IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Restoration of omissions in the quarterly indicators of financial statements for the other financial institutions in the Bank of Russia[1]

Anna Borisenko, Denis Koshelev, Petr Milyutin and Alieva Piruza,
Bank of Russia

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Restoration of omissions in the quarterly indicators of financial statements for the Other Financial Institutions in the Bank of Russia

Piruza Alieva, Anna Borisenko, Petr Milyutin, Denis Koshelev

Abstract

Quarterly financial accounts and sectoral balance sheets' statistics in the context of financial instruments and sectors of the economy, formed on the basis of microdata, is a reliable information basis for a comprehensive and deep macroeconomic analysis. Most organizations in the Financial Corporations sector (S12) report on an annual and quarterly basis, but for some of them, including organizations of the Other Financial Institutions subsector (S125), which perform non-licensed activities, data is only available on an annual basis. On a quarterly basis, only a small part of these organizations' reporting is available. Therefore, to ensure the completeness of the range of companies in the formation of statistics of financial accounts and sector balance sheets, it is necessary to restore gaps in the quarterly indicators of financial statements of organizations. In this article the results of restoring omissions in the quarterly indicators of financial statements for the Other Financial Institutions subsector (S125) in the Russian Federation, which perform non-licensed activities, are presented. In particular, the results of the traditional methods (regression analysis, individual growth rates, cluster analysis) and Machine learning-based methods, that can be applicable to recover data, such as random forest and generative neural network.

Keywords: financial accounts and sectoral balance sheets' statistics, Other Financial Institutions subsector in the Russian Federation, restorations of omissions in the quarterly indicators, Machine learning-based methods

JEL classification: C8

Contents

Introduction

In accordance with Clause 16.1 of Article 4 of Federal Law No. 86-FZ, dated 10 July 2002, 'On the Central Bank of the Russian Federation (Bank of Russia)', the Bank of Russia develops the methodology for compiling the Russian Federation financial accounts in the System of National Accounts (SNA) and organises the compilation of the Russian Federation financial accounts (hereinafter, SNA financial accounts). The Bank of Russia's obligations to form the measures of the financial accounts and financial balance sheets of the SNA on a quarterly and annual basis are also stipulated in Recommendation No. 8 'Sectoral Accounts' within the G20 Data Gaps Initiative (DGI-II). The Bank of Russia has been publishing the financial accounts and financial balance sheets, being part of the System of National Accounts of the Russian Federation, since 2015 on a quarterly and annual bases. The Bank of Russia relies on the System of National Accounts 2008 (2008 SNA)[1] manual as a conceptual and methodological framework for compiling financial accounts and financial balance sheets.

Statistics provided in quarterly financial accounts and sectoral balance sheets, broken down by financial instrument and economic sector, which are compiled based on microdata, expand the opportunities for enhancing the efficiency and the depth of macroeconomic research. This in turn improve the understanding of interconnections between the real sector and the financial industry of the country's economy.

A significant advantage of SNA financial accounts is the fact that they are an important source of data used to analyse activity in the economic sectors failing to provide detailed information, e.g. in the subsector 'Other financial corporations' (of the sector 'Financial corporations') comprising a large number of organisations not reporting to the Bank of Russia.

The key challenge in compiling statistics on organisations in the subsector 'Other financial corporations' is that, in contrast to the majority of organisations in the sector 'Financial corporations' (S12) regularly reporting to the Bank of Russia on a monthly, quarterly and annual basis, information on all organisations in the subsector 'Other financial corporations' (S125) can be obtained only on an annual basis. The main source of data for compiling statistics on the subsector 'Other financial corporations' is annual accounting (financial) statements producing the main portion of processed statistics (whereas quarterly statements are only submitted by a small number of these organisations).

In this regard, when quarterly SNA financial accounts are compiled, it is essential to have comprehensive information as of quarterly dates when only a part of organisations submit their reporting, as well as to ensure that annual and quarterly statistics are comparable. For this purpose, it is needed to close data gaps in organisations' accounting statements as of quarterly dates.

[1] System of National Accounts 2008 (European Commission, United Nations, Organisation for Economic Cooperation and Development, International Monetary Fund, World Bank).

This paper outlines statistical and machine learning methods which will be used to close data gaps in the measures of accounting statements as of quarterly dates for the subsector 'Other financial corporations' of the sector 'Financial corporations' (hereinafter, OFCs). In particular, the paper considers the results of using the individual growth method and cluster analysis, and describes the currently applied method for restoring missing data (the '1/4' and dynamics extension methods). Among machine learning methods, the authors consider the random forest method for regression, gradient boosting model and the generative adversarial network (GAN).

The authors explore the following balance sheet measures (Form No. 0710001) as the data to be restored on the quarterly bases, because these measures are the key ones for compiling SNA financial accounts:

- Loans (short- and long-term ones),
- Accounts receivable,
- Accounts payable,
- Equity and investment fund shares.

The methods employed were compared based on the following criteria.

1) Discrepancies between restored and actual values in an artificial sample as of 1 January 2017 and 1 January 2020.
2) Interpretable dynamics (identification of organisations accounting for a rise or a decline in a particular measure).

Overview of methods for closing data gaps in quarterly measures of accounting statements

There is a vast number of statistical methods generating quarterly data based on annual values. Specifically, the quarterly financial accounts for the 1950s published by the Federal Reserve System were obtained based on the interpolation method.[2] This method estimates quarterly values of various financial measures in the form of a fixed weighted linear combination of annual values for the periods t-1, t and t+1. Thus, interpolation determines unknown intermediate values in a time series as a linear combination of bordering annual values.

To estimate unknown quarterly values of financial measures, the Federal Reserve System also applies the ratio method.[3] Under this method, the first step is to calculate the ratio R_h for each quarter using known data, according to the formula:

$$R_{t,h} = \frac{x_{t,h}}{\sum_{h=1}^{4} x_{t,h}}$$

[2] Board of Governors of the Federal Reserve System, 2000.
[3] Financial Accounts: History, Methods, the Case of Italy and International Comparisons, 2008, p. 118–122.

where $x_{t,h}$ is the base time series with known quarterly values, $h = 1, \dots, 4$ is the order number of a quarter, and $t = 1, \dots, N$ is the order number of annual values.

The resulting ratio is then used to derive the unknown value of a particular measure for the relevant quarter, according to the formula:

$$y_{t,h} = R_{t,h} y_t$$

where $y_t,\ t = 1, \dots, N$ is a series of annual values.

The above methods are easy to use and are in line with time aggregation limits. However, the quarterly series generated by these methods omit possible dynamics of an unknown quarterly series.

Most central banks use the Chow–Lin method[4] to derive quarterly values from annual figures. This method assumes that a time series of annual figures and the dynamics of quarterly values are strongly correlated with each other and have the same order of integration equal one, which means that the above time series are cointegrated. According to the Chow–Lin method, the coefficients of a linear regression model are estimated with the autocorrelation of order 1 errors using the generalised least squares method. Then, based on the estimated coefficients of the model, the quarterly values of variables are derived.

If the assumption of the correlation between the annual values of a particular measure and its disaggregated values (quarterly values) is rejected, Fernández's method[5] or Litterman's method[6] is then employed. Fernández's method estimates the coefficients of a linear regression model with random errors. Contrastingly, Litterman estimates a linear regression model with the autocorrelation of order 2 errors. It should be noted that Litterman extended the method proposed by Fernández with the autocorrelation coefficient equalling zero.

To restore missing values in the measures of accounting statements as of quarterly dates, it is also possible to apply machine learning methods. Specifically, the generative adversarial network (GAN) is one of the most widespread methods used to close data gaps. The GAN is an algorithm of unsupervised machine learning built on a combination of two neural networks. One of them (the generative network) generates candidates (the generative model), whereas the other evaluates them (the discriminative network) trying to distinguish generated candidates from true data (the discriminative model). Hence, the idea of this method is to produce objects that would resemble true ones. This method is often applied to restore missing parts in images. However, the technique of producing realistic objects can also be applied to generate the vectors whose elements are balance sheet measures in accounting statements corresponding to companies' actual behaviour in the market. The Wasserstein GAN (WGAN), which is an extension to the conventional GAN offering an alternative way to train the model, improves the approximation of the distribution of data

[4] Chow, G. and Lin, A. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. The Review of Economics and Statistics 53, 1971, p. 372–375.
[5] Fernández, R. A Methodological Note on the Estimation of Time Series. The Review of Economics and Statistics 63, 1981, p. 471–478.
[6] Litterman, R. A Random Walk, Markov Model for the Distribution of Time Series. Journal of Business and Economic Statistics 1, 1983, p. 169–173.

observed in the training sample. The benefit of the WGAN is that the training process is more stable and less sensitive to the model architecture and the choice of hyperparameter configurations.[7]

It is also possible to find a direct functional dependence between balance sheet measures by using ensemble trees algorithms to solve regression problem. Two of the most popular algorithms used: random forest[8] and gradient boosting[9]. Both models are ensemble decision tree models. In random forest model trees are built independently and final result is mean output of all decision trees. In gradient boosting every new tree helps to correct errors made by previously trained tree, thus, final model is sequentially connected trees. Gradient boosting often is more accurate but prone to overfitting. So, both models were used.

Description of data

Other financial corporations are financial institutions providing financial services, except credit institutions, insurers, pension funds, and financial auxiliaries. OFCs comprise leasing companies, financial holdings, factoring companies, investment companies, mortgage companies, mortgage agents, and other organisations.

A part of OFCs perform activities supervised by the Bank of Russia. These OFCs include pawnshops, microfinance organisations, consumer credit cooperatives and agricultural consumer credit cooperatives, professional securities market participants, and housing savings cooperatives. However, the largest portion of OFCs' financial operations are performed by organisations that are not subject to the Bank of Russia's supervision. This is the group of financial institutions that this paper deals with.

OFCs that are not subject to the Bank of Russia's supervision account for a rather significant portion of information impacting the dynamics of released official statistics on SNA financial accounts. OFCs carrying out activities not supervised by the Bank of Russia account for over 90% of the total balance of OFCs (line 1600 in Form No. 0710001 'Balance sheet' and make more than 70% of their overall number over the entire period under review (Table 1).

[7] Martin Arjovsky, Soumith Chintal, Léon Bottou. Wasserstein GAN, 2017.

[8] Leo Breiman. Random Forests, 2001

[9] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, p. 1189–1232

Table 1. Portion of the total balance and number of OFCs, whose activities are beyond the scope of the Bank of Russia's supervision, in the total balance and number of all organisations classified as OFCs

	Portion of unsupervised OFCs' total balance in all OFCs' total balance	Ratio of the number of unsupervised OFCs to the overall number of OFCs
01.01.2015	97.34%	72.75%
01.01.2016	97.48%	76.97%
01.01.2017	96.07%	77.31%
01.01.2018	95.29%	83.45%
01.01.2019	99.28%	84.47%
01.01.2020	99.54%	85.52%

The main descriptive statistics for the measures of accounting statements as of annual dates (1 January 2016–1 January 2020) for OFCs not supervised by the Bank of Russia are presented in Table 2.

Table 2. Descriptive statistics for the main measures of accounting statements of the subsector of OFCs not supervised by the Bank of Russia, as of annual dates, mln of rubles

		Maximum	Mean value	Standard deviation
01.01.2016	Accounts receivable	360 629.53	60.44	2 081.14
	Accounts payable	345 204.79	54.96	1 948.86
	Loans	420 355.01	156.40	3 898.93
	Equity and investment fund shares	675 510.8	153.52	4 348.28
01.01.2017	Accounts receivable	713 452.34	69.66	3 479.26
	Accounts payable	724 855.60	66.87	3 454.53
	Loans	764 417.40	166.15	5 484.51
	Equity and investment fund shares	522 092.58	153.92	3 760.17
01.01.2018	Accounts receivable	672 230.23	72.46	3 618.33
	Accounts payable	607 838.35	64.71	3 413.29
	Loans	784 792.92	200.77	7 029.77
	Equity and investment fund shares	570 229.35	175.67	4 211.96
01.01.2019	Accounts receivable	644 381.76	97.04	4 056.20
	Accounts payable	645 739.15	90.86	3 934.06
	Loans	1 079 909.35	274.70	8 979.45
	Equity and investment fund shares	1 003 973.45	249.19	6 820.11

01.01.2020	Accounts receivable	698 177.71	110.31	4 002.05
	Accounts payable	684 351.64	96.69	3 660.11
	Loans	1 022 043.26	325.50	9 538.12
	Equity and investment fund shares	1 163 698.12	297.53	8 302.93

According to Table 2, organisations of the subsector of OFCs whose activities are beyond the scope of the Bank of Russia's supervision are rather heterogeneous over the entire period under review. This is evident from the high values of variation indicators (variance, standard deviation, variation coefficient).

The main source of information for compiling quarterly SNA financial accounts of OFCs whose activities are not supervised by the Bank of Russia is primary statistics submitted according to federal statistical forms No. P-3 'Data on organisations' financial standing' (hereinafter, form No. P-3) and No. P-6 'Data on financial investment and liabilities' (hereinafter, form No. P-6). However, the coverage in the above forms had been low during several years (see Table 3). Loans have the highest coverage ratio, while Equity and Accounts payable – the lowest coverage ratios.

Table 3. Dynamics of the coverage according to forms No. P-3 and No. P-6 of the main financial instruments for OFCs whose activities are not supervised by the Bank of Russia, as of annual dates, %

	01.01.2016	01.01.2017	01.01.2018	01.01.2019	01.01.2020
Accounts payable	4.77	18.60	11.46	10.74	13.38
Accounts receivable	16.67	17.90	14.91	22.67	29.02
Loans	30.76	36.36	31.90	56.60	58.79
Equity and investment fund shares	3.37	15.60	16.16	12.98	24.31

As regards OFCs whose activities are beyond the scope of the Bank of Russia's supervision, data on all these organisations are only available on an annual basis. As to quarterly reporting of these organisations, only a small part of it is available. Chart 1 shows changes in the main financial measures of accounting statements of OFCs whose activities are not supervised by the Bank of Russia.

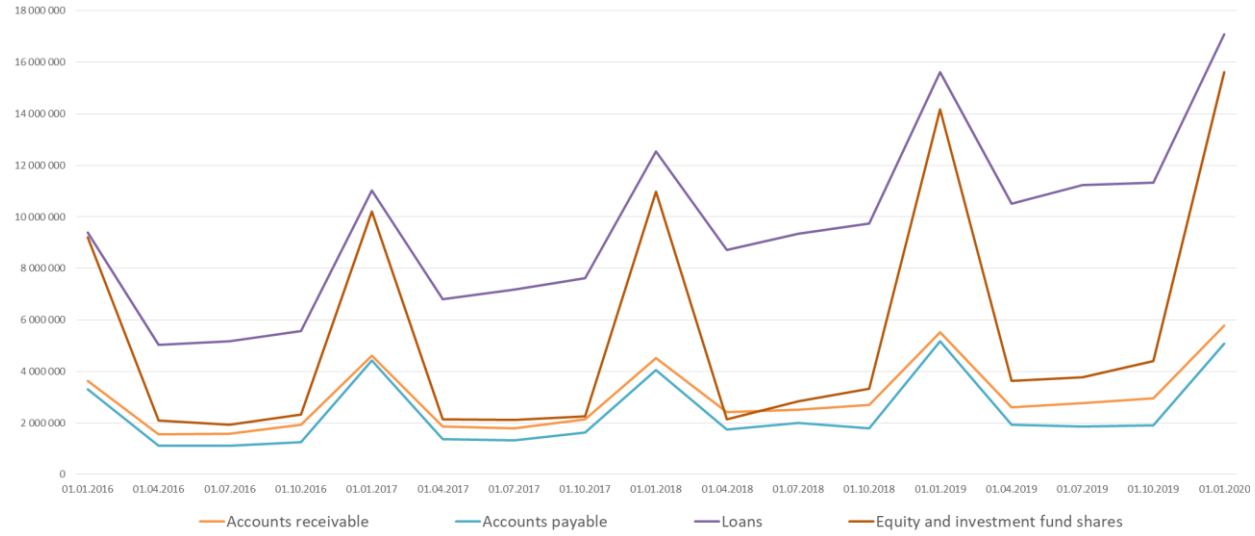

Chart 1. Changes in the main financial measures of accounting statements of OFCs not supervised by the Bank of Russia, 01.01.2016–01.01.2021, mln of rubles

Moreover, the number of OFCs submitting statements as of quarterly dates, including federal statistical forms (forms No. P-3 and No. P-6), is considerably smaller than the number of organisations submitting reporting on an annual basis (see Chart 2). These forms are submitted predominantly by large organisations of the OFC subsector.

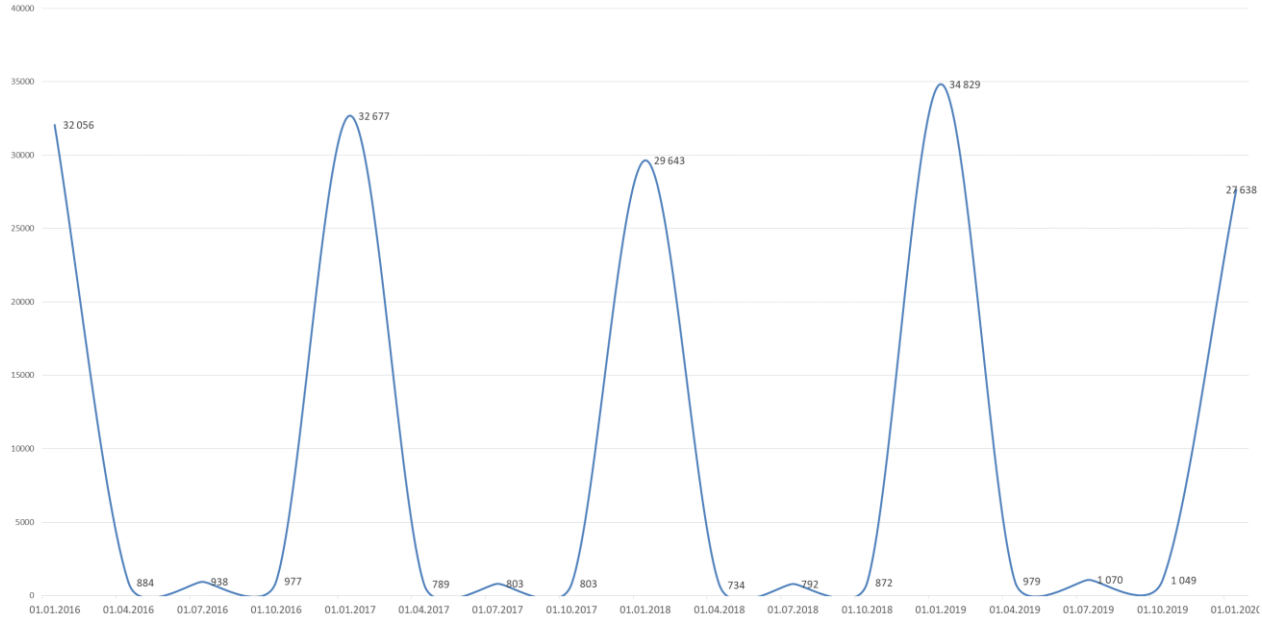

Chart 2. Changes in the number of statements submitted by OFCs not supervised by the Bank of Russia, 01.01.2016–01.01.2020

Chart 2 evidences that OFCs mostly submit statements as of annual dates. According to the analysis of changes in the main measures of accounting statements, the most widespread pattern of gaps over the period from 1 January 2016 to 1 January 2020 is missing data only as of quarterly dates. Specifically, over the period under review, 22.08% of OFCs provided data on Accounts receivable only as of annual dates, 22.04% – on Accounts payable, 20.58% – on Loans, and 27.66% – on Equity.

Chart 3 shows the behaviour of stable organisations of the OFC subsector not supervised by the Bank of Russia, that is, of the organisations that submitted data on the main measures of accounting statements as of all quarterly dates over the period from 1 January 2016 to 1 January 2020. Stable organisations primarily demonstrate bucket dynamics of the main financial measures over the considered period, with declines as of quarterly dates.

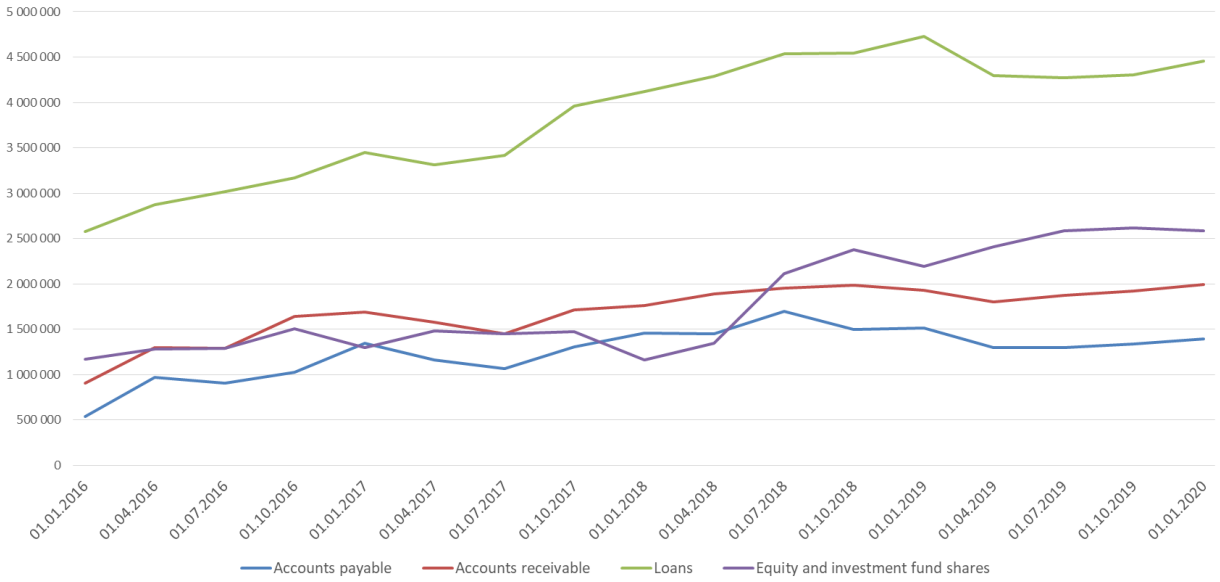

Chart 3. Changes in OFCs' main financial measures, 01.01.2016–01.01.2020, mln of rubles

Results of different methods used to close data gaps

The previous stage of the research made it clear that it is impossible to compile financial balance sheets on a quarterly basis for the subsector of OFCs whose activities are not supervised by the Bank of Russia relying solely on the data submitted to the Bank of Russia. Only a small part of reporting is available on a quarterly basis. Therefore, to obtain comprehensive information on all companies when compiling statistics on financial accounts and sectoral balance sheets, it is necessary to restore missing values in the measures of organisations' accounting statements as of quarterly dates. Below are the main results of various statistical and machine learning methods employed to restore missing data in the measures of accounting statements as of quarterly dates. It should be noted that the analysis encompassed all OFCs, including those subject to the Bank of Russia's supervision, in order to improve the quality of data restoration.

Cluster analysis

As evident from the analysis of descriptive statistics for the main measures of accounting statements, the subsector of OFCs whose activities are not supervised by the Bank of Russia is highly heterogeneous. For this reason, k-means cluster analysis was carried out at the first stage. The scree test formed 11 clusters (see Chart 4).

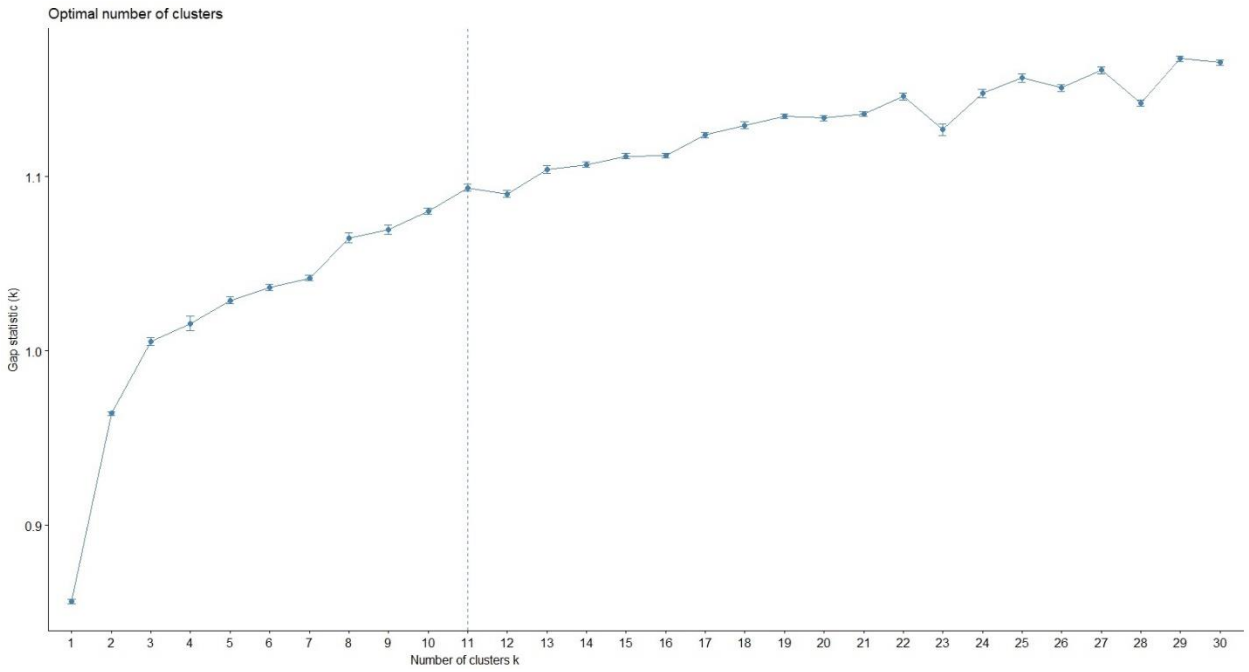

Chart 4. Determining the optimal number of clusters

The variation coefficient for the measures of accounting statements for OFCs not supervised by the Bank of Russia by cluster is presented in Table 4.

Table 4. Variation coefficient for clusters

#	Accounts payable	Accounts receivable	Loans	Equity and investment fund shares
1	6.09	7.39	19.97	7.68
2	17.45	6.65	29.76	7.78
3	15.96	7.13	11.39	5.44
4	19.83	4.95	12.81	12.32
5	17.03	11.97	6.73	17.84
6	5.52	6.54	4.89	4.34
7	10.08	11.54	13.13	4.57
8	6.24	4.83	5.05	9.18
9	6.94	4.04	7.79	6.54
10	21.47	17.25	16.53	8.74
11	8.73	7.46	5.62	6.32

According to Table 4 cluster analysis doesn't solve the problem of heterogeneity, as the value of variation coefficient for each financial instrument in each cluster remain high.

Individual growth rates

As demonstrated by the analysis of missing values in the dynamics of the main measures of accounting statements, it is possible to apply interpolation. This method estimates quarterly values of various financial measures in the form of a fixed weighted linear combination of bounding annual figures.

Let us assume that the value as of the end of the third quarter (y_3) is missing, whereas all other values are known. In this case, the value of the unknown quarterly measure is calculated according to the formula:

$$y_3 = y_2 + \frac{Y_3 - y_2}{2},$$

where y_3 is the value of the measure at the end of the third quarter, Y_3 is the value of the measure as of the end of the year, and y_2 is the value of the measure as of the second quarter. Other cases are considered in a similar way.

The results of interpolation are presented in Chart 5. Changes in the resulting main financial measures resemble the dynamics of measures in accounting statements submitted by stable companies of the OFC subsector shown in Chart 3. The reason for this is that the coverage ratio as of quarterly dates is very low, due to which the result of data restoration is very similar to stable companies' behaviour. However, there is no reason to believe that the behaviour of organisations not submitting statements on a quarterly basis repeats the behaviour of large market participants that submit such reporting.

Chart 5. Results of restoring the main financial measures using the individual growth method, mln of rubles

'1/4' and dynamics extension methods (current method)

Currently, the main method to restore missing values on a quarterly basis is the so-called '1/4' method. It assumes that all quarterly dynamics of balance sheet measures in the OFC subsector are proportionately equal to the annual dynamics of the same financial measures. In particular, the values of unknown quarterly measures are calculated according to the formulas: $y_{1,i} = 0{,}25(Y_{2,i} - Y_{1,i})$, $y_{2,i} = 0{,}5(Y_{2,i} - Y_{1,i})$, $y_{3,i} = 0{,}75(Y_{2,i} - Y_{1,i})$, where $y_{1,i}, y_{2,i}, y_{3,i}$ are unknown values of the measure as of the first,

second, and third quarters, respectively, for the i organisation and $Y_{1,i}, Y_{2,i}$ are known annual values as of the beginning and the end of the year, respectively, for the i organisation. When the value as of the end of the relevant year is unknown, the latest known value of a given measure over the year is taken as the unknown quarterly value.

By attributing a proportionate change over the year to each quarter, it is possible to uniformly distribute the annual growth or decline of a particular financial measure, thus smoothing the dynamics. Furthermore, when the latest known values are attributed as of quarterly dates where there is no closing annual date, the balances of the additionally calculated measure change only based on known data. We thus avoid significant errors in the dynamics when data as of the next annual date are received. The results of current method are presented in Chart 6.

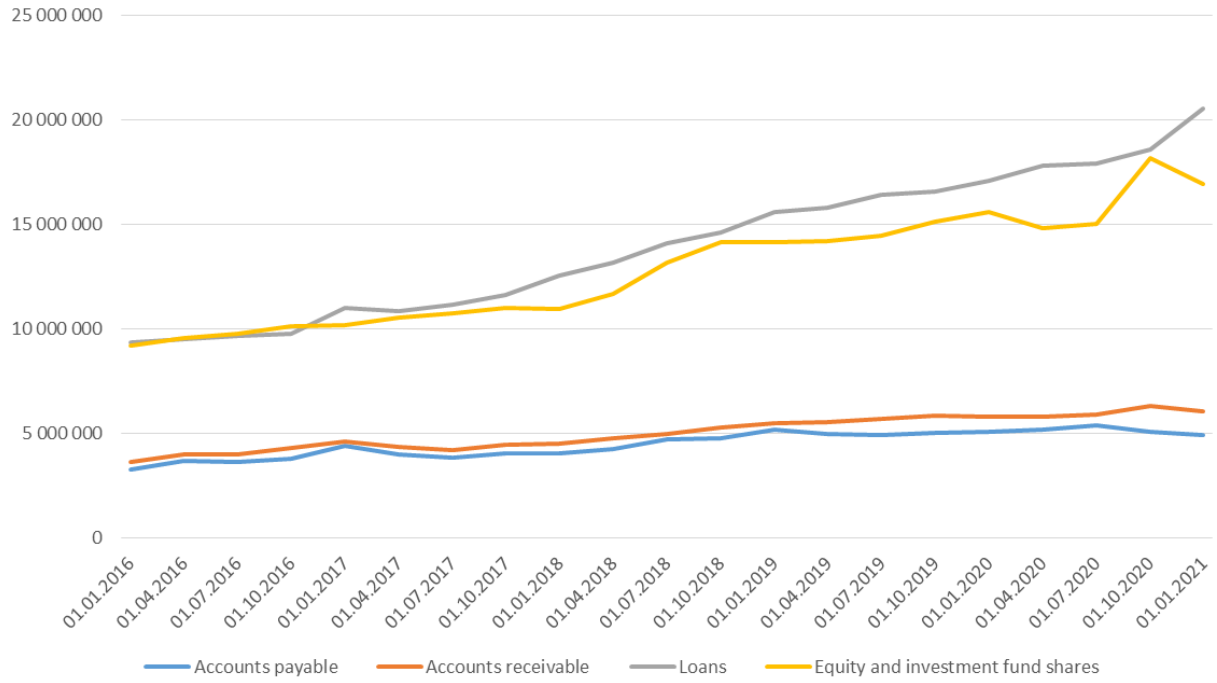

Chart 6. Results of restoring the main financial measures using '1/4' and dynamics extension methods, mln of rubles

Random forest and gradient boosting

As an alternative to classic models we use machine learning algorithms for data recovery such as random forest and gradient boosting. We assume that there is functional relationship between current value of balance indicator and previous values of company's balance indicators. So the relationship for indicator i in period t (y_t^i) looks like this:

$$y_t^i = f_t^i(y_{t-1}^1, .., y_{t-1}^N, y_{t-2}^1, .., y_1^N)$$

Small training sample and large variance in sizes and structures of balances lead to the fact, that fitting this model doesn't give good results. Instead we use following relationship:

$$\frac{y_t^i - y_{t-1}^i}{S_{t-1}} = f_t^i(\frac{y_{t-1}^1}{S_{t-1}}, .., \frac{y_{t-1}^N}{S_{t-1}}, \frac{y_{t-2}^1}{S_{t-1}}, .., \frac{y_{t-2}^N}{S_{t-1}}, S_{t-1})$$

Where S_t – aggregate balance of company. The dependence of the share of the increase in the balance sheet indicator is estimated depending on the distribution of the shares of indicators in the two previous periods. For each indicator, training sample contains companies, with this indicator filled in in the current period, and all indicators filled in past and before last periods. Thus, part of the data recovered by the algorithm before is not included in the learning process.

Random forest and gradient boosting models are used for estimation f_t^i. To tune model hyperparameters (number of trees, maximum tree depth, etc.), the assumption is used that the dependence of the value of any indicator on the values of indicators in past dates has a similar structure to the dependence of the annual value on past annual values. Thus, to select the hyperparameters for each indicator, tests were made on the annual data which is fully completed.

The results of these methods are presented in Chart 7 and Chart 8 respectively.

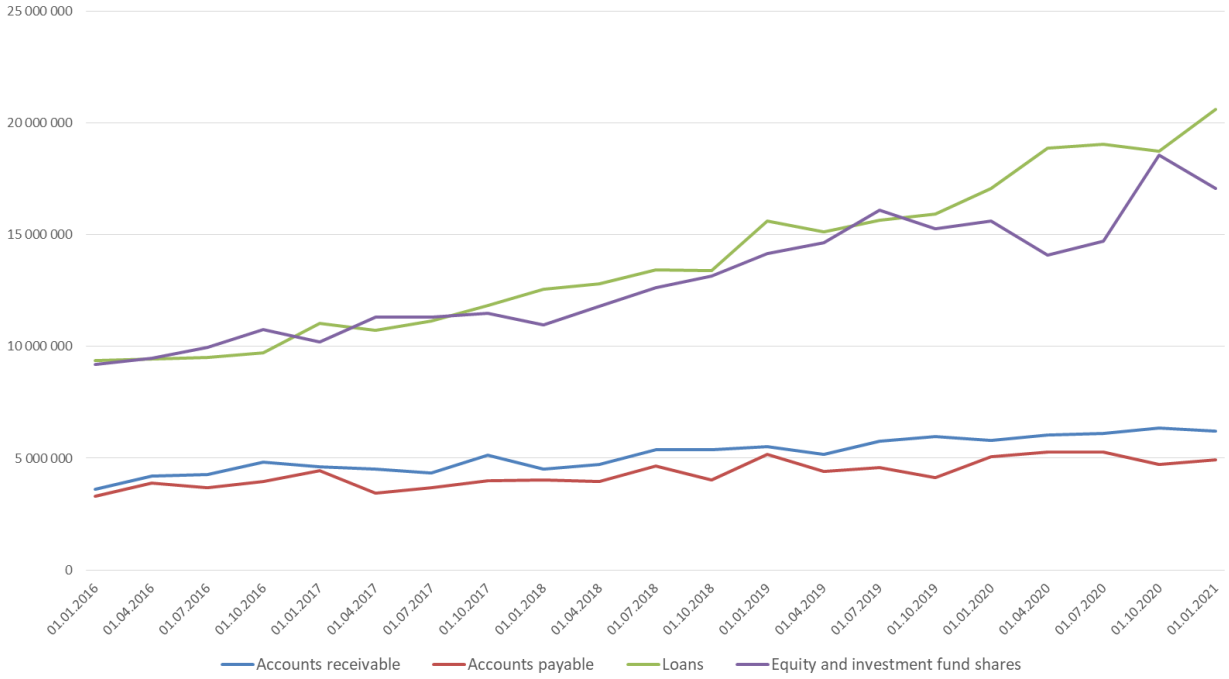

Chart 7. Results of data restoration using random forest, mln of rubles

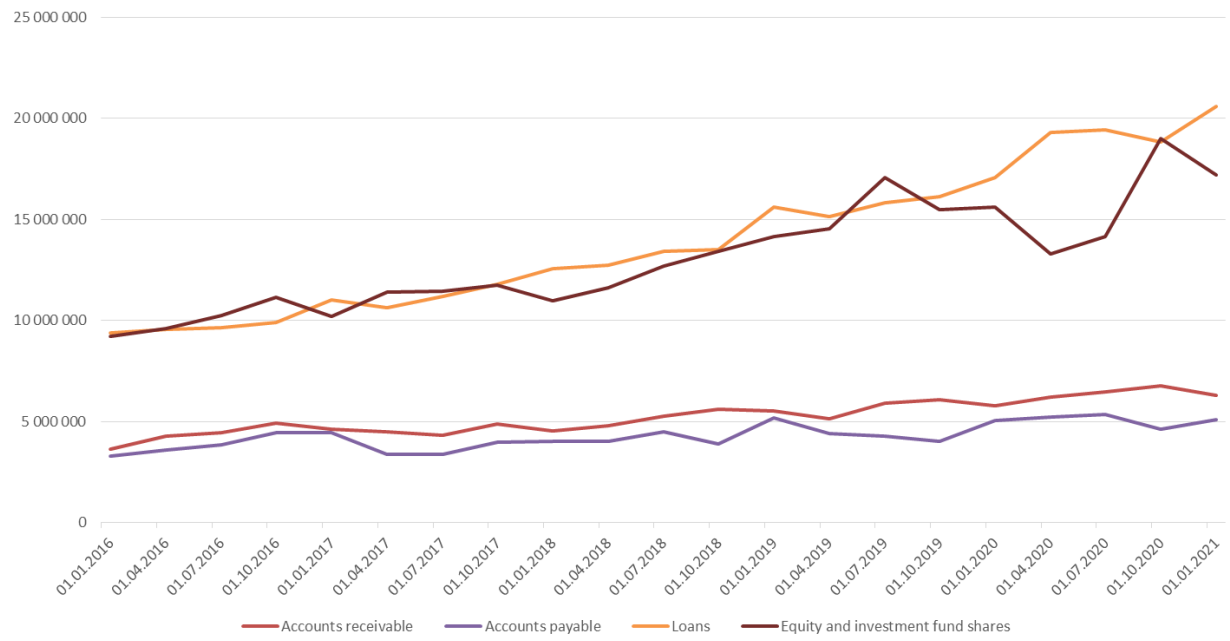

Chart 8. Results of data restoration using Gradient boosting model, mln of rubles

Generative adversarial network

Each measure of financial statements is considered separately. The vector of values for each company is a time series of 17 elements, each of which is a quarterly value of the measure.

The generative network should be trained using the maximum possible number of real companies will all data filled in. However, the array of source data is very sparse as a large portion of companies' quarterly information is unavailable. To increase the number of companies in the training sample, it is usual to also include companies with 'almost all' data filled in. In this case, these are companies having three missing values at most. These missing values are filled in based on mean (quarterly) growth rates calculated for all companies (except the outliers where a specific company's quarterly growth rates are beyond the range [0.7, 1.5]). This generates a training sample that is significantly larger than the original sample of all filled-in data, which uses 'almost all' filled-in data with the minimal impact of growth rates.

After the GAN is trained, the noise generated by it as inputs transforms into the vector of financial measures of a simulated company. To achieve a high accuracy, the input noise is changed iteratively using the Adam optimiser so as to make inputs closer to the filled-in values of the vector. The generative network thus simulates a company which is most similar to the one having data gaps. The generative network outputs are used to close data gaps.

Missing data for all companies are restored sequentially for each financial measure. The results of this method are presented in Chart 9.

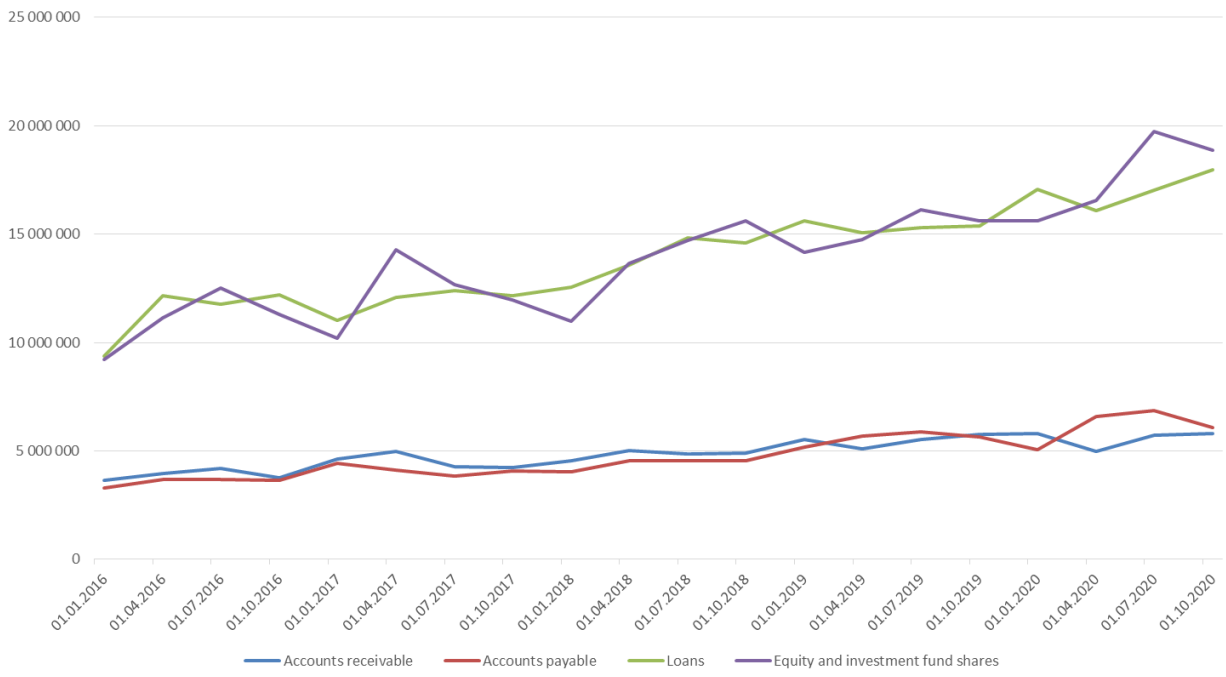

Chart 9. Results of data restoration by the GAN, mln of rubles

Comparative analysis of the methods used to restore quarterly values of the measures of accounting statements and conclusions

At the previous stage of the research, we presented the main results of the data restoration methods employed. Below are the results of the comparison between the currently applied '1/4' and the dynamics extension method, as well as machine learning methods, namely the Random Forest algorithm for regression, gradient boosting model and the generative adversarial network (GAN).

To compare the results of the methods used to close data gaps for each of the reviewed measures as of 1 January 2017 and 1 January 2020, we formed a sample of organisations from the ensemble of other financial corporations carrying out unlicensed activities and made an additional calculation for this sample. The sample for each financial measure included organisations that had not submitted their statements according to federal statistical forms No. P-3 and No. P-6 as of the dates under review.

To compare the methods used, we calculated the ratio of the deviation of the restored values from the actual ones to the total value of each financial measure. The results of closing data gaps in the main financial measures are given in Table 5.

Table 5. Comparison of the results of different methods used to close data gaps, as of 1 January 2017 and 1 January 2020, %

		01.01.2017	01.01.2018	01.01.2019	01.01.2020
Current method	Accounts receivable	-7.13	6.80	-8.16	-2.90
	Accounts payable	-11.01	8.48	-13.13	-9.84
	Loans	-3.44	2.07	-10.50	-4.43
	Equity and investment fund shares	-9.93	-10.33	-16.05	-3.88
Random forest	Accounts receivable	-2.73	11.23	1.96	0.08
	Accounts payable	-4.88	13.26	-1.81	5.25
	Loans	-3.95	7.27	-8.14	-0.83
	Equity and investment fund shares	-0.06	-6.45	-13.33	-8.11
Gradient boosting model	Accounts receivable	3.46	10.30	7.77	2.74
	Accounts payable	-6.63	6.65	2.88	23.44
	Loans	-4.23	7.98	-8.11	2.44
	Equity and investment fund shares	0.19	-5.14	-12.13	-8.25
GAN	Accounts receivable	-18.32	-1.18	-5.19	11.03
	Accounts payable	-14.27	-1.51	2.24	3.76
	Loans	12.90	12.25	-8.97	-4.59
	Equity and investment fund shares	8.94	39.72	-25.13	-8.00

Source: the authors' calculations.

As demonstrated by the analysis of the results presented in Table 5, we can't distinguish the best approach to close data gaps. According to the Table 5 random forest was the best algorithm in many cases. Contrastingly, generative adversarial network method showed the highest ratio of the deviation. This is so, because, first of all, there are not enough points to fit on the GAN model. Secondly, the learning set is more likely to have missed many maxima and minima of the loss function (that finds the weights of neural network), i.e. points where the gradient is zero. This is occured when the domain of inputs is not dense, i.e. the input values are not closely clustered (typical case for sparse data).

The main disadvantage of machine learning methods is their bad interpretive properties, whereas the current method identify organisations accounting for a rise or a decline in a particular measure. Unfortunately, it is hard to tell which method will be preferable for a particular indicator at a certain point in time. Furthermore, there is a significant difference in the results of the calculations of each measure: on average, the best results in data restoration can be achieved in Accounts receivable, whereas the worst results – in Equity. This may be associated with both the low coverage as of quarterly dates, as compared to annual dates, and high volatility of this indicator.

Literature

1. Board of Governors of the Federal Reserve System, 2000.
2. Leo Breiman. Random Forests, 2001
3. Chow, G. and Lin, A. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. The Review of Economics and Statistics 53, 1971, p. 372–375.
4. Fernández, R. A Methodological Note on the Estimation of Time Series. The Review of Economics and Statistics 63, 1981, p. 471–478.
5. Financial Accounts: History, Methods, the Case of Italy and International Comparisons, 2008, p. 118–122.
6. Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001, p. 1189–1232
7. Litterman, R. A Random Walk, Markov Model for the Distribution of Time Series. Journal of Business and Economic Statistics 1, 1983, p. 169–173.
8. Martin Arjovsky, Soumith Chintal, Léon Bottou. Wasserstein GAN, 2017.
9. System of National Accounts 2008 (European Commission, United Nations, Organisation for Economic Cooperation and Development, International Monetary Fund, World Bank).

Bank of Russia

RESTORATION OF OMISSIONS IN THE QUARTERLY INDICATORS OF FINANCIAL STATEMENTS FOR THE OTHER FINANCIAL INSTITUTIONS IN THE BANK OF RUSSIA

PIRUZA ALIEVA AND ANNA BORISENKO,
STATISTICS DEPARTMENT
PETR MILYUTIN AND DENIS KOSHELEV,
RESEARCH & FORECASTING DEPARTMENT

Workshop on Data Science in Central Banking
19-22 October 2021

Agenda

1. Overview of the Other Financial Intermediaries in the Russian Federation
2. The results of cluster analysis for the Other Financial Intermediaries
3. Current approach of the restoration of omissions in the quarterly indicators of financial statements for the Other Financial Intermediaries
4. Results of methods
    1. Individual growth rates method
    2. Random forest
    3. Gradient boosting model
    4. Generative adversarial networks
5. Comparison of methods
6. Conclusions

Overview of the Other Financial Intermediaries in the Russian Federation

Table 1. Unsupervised OFIs' statistics

	Portion of unsupervised OFIs' total balance in all OFIs' total balance	Ratio of the number of unsupervised OFIs number to the overall number of OFIs
01.01.2015	97,34%	72,75%
01.01.2016	97,48%	76,97%
01.01.2017	96,07%	77,31%
01.01.2018	95,29%	83,45%
01.01.2019	99,28%	84,47%
01.01.2020	99,54%	85,52%

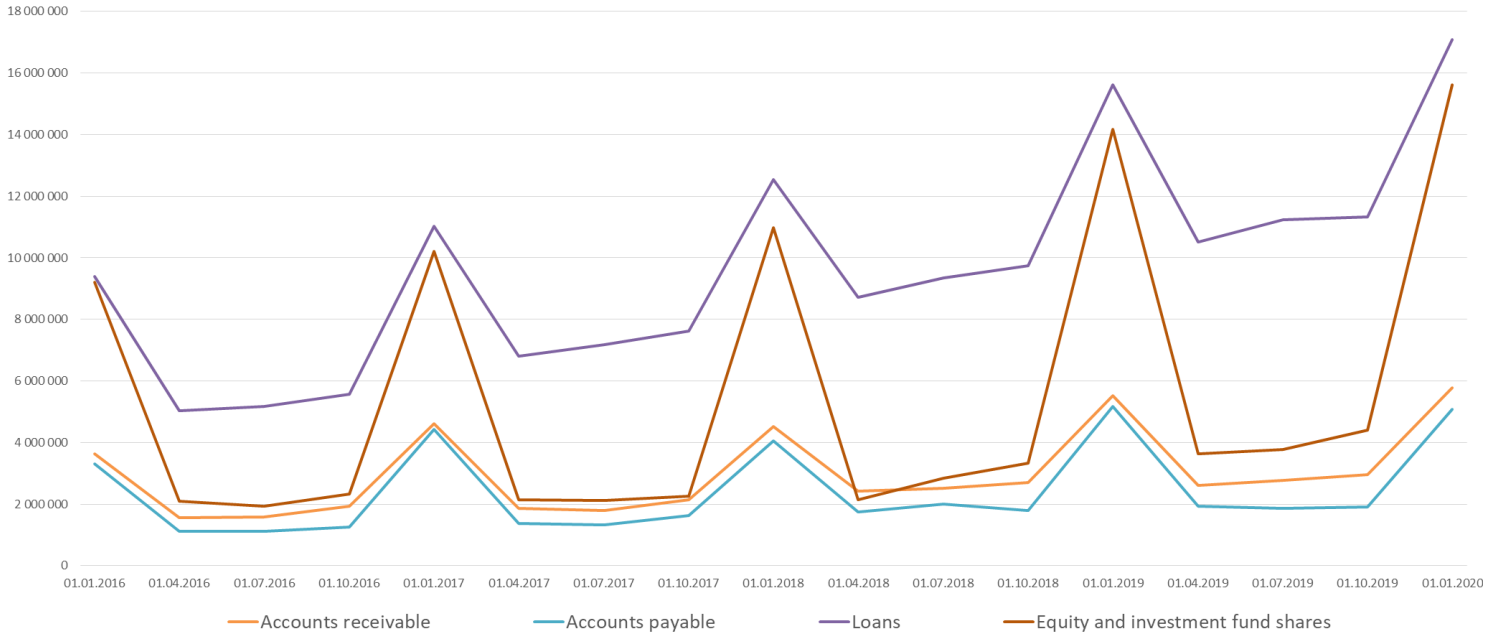

Figure 1. Dynamics of the main financial measures of accounting statements of unsupervised OFIs', 2016 – 2020, mln of rubles

Cluster analysis

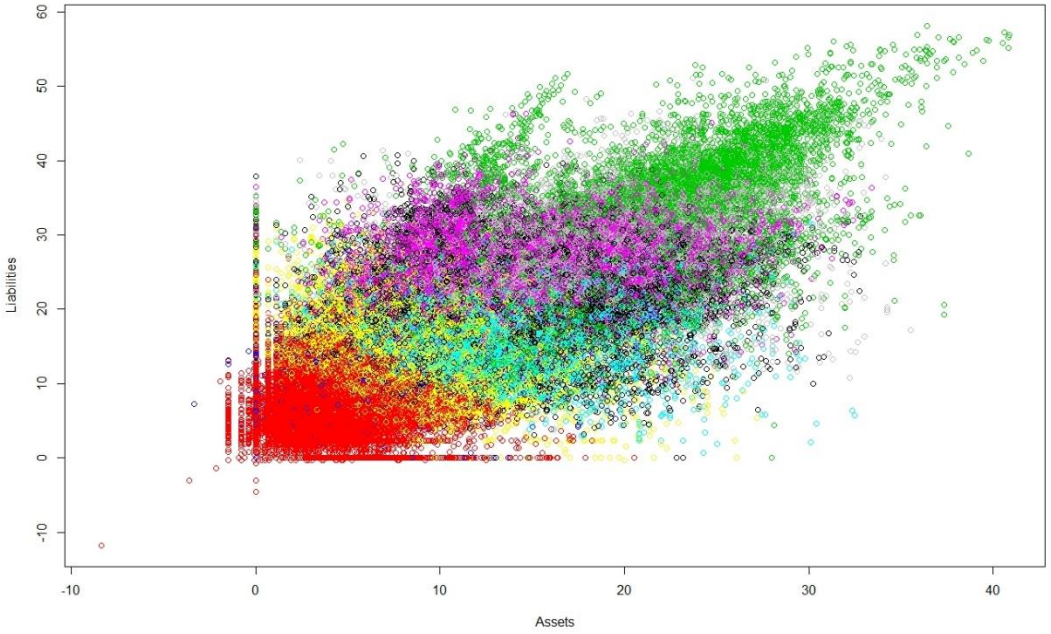

Figure 2. Results of cluster analysis

Table 2. Variation coefficient for clusters

#	Accounts payable	Accounts receivable	Loans	Equity and investment fund shares
1	6,09	7,39	19,97	7,68
2	17,45	6,65	29,76	7,78
3	15,96	7,13	11,39	5,44
4	19,83	4,95	12,81	12,32
5	17,03	11,97	6,73	17,84
6	5,52	6,54	4,89	4,34
7	10,08	11,54	13,13	4,57
8	6,24	4,83	5,05	9,18
9	6,94	4,04	7,79	6,54
10	21,47	17,25	16,53	8,74
11	8,73	7,46	5,62	6,32

Current approach of the restoration of omissions in the quarterly indicators

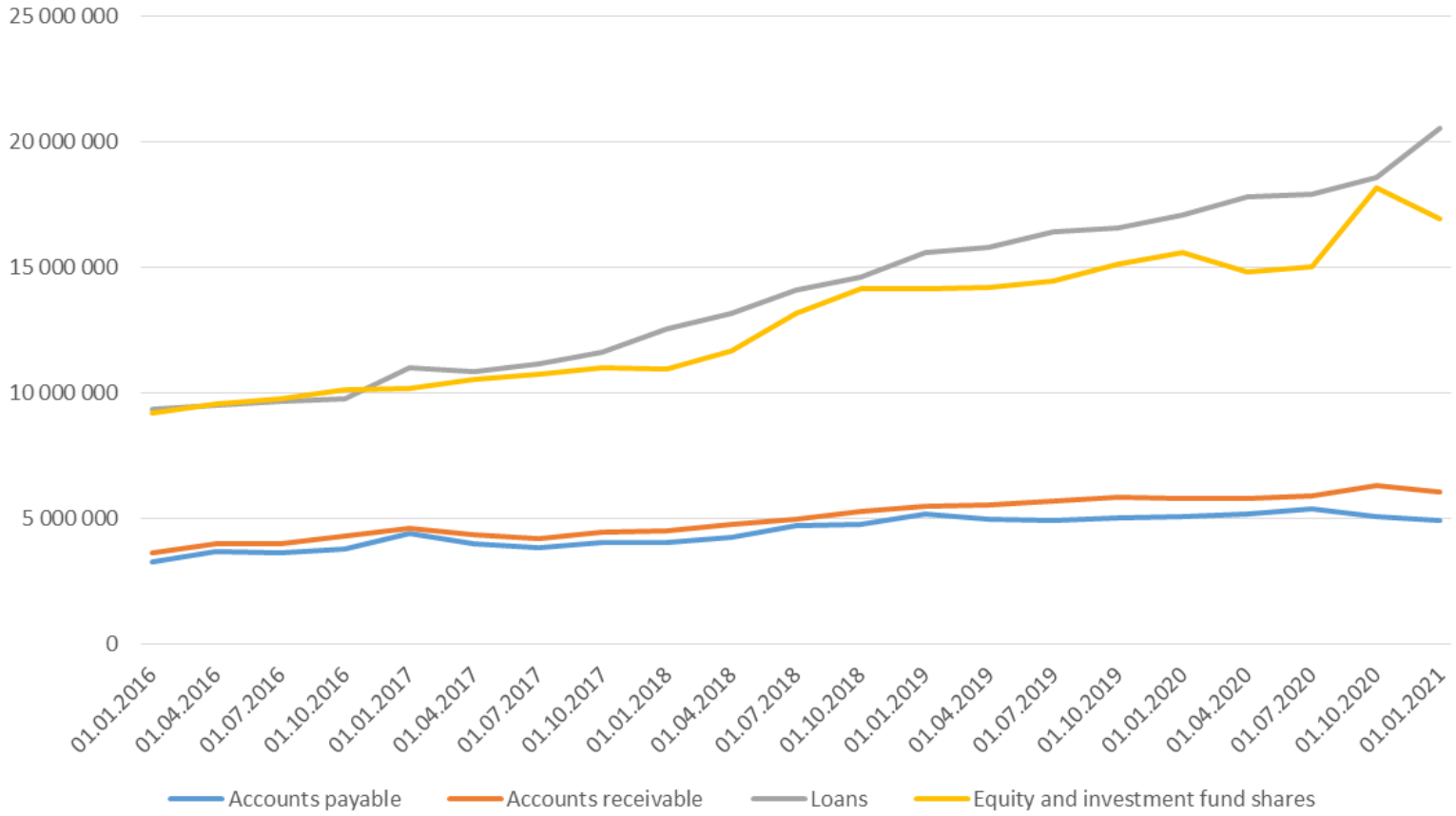

Figure 3. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Individual growth rate method

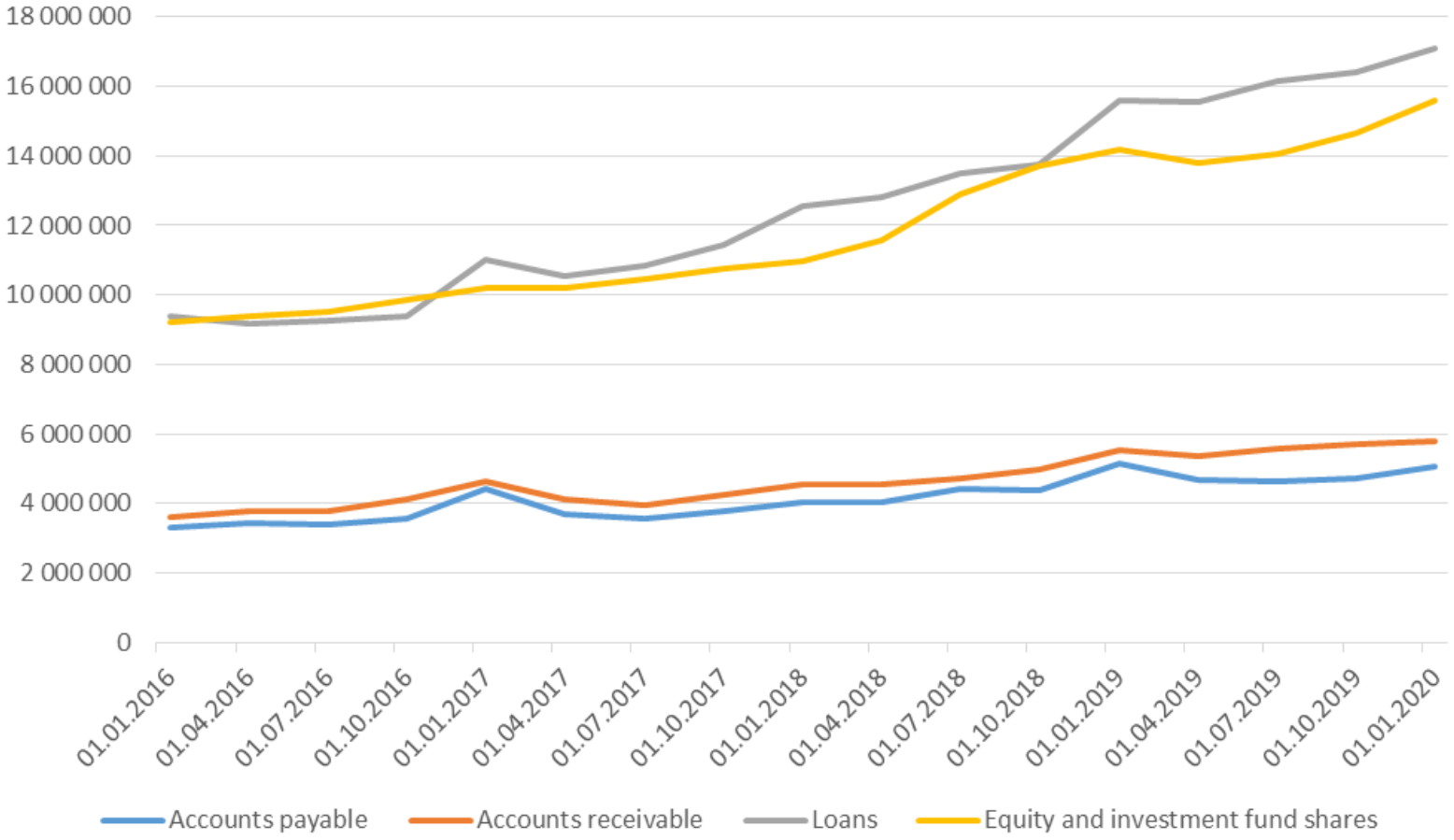

Figure 4. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Stable OFIs' main financial measures

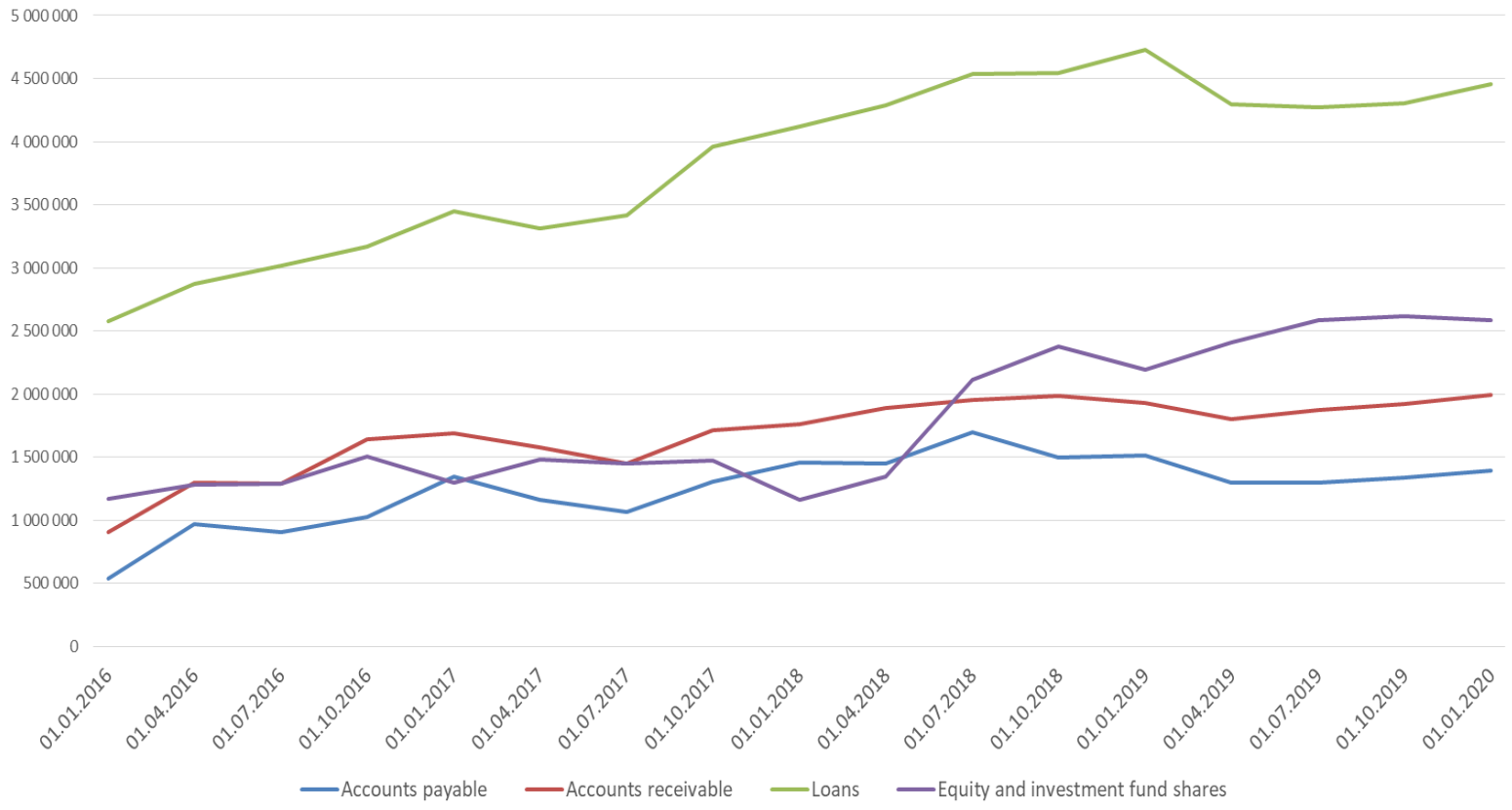

Figure 5. Changes stable OFCs' main financial measures, 01.01.2016–01.01.2020, mln of rubles

Random forest

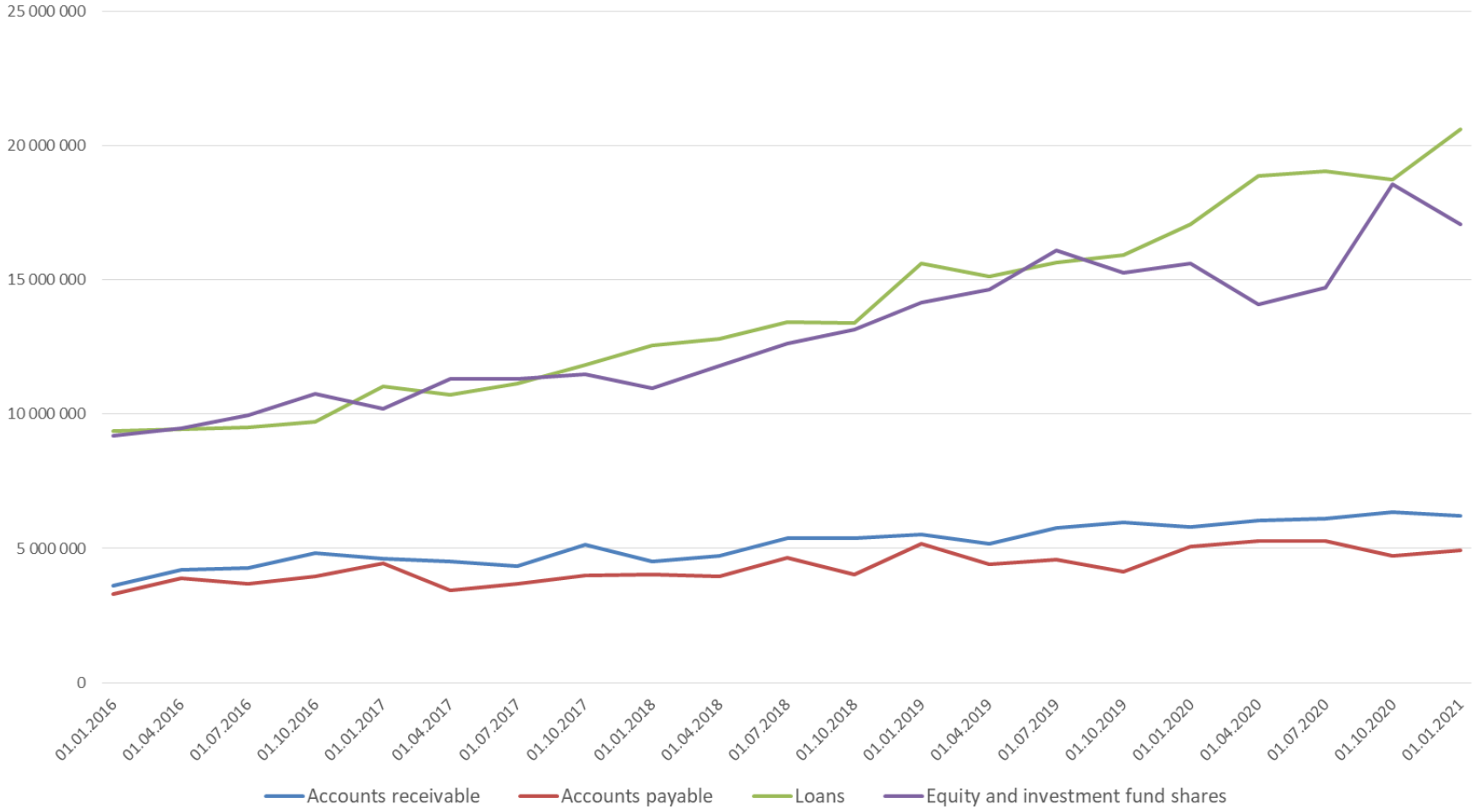

Figure 6. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Gradient boosting model

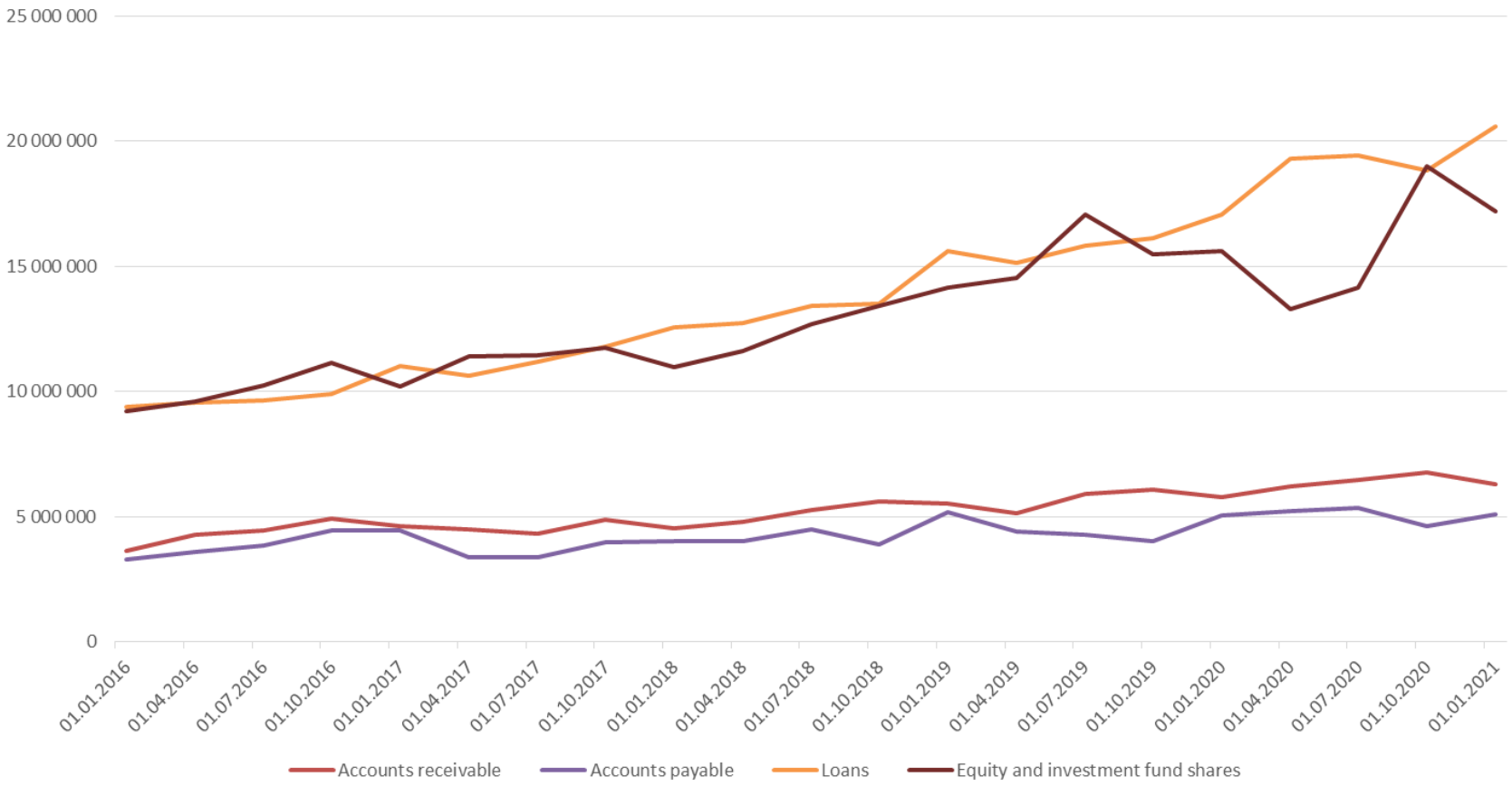

Figure 7. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Generative adversarial networks (GAN)

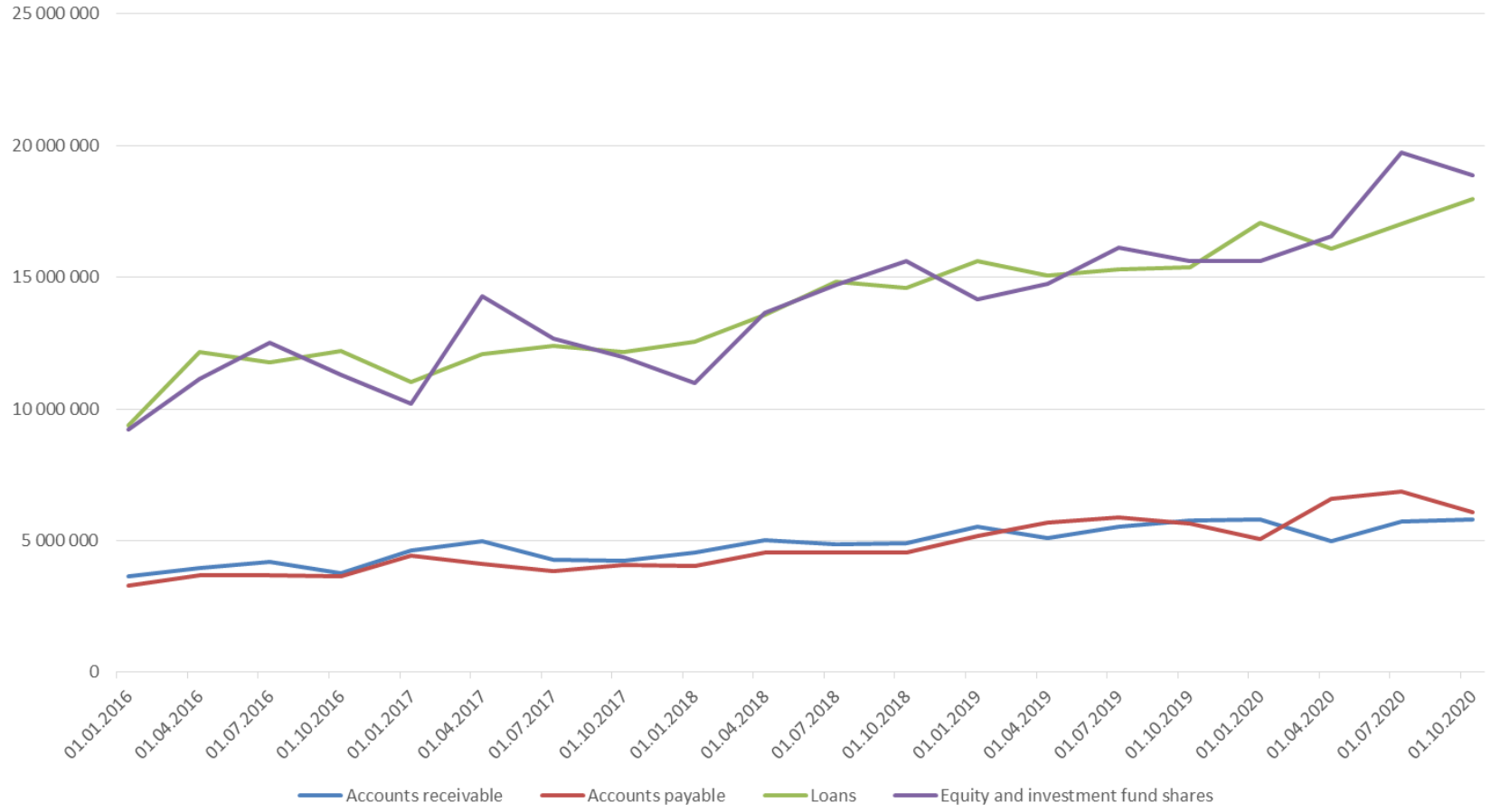

Figure 8. Results of the restoration of omissions in the quarterly indicators (mln of rubles)

Comparison of methods and conclusions

Table 3. Results of comparison of methods (deviation of estimated value from real number)

		01.01.2017	01.01.2018	01.01.2019	01.01.2020
Current method	Accounts receivable	-7,13%	6,80%	-8,16%	-2,90%
	Accounts payable	-11,01%	8,48%	-13,13%	-9,84%
	Loans	**-3,44%**	**2,07%**	-10,50%	-4,43%
	Equity and investment fund shares	-9,93%	-10,33%	-16,05%	**-3,88%**
Random forest	Accounts receivable	**-2,73%**	11,23%	**1,96%**	**0,08%**
	Accounts payable	**-4,88%**	13,26%	**-1,81%**	**5,25%**
	Loans	-3,95%	7,27%	-8,14%	**-0,83%**
	Equity and investment fund shares	**-0,06%**	-6,45%	-13,33%	-8,11%
Gradient boosting model	Accounts receivable	3,46%	10,30%	7,77%	2,74%
	Accounts payable	-6,63%	6,65%	2,88%	23,44%
	Loans	-4,23%	7,98%	**-8,11%**	2,44%
	Equity and investment fund shares	0,19%	**-5,14%**	**-12,13%**	-8,25%
Generative adversarial networks	Accounts receivable	-18,32%	**-1,18%**	-5,19%	11,03%
	Accounts payable	-14,27%	**-1,51%**	2,24%	3,76%
	Loans	12,90%	12,25%	-8,97%	-4,59%
	Equity and investment fund shares	8,94%	39,72%	-25,13%	-8,00%

Bank of Russia

THANK YOU FOR YOUR ATTENTION

RESTORATION OF OMISSIONS IN THE QUARTERLY INDICATORS OF FINANCIAL STATEMENTS FOR THE OTHER FINANCIAL INSTITUTIONS IN THE BANK OF RUSSIA

PIRUZA ALIEVA AND ANNA BORISENKO, STATISTICS DEPARTMENT
PETR MILYUTIN AND DENIS KOSHELEV, RESEARCH & FORECASTING DEPARTMENT

2021