This confirms the fact that the test scores have not changed much. With the introduction of distance learning, strong students have sharply decreased their motivation to learn. Probably, for successful students, the opportunity to express themselves, communicate, and be included in the educational process as much as possible plays a big role. With the introduction of a distance learning format, these opportunities are reduced. Therefore, those students who had high and very high academic performance were no longer included in the learning process due to a decrease in interest in it. This was especially pronounced in basic subjects, as mathematics and the Russian language.

Thus, it can be concluded that the introduction of quarantine and distance learning format had a greater impact on students with high academic performance. This is especially true for such subjects as algebra, geometry, Russian language. It can be assumed that the online format does not allow to fully integrate into the educational process, and those students who can be actively involved in the traditional format, are interested in the learning process, lose this opportunity and the ability to actively participate in the online format.

References

1. *Amerise, I.L.* Predicting Students Academic Achievement: A Quantile Regression Approach. International Journal of Statistics and Systems 13(1), 9–14 (2018).
2. *Aspachs O, Durante R, Graziano A, Mestres J, Reynal-Querol M, et al.* (2021) Tracking the impact of COVID-19 on economic inequality at high frequency. PLOS ONE 16(3): e0249121. https://doi.org/10.1371/journal.pone.0249121
3. *Chen, L., Zhou, Y.* Quantile regression in big data: A divide and conquer based strategy. Computational Statistics & Data Analysis 144, 106892 (2020). https://doi.org/10.1016/j.csda.2019.106892 .
4. *Costanzo, A., Desimoni, M.* Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data. Large-scale Assess Educ 5, 14 (2017). https://doi.org/10.1186/s40536-017-0048-4
5. *Gafarov F, Minullin D, Gafarova V.* Dask-based efficient clustering of educational texts. CEUR Workshop Proceedings, 3036, 362–376 (2021).
6. *Gürsakal, Necmi & Murat, Dilek.* (2018). Assessment of PISA 2012 Results With Quantile Regression Analysis Within The Context of Inequality In Educational Opportunity. alphanumeric journal. 4. 41-54. https://doi.org/10.17093/aj.2016.4.2.5000186603 .
7. *Hao, L., Naiman, D.* Quantile regression. Sage, London (2007).

8. *Hu, A., Li, Ch., Wu, J.* Communication-Efficient Modeling with Penalized Quantile Regression for Distributed Data. Complexity, 2021, 6341707 (2021). https://doi.org/10.1155/2021/6341707
9. *Henriques, J., Caldeira, F., Cruz, T., Simões, P.* Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. Electronics 9, 1164 (2020). https://doi.org/10.3390/electronics9071164
10. *Koenker, R., Basset, G.* Regression quantiles. Econometrica, 46, 33–50 (1978). https://doi.org/10.2307/1913643
11. *Konstantopoulos S., Li W., Miller S., van der Ploeg A.* Using Quantile Regression to Estimate Intervention Effects Beyond the Mean. Educational and Psychological Measurement 79(5), 883–910 (2019). https://doi.org/10.1177/0013164419837321
12. *Li J., Jiang Y.* The Research Trend of Big Data in Education and the Impact of Teacher Psychology on Educational Development During COVID-19: A Systematic Review and Future Perspective. Front. Psychol. 12, 753388 (2021). https://doi.org/10.3389/fpsyg.2021.753388
13. *Park Y.-E.* Uncovering trend-based research insights on teaching and learning in big data. Journal of Big Data 7 (93), 1–17 (2020). https://doi.org/10.1186/s40537-020-00368-9
14. *Porter, S.R.* Quantile regression: Analyzing changes in distributions instead of means. In: M. B. Paulsen (Ed.), Higher education: Handbook of theory and research, vol. 30, 335–381. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-12835-1_8
15. *Rangvid, B.* School composition effects in Denmark: quantile regression evidence from PISA 2000. Empirical Economics 33, 359–388 (2007). https://doi.org/10.1007/s00181-007-0133-6
16. *Rocklin M.* Dask: Parallel Computation with Blocked algorithms and Task Scheduling. In: Proceedings of the 14th Python in Science Conference, pp. 126–132, (2015) https://doi.org/10.25080/Majora-7b98e3ed-013
17. *Sorensen, L.* "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk. Educational Administration Quarterly 55, 404–446 (2019). https://doi.org/10.1177/0013161X18799439
18. *Tian, M.* A Quantile Regression Analysis of Family Background Factor Effects on Mathematical Achievement. Journal of Data Science 4, 461–478 (2006). https://doi.org/10.6339/JDS.2006.04(4).283
19. *Ustin, P., Sabirova E., Alishev T., Gafarov F.* Key Factors of Teacher's Professional Success in the Digital Educational Environment. ARPHA Proceedings 5: 1747-1761 (2022) https://doi.org/10.3897/ap.5.e1747
20. *Yu, K.* Quantile Regression: Applications and Current Research Areas. Journal of the Royal Statistical Society Series D (The Statistician) 52(3), 331–350 (2003). https://doi.org/10.1111/1467-9884.00363
21. *Yuan, X., Li, Y., Dong, X., Liu T.* Optimal subsampling for composite quantile regression in big data. Statistical Papers (2022). https://doi.org/10.1007/s00362-022-01292-1

Yunusov V.A. Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: valentin.yunusov@gmail.com

Gilemzyanov A.F. Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: gilemal59@gmail.com

Gafarov F.M., PhD**,** Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: fgafarov@yandex.ru (correspondent author)

Ustin P.N. PhD, Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: pavust@mail.ru

Khalfieva A.R. PhD, Kazan Federal University, Kremlyovskaya St, 18, Kazan, Respublika Tatarstan, Russia, 420008, e-mail: khalfieva@inbox.ru