NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Marat Z. Kurbangaleev, Victor A. Lapshin,
Zinaida V. Seleznyova

STUDYING THE REPLICABILITY OF AGGREGATE EXTERNAL CREDIT ASSESSMENTS USING PUBLIC INFORMATION

BASIC RESEARCH PROGRAM
WORKING PAPERS

SERIES: FINANCIAL ECONOMICS
WP BRP 71/FE/2018

Marat Z. Kurbangaleev[1], *Victor A. Lapshin*[2],

Zinaida V. Seleznyova[3]

STUDYING THE REPLICABILITY OF AGGREGATE EXTERNAL CREDIT ASSESSMENTS USING PUBLIC INFORMATION[4]

In this paper, we examine whether an aggregate rating can be accurately predicted with publicly available information about a company's individual characteristics. We propose an algorithm that shows how efficient and replicable an arbitrary aggregate rating is respectively to the widely used credit risk models and to what extent an aggregate rating can be extrapolated to the non-rated companies as a valid indicator of their credit risk. Using this algorithm, we empirically study the aggregate ratings constructed as a consensus of ratings assigned by seven credit rating agencies for Russian banks on a national scale and compare it with several alternatives and proxies based on the publicly available characteristics of those banks. We measure how well the aggregate (consensus) rating and the proxies are agreed in terms of ordering banks by their credit quality and predicting defaults over a one-year horizon. We show that the aggregate (consensus) rating is comparable to a standard logit default model in terms of discriminatory power, but for ordering, the former is in low agreement with the latter. We also found that using models for predicting initial credit ratings allows the building of a proxy that is in high agreement with the original aggregate rating, but the original aggregate rating outperforms the proxy in terms of discriminatory power. It was also found that greater agreement between the original aggregated rating and the proxy can be achieved on a subsample of investment grade ratings.

JEL Classification: B23, G21, G24

Keywords: credit rating agency, credit ratings, rating aggregation, consensus ordering, logit model

1 National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab. E-mail: mkurbangaleev@hse.ru

2 National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab. E-mail: vlapshin@hse.ru

3 National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab. E-mail: zseleznyova@hse.ru

4 Study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2017.

1 Introduction

Credit rating agencies (CRA) play a significant role in the modern financial market, presenting professional opinions on the financial stability or creditworthiness of companies or other legal entities. As opinions, CRA assessments do not always match due to conceptual differences (such as in rating philosophies[5], the factors analyzed, and the models applied) or because of occasional calculation errors, delayed reactions or intentional misrepresentations.

Disagreements between CRA assessments can be significant. For example, the same entity is rated by S&P as a BBB and Fitch as a BB, i.e. is the assigned ratings implying significantly different credit quality. The first CRA states that the entity "has adequate capacity to meet its financial commitments" (Standard&Poors, (2016), p.6), whereas second one says that the entity "indicates an elevated vulnerability to default risk" (Fitch Ratings, (2014), p.9). Such an ordering can be reversed by different agencies with respect to that entity's peers, therefore such CRA opinions are inconsistent and cannot be seen as a reliable indicator of the relative risk of a particular entity.

The natural question arises. Is it possible to construct a collective indicator of credit quality from two or more independent (maybe partially contradictory) opinions and what properties should this indicator have to be meaningful, robust and useful? Such an indicator is an aggregation. In the context of credit ratings, the result of such an aggregation is an **aggregate rating**.

There are three main reasons for studying aggregate credit ratings and the quality of aggregation.

1. **Reducing the reliance on a single external credit rating**. In 2010, responding to the lessons learnt from the global financial crisis 2007–8, the G20 Financial Stability Board issued a resolution encouraging reduced reliance (especially mechanistic) on CRA ratings by banks in order to prevent the negative consequences of rating inaccuracy and the "cliff effect". However, in emerging markets such as Russia, CRA ratings play a key role in providing information on company creditworthiness. Therefore, regulators and market participants cannot abandon CRA ratings, at least for now.

   An aggregate rating would reduce the dependency of the financial system on any single CRA in the spirit proposed by the Financial Stability Board resolution. It can also be used by the financial regulator as a reference scale in order to provide a mapping of external rating scales

[5] Point-in-time (PIT), through-the-cycle (TTC) or hybrid.

to the regulatory scale thus creating a fair, level playing-field, and most importantly a robust framework.

2. **Meaningful aggregation.** Banks use internal rating-based models to assess the creditworthiness of their counterparties. These models routinely deal with heterogeneous information such as external credit ratings, credit spreads, internal models, expert estimates. Aggregating this heterogeneous information into one internal rating is exactly what an aggregate credit rating does with CRA ratings. A meaningful internal aggregation model is very important for regulatory compliance, especially in a low-default environment, where models cannot be validated using statistical data on defaults.
3. **More data of better quality for training other models**. In a default-rich environment, models for estimating credit quality are usually trained on samples of companies which default and those which do not (with a suitable definition of default, usually including a time horizon). However, in low-default environments the set of companies which have defaulted is almost empty, which makes training problematic. A common solution to this problem is to train the model not to identify defaults from non-defaults, but to replicate an external credit quality proxy such as external credit ratings (BCBS (2005), p.96-102). However, the question of which external credit rating to choose immediately arises. An aggregate credit rating is a natural candidate for such a credit quality proxy. It is preferable to a single external credit rating, because:
    a. It is more robust to any outliers in the rating data. Quality input data allow the training of credit quality models with more precision.
    b. It encompasses more companies than any single external credit rating. More input data also increase training quality and possibilities. For some kinds of models, more data make all the difference.
    c. It eliminates the variability of model estimates. Training the same model to replicate different external credit ratings results in different models, because the sets of rated companies differ for different external credit ratings, and also because some companies are rated differently by different CRA.

However, points 2, 3a, 3b and 3c above explicitly depend on the availability of independent studies of such aggregation methodologies.

This paper, while not intending to close this question, contributes to the topic by proposing a methodology for studying the quality and applicability limits of an arbitrary aggregate rating relative

to existing credit risk models in order to determine if such an aggregate rating can be seen as a suitable tool for the issues described in point 3. The methodology is based on a two-step algorithm which helps to determine, 1) if the aggregate rating performs as well as best-practice credit risk models on its own domain of rated companies, 2) if the aggregate rating can be accurately modeled and, therefore, extrapolated to non-rated companies without loss of competitiveness. We argue that an aggregate rating which successfully passes these two steps is suitable as a candidate for point 3.

Various approaches to rating aggregation can be found in the academic and practical literature. "Ad hoc" methods[6] are usually intuitive and easy to implement, but they typically lack for conceptual soundness. More comprehensive approaches can be roughly divided into two classes: model (generally parametric) and model-independent (generally normative and non-parametric) approaches. Model approaches are understandable and tractable, but are typically more complex than "ad hoc" methods and depend heavily on assumptions about the information they aggregate (e.g. ratings are assigned by rating agencies independently of each other). Elements of the model-based aggregation of ratings can be found in Karminsky, Peresetsky (2008), Aivasyan et al. (2011), Hornik et al. (2010), Karminsky et al. (2013), Grun et al. (2013) to name a few. The implementation of model-independent approaches is primarily based on the interpretation of data and the desirable properties of the aggregate rating. Examples of such approaches are Eisl et al. (2013), Buzdalin et al. (2017). In particular, Buzdalin et al. (2017) adopts the concept of consensus from social choice theory as a basic principle of rating aggregation.

In this paper, we apply the proposed methodology to studying the aggregate rating constructed in Buzdalin et al. (2017) as a consensus of individual credit ratings assigned to Russian banks by three international rating agencies (Fitch, Moody's and Standard&Poors) and four Russian national rating agencies (AKM, NRA, RA Expert and RUS Rating) from the third quarter of 2010 to the first quarter of 2016. We use one particular method of aggregation, therefore, we use the terms **aggregate rating** and **consensus rating** interchangeably. The original paper by Buzdalin et al. (2017) shows that consensus rating demonstrates good discriminatory power and robustness, the method of its construction is computationally hard, so it is important to make sure that the aggregate rating provides information that is worth its complexity.

The paper has the following structure. Section 2 describes our methodology. Section 3 briefly describes the data and Section 4 presents and discusses the results. Section 5 concludes.

[6] E.g. Russian information service company Interfax calculate and disseminates such aggregates.

2 Methodology

The proposed methodology is a two-step algorithm.

1. The first step is to determine if an aggregate rating perform better or comparative to widely used credit risk models for the rated companies. Relatively poor performance would make the aggregate rating worthless and not deserving of further study.

2. If the aggregate rating performs relatively well in its own domain, the replicability of the aggregate rating is examined at the second step of the algorithm. We call the rating replicable if (1) an accurate predictive model of such a rating can be built, (2) the model can be used to extrapolate ratings to non-rated companies, (3) the performance of the extrapolated ratings is still better or comparable to the performance of the best-practice credit risk models.

Assuming the robustness of the aggregation method itself, successfully passing these two steps means that the considered aggregate rating is suitable, since it is determined for all companies, robust and performs at least as well as best-practice models. A successful pass of only the first step means that the aggregate rating has limited applicability outside its own domain, but still can be used for credit risk analysis, for example, for validation purposes.

In this paper the first step is carried out in the following order.

1. We construct the aggregated rating (see Subsection 2.1 Consensus-based aggregation of ratings), which is defined for the companies assigned two or more ratings from different CRA (further below we refer to such a data set as a Consensus sample, see Section 3 Data).
2. We build a logit default model that utilizes the individual (financial, business) characteristics of the company and is calibrated to the default data (see Subsection 2.2 Econometric default model). This model is defined for all companies in our data. The model is estimated and tested on the Training and Test Samples respectively (see Section 3 Data).
3. We then compare the discriminatory powers of the aggregate rating and the logit model to determine if the aggregate rating provides as much information on the credit quality of Russian banks as a standard, purely default-based, econometric model. We measure the discriminatory power with the Accuracy Ration (AR) indicator (see Subsection 2.4 Discriminatory power).

The second step builds a series of predictive models for the consensus rating to study if it can

be replicated and extrapolated outside its domain.

1. First, we map the level of the consensus rating directly to the individual (financial, business) characteristics of the companies via an ordered logit model (see Subsection 2.3 Rating models). We compare the predicted consensus with a real one in terms of the discriminatory power and degree of agreement (see Subsection 2.5 Degree of agreement) in order to determine if the model fits the real aggregate rating well and has comparable quality inside and outside its domain.
2. Second, we build CRA ratings as inputs for the aggregation via an ordered logit model (see Subsection 2.3 Rating models). Then we compare the consensus of the modeled ratings with the real one in terms of the discriminatory power and degree of agreement, as for the predictive model of consensus in the previous point. In order to ensure that the result is robust and is not subject to heterogeneous data (different rating methodologies, different rating class), we carry out this exercise for different combinations of CRA and ratings:
    a. all ratings of all seven CRA;
    b. all ratings of Russian national rating agencies;
    c. investment grade ratings of all seven CRA.

2.1 Consensus-based aggregation of ratings

The approach to aggregation that we study in this paper considers credit ratings as relative orders of entities according to CRA opinions about their relative credit quality. Such a rating interpretation allows the application of some widely used concepts from social choice theory to the rating aggregation problem. The paper adopts the Kemeny median concept which formalizes a fair (consensus) aggregation of orders and from a practical perspective has some natural properties (see Brandt et al. (2016)). Kemeny median is a solution of the following problem

$$R^* = arg \min_R \sum_{k=1}^{m} d(R, R_k), \quad (1)$$

where R^* is the Kemeny median, m is the number of input orders, R_k is the k-th individual (input) orders, $d(R', R'')$ is the Kemeny-Snell distance metric between orders R' and R''.

Although the Kemeny median concept is relatively well studied and developed, its application to the ratings aggregation problem has some specific features, such as high dimensionality and a partial input order, i.e. orders may not be defined for all objects (not every CRA rates each entity). Moreover, the original Kemeny median generally provides a set of aggregations rather than a unique solution. Together these features make the rating aggregation problem computationally complex.

In order to obtain a single solution within a practically acceptable time, the original optimization problem is modified by adding supplementary criterion and setting it in the spirit of the Tikhonov regularization. A genetic optimization algorithm is adopted for the numerical solution. Therefore, a consensus rating is:

$$R^{cons} = arg \min_{R} \sum_{k=1}^{m} \phi_k [\tilde{d}(R, R_k) + \lambda \delta^2(R, R_k)], \quad (2)$$

where R^{cons} is the aggregate (consensus) rating, m is the number of input orders, R_k is the individual (partial) order of entities according to the ratings assigned by the k-th CRA; $\varphi_k > 0$, $\sum_{k=1}^{m} \varphi_k = 1$ is the weight representing relative CRA credibility (if all agencies are equally credible, then $\varphi_k = 1/m$, for all k); $\tilde{d}(R', R'')$ is the modified Kemeny-Snell distance metric; $\delta^2(R', R'')$ is a supplementary criterion. Having λ is small, the consensus rating is still optimal according to criterion $\tilde{d}(R', R'')$, but also the best one according to supplementary criterion $\delta^2(R', R'')$.

Such an approach, applied to real rating data, provides an aggregate rating with good discriminatory power, therefore it can be considered a fair and robust benchmark in a multi-rating environment. For more details on method and its properties see Buzdalin et al. (2017).

2.2 Econometric default model

Econometric models (such as logit or probit models) are widely used to build up a multi-variable scoring/rating system calibrated to default data. These models are fairly simple, easy to implement and recognized by the Basel Committee (see BCBS (2005), p.33, p.37). They are also frequently used for research purposes, such as Campbell et al. (2008), Agarwal, Taffler (2008), Kavussanos, Tsouknidis (2016). These papers use data from financial statements and qualitative (typically categorical) indicators as input for those models in order to assess the default probability of entities from financial and non-financial sectors.

In these models the default event of *i*-th entity is modeled by binary variable Y_i depending on variable Y_i^* which represents entity's credit quality:

$$Y_i = \begin{cases} 1, \text{if } Y_i^* \geq 0 \quad \text{(default)} \\ 0, \text{else} \quad \text{(no default)} \end{cases}. \tag{3}$$

If Y_i^* linearly depends on some observable variables X (entity characteristics, macroeconomic factors etc.) and some unobservable random component ε with distribution F, the probability of default can be written as follows

$$P(Y_i = 1) = P(Y_i^* \geq 0) = P(X_i'\beta + \varepsilon \geq 0); \tag{4}$$

$$P(Y_i = 1) = 1 - F(X_i'\beta). \tag{5}$$

where F(*z*) is usually chosen to be a logistic cumulative distribution function and the model is called logit.[7]

In this paper, an order of entities according to a logit regression is considered an independent alternative to the consensus rating and used for benchmarking its discriminatory power.

2.3 Rating models

If the consensus rating and its independent (default based) alternative show low agreement, a proxy of the consensus rating can be constructed in a way close in spirit to the original consensus rating. As the original consensus rating consists of two components – the data (ratings) and the method (algorithm) of aggregation – it is reasonable to ask if a close proxy can be obtained by altering these components. In particular, can a close proxy be constructed from non-rating data, for example, by predicted ratings?

One of the generally accepted tools for assessing and predicting ratings is the econometric models of ordered choice, for example, an ordered logit model. The credit rating of the *i*-th entity assigned by a particular CRA is modeled by variable y_i depending on the variable.

$$y_i = \begin{cases} 0, \text{if } y'_i \leq c_0 \\ 1, \text{ if } c_0 < y'_i \leq c_1 \\ \quad \ldots \\ n, \text{ if } y'_i > c_{n-1} \end{cases}, \tag{6}$$

[7] A popular alternative to logit model is probit model, which applies normal distribution instead of logistic. Usually logit and probit models provide fairly close results.

where y'_i represents an entity's credit quality, c_j are the endpoints of the observable rating categories in terms of y'_i values, *n* is number of observed rating categories

If y'_i linearly depends on X (some entity's observable characteristic or macroeconomic factor), the probability of falling into some rating category can be written as:

$$\begin{aligned} P(y'_i = 0) &= F(c_0 - X'_i\beta), \\ P(y'_i = 1) &= F(c_0 - X'_i\beta) - F(c_1 - X'_i\beta), \\ P(y'_i = n) &= 1 - F(c_{n-1} - X'_i\beta). \end{aligned} \tag{7}$$

One option to measure the goodness of fit of an ordered selection model is to measure MacFadden's R^2 (Likelihood Ratio Index, LRI) which is the following:

$$LRI = 1 - \frac{l_1}{l_0}, \tag{8}$$

where l_1 is the log-likelihood function value for the estimated regression, l_0 is the log-likelihood function value if all coefficients except the "constant" are assumed to insignificant. As can be seen from the formula, it is almost a direct analogue of OLS R^2, and its meaning is the same. The larger the LRI, the more accurately the model predicts ratings. However, it is argued that an integral goodness-of-fit measure should be considered along with a more detailed indicator in order to provide a more granular representation of fitting results (see, for example, Hosmer et al. (2013)). In this regard, we use a classification table and present the consolidated results of the accuracy of the predictions by models to refine the results.

2.4 Discriminatory power

For the purpose of this study we need to measure the discriminatory power of the scoring variables. Since a consensus rating in essence is a scoring variable and the discriminatory power is a generally accepted indicator of rating quality, it is natural to ask if the discriminatory power of the consensus rating is comparable to the discriminatory power of popular scoring default models calibrated to default data.

Generally, discriminatory power is represented by the ROC (CAP)-curve[8] and measured by AUC (area under curve) and/or AR[9] (accuracy ratio) (see Tasche (2010)). AUC ranges from 0 to 1, and the greater the value, the greater the discriminatory power of the model. For ROC-curve AR = 2

[8] Receiver Operating Characteristic and Cumulative Accuracy Plot respectively. By their nature and function these two plots are close to Lorenz curve.
[9] In essence AR is a Gini coefficient.

AUC-1 and ranges from 0 to 1. The quality for the scoring/rating model is also measured in terms of AR: the larger the AR, the better the model predicts defaults (see Pomazanov (2016), p.54 or Hosmer et al. (2013) p. 177). All these indicators have also been recommended by the Basel Committee (see BCBS (2005), p.36-39) and have been used repeatedly in research.

2.5 Degree of agreement

For the purpose of this study we need to measure the agreement between ratings (orders). We do this using the modified Kendall correlation coefficient of τ_x (see Emond, Mason (2002)). There are a few reasons for such a choice. First, as Emond and Mason shown, τ_x is the unique rank correlation coefficient, which is equivalent to the Kemeny-Snell distance metric, which is the key component of consensus rating construction. Second, like any other concordance coefficient, it represents the degree of agreement between two orders, while it does not lend itself to an accurate quantitative estimate. Unlike the other concordance coefficients, the same dimension assessment scales are unnecessary. τ_x is calculated as follows:

$$\tau_x = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} r'_{ij} r''_{ij}}{n(n-1)}, \tag{9}$$

where r'_{ij} and r''_{ij} are the signs reflecting the ratio of ratings R' and R'' for banks i and j, n is the total number of banks which have 2 ratings simultaneously.

The closer τ_x is to 1, the more the ratings agree. So if τ_x is close to 1, then the consensus rating and its alternative/proxy agree highly; if τ_x is close to -1, then consensus rating and its alternative/proxy contradict each other. Values of τ_x close to 0 mean no or a low correlation between ratings.

3 Data

We study if the consensus ratings of Russian banks are practically useful and can be well replicated using information from financial statements and other publicly available characteristics of those banks. We consider data from 2010, when the first regulation of the credit rating industry was introduced by the Russian Ministry of Finance, to 2016, when new industry regulations changed the landscape drastically (international agencies of the Big Three left the market, the ratings of most national agencies were excluded as elements in financial regulations, and one new agency entered the market).

Information on bank ratings is obtained from RU.Data.[10] Information on bank financial indicators and defaults is obtained from the Central Bank of the Russian Federation[11], mainly from bank report №101 (containing data on banks' key balance sheet items), №102 (income statements, published quarterly), №135 (containing data on capital requirements, liquidity requirements etc.). All explanatory variables referred to in the next section are based on parameters from these forms. Variable meanings, correlations between variables and descriptive statistics are in Appendix A, Appendix E and Appendix F respectively.

There are 134 Russian banks in our sample rated by at least two different agencies during the period. These banks and the information on them we call the Consensus Sample, since the real consensus is defined only for these banks. The statistics of the banks from the Consensus Sample can be seen in Figure 1. 17 of these banks defaulted during the considered period.

[10] http://www.ideal.ru/text.asp?Rbr=117
[11] https://www.cbr.ru/credit/

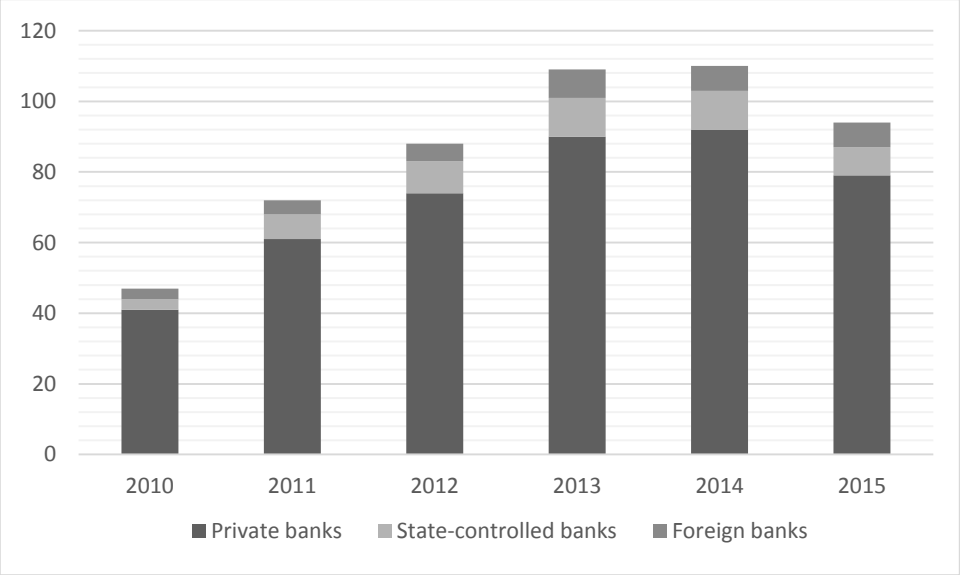


Figure1. Number of banks grouped by year and ownership.[12]
Source: authors' calculations.

The sample includes quarterly financial data and data on ratings in national scale assigned by seven CRAs. These CRAs are Fitch, S&P, Moody's, RA Expert, AKM, NRA, RUS (for the details see Appendix G). The sample size is 1,700 observations.

Consensus rating, constructed as described in Subsection 2.1, categorizes the 1,700 observations into 314 categories. Each (consensus) category consists of observations where banks have exactly the same set of ratings. Observations are quite evenly distributed over consensus categories – generally each category contains from three to ten observations.

Since we intend to compare the consensus rating with the series of econometric models we also utilize data on the whole banking system in the period. These data are divided into two samples for model estimation and validation. In order to ensure the homogeneity of the Training and Test Sample respectively to the Consensus Sample, we divide the entire data sample by time.

1. The Training Sample consists of quarterly financial data on 946 Russian banks and their defaults from 01.07.2010 to 01.07.2014. The number of defaults is 144. This sample is used as a training set to construct the scoring logit model.
2. The Test Sample consists of data on 782 Russian banks from 01.10.2014 to 01.01.2016. The number of defaults is 91. This sample is used as a test set in order to control the overfitting of the scoring logit model.

[12] State-controlled bank is defined as bank with 25% or more state ownership. Foreign bank is defined as bank with 25% or more foreign ownership, its ultimate owner is not a foreign government or any public-sector entity. Information on ownership structure has been verified by banks' and the Bank of Russia websites.

4 Results

4.1 Comparison with Logit model

Here we build the scoring logit model based on banks' public information and fit it directly to the banks' default data. Since the logit model is derived from different data by different methodology we treat the ordering of the banks which this model implies as an independent alternative for the consensus rating (see Appendix C, model I and model II).[13]

In practice the rating or scoring model is considered suitable under the following conditions:

1. the model has a quality not lower than "good" (see Appendix B);
2. the quality of the model is stable on different samples.

The logit model has been trained and tested on Sample 1 and Sample 2 respectively. Then the final model is replicated on the Consensus sample. The dataset is assigned to the three samples described in Section 3. The financial indicators used in final logit model are presented in Table 1. The fitting results are presented in Table 2.[14]

Table 1. Explanatory variables description.

Variable	Description	Sign of coefficient in model
H3	normative indicator of current liquidity; ratio of liquid assets to demand deposits and other liabilities with 30 day to maturity(without min.balance)	Negative
H4	normative indicator of long-term liquidity; ratio of loans with 365 day maturity to sum of bank capital, liabilities with 365 day to maturity with min.balance	Negative
PA5 (H7)	indicator of large credit risks concentration; ratio of sum of loans with the highest credit risk(without reserves) to bank capital	Positive
PK3	indicator of capital quality assessment; ratio of supplementary capital to sum of share capital and disclosed reserves	Positive

[13] Correlation matrix of models variables are in Appendix E.
[14] All variables are significant on 5% confidence level.

SIZE	ln(total assets)	Positive
LN(DEP/ST_LOANS)	ln((total deposits)/(total short-term loans))	Positive
TOTAL LOAN	total loan, including interbank loan	Negative
E_F	(deposits of legal entities)/(total assets)	Negative
TR RATIO	(total reserves)/(total assets)	Positive

Source: authors' calculations.

Table 2. Summary of discriminatory power metrics.

	Logit model			**Consensus rating**
	Training Sample	Test Sample	Consensus Sample	Consensus Sample
AUC	80.70%	78.68%	84.70%	80.6%
AR	61.4%	57.36%	69.4%	61.2%

Source: authors' calculations.

Table 2 shows that the logit model is not overfitted because the AR of the logit model on different sets are comparable. The AR of the logit model and the consensus rating are also comparable. Recall that the logit model is directly fitted to the default data, while the consensus ranking is constructed from *ex ante* credit assessments, therefore the consensus rating may be considered a decent alternative to econometric default models from a discriminatory power perspective.

However τ_x between the logit model and the consensus rating is 0.33. It means that according to the logit model, banks are ordered significantly differently in respect to the original consensus[15]. The reasons for this disagreement may lie in both the data and the methodology. First, the consensus rating was constructed from credit ratings assigned by different CRA. Each CRA has its own methodology and each methodology likely requires its own set of explanatory factors, so a more diverse set of financial indicators may be required. Second, the non-trivial algorithm for building the consensus rating may create interdependences between inputs and the result of the aggregation, which are difficult to detect for the logit model. Finally, the low default sample[16]

[15] Obtained with a similar logit regression, where the independent variable is real consensus rating.
[16] Rated entities defaults less often that non-rated ones.

(around 2.5% of all observations in the Consensus Sample experienced default for a one-year horizon) provides more options for shuffling observations without impacting the discriminatory power.

Taking methodological nuances into account, we study whether the consensus rating can be reproduced with a practical degree of precision by the same methodology but using alternative data. This also allows us to estimate how much information is needed to reproduce the consensus rating.

4.2 Comparison with the consensus of modeled ratings

In this subsection a proxy for the consensus rating is constructed from the model estimates of CRA ratings (see Appendix C, model I and model IV).

The idea is to build a model utilizing observable characteristics of entities for predicting CRA ratings and use these estimates as inputs in the consensus algorithm. We model credit ratings with the ordered logit model as described in Section 2. As Table 4[17] shows, the composition of independent variables is quite different in regressions models describing the ratings of different CRA. The sign of the coefficients in a few rating models is also different. This could be explained by different rating methodologies because the coefficient sign is stable for different subsample (see Table 6). It should be noted that these models also take into account factors of state support, foreign brand and ownership of the bank. Table 5 summaries rating models' accuracy measures.

The average τ_x between the proxy and original consensus increases significantly and is 0.685. Such a correlation is high, since observed τ_x between CRA ratings ranges from 0.5 to 0.8 in our sample. Therefore the result indicates that using the same algorithm and the same data composition, but not the same data quality (ratings are only predicted values), is enough to construct the model rating, which has a good degree of agreement.

Considering a wide set of variables in rating models may provide a more accurate rating estimate and a more accurate consensus replication. However, the example of international CRA shows that factors obtained on the basis of financial statements would not be enough for more precisely predicting rating estimates.

[17] Description of variables can be seen in Appendix A and correlation matrix in Appendix E.

Table 4. Explanatory factors for regression (by CRA).[18]

Factor	H1	H2	H3	H4	H7	H9.1	H12	H1_i	H2_g	H3_g	H3_i	PA3	PL5	PK3	Di	Size	e_f	T_L	L_s_t_g
AKM			+		+							+							
EXP	-		+	+		+	+					+	+				-	-	
FCH					+			-	-			+	+			-	-		
MDS	+	+			+	+					-				-	-	-		
NRA	-		-	-			-					+		+	-	-			
RUS												+		-		-	-	-	
SNP	+			-	+	+				-		+	-		-		-		+

Factor	Tr ratio	L_s_t	NonF	Res	D/L	CS	Tr ratio_g	Size_i	Di_g	Res_i	CS_g	Res_g	PK3_g	PL5_g	H10.1	e_f_i	e_f_g	F
AKM		+	+		-										-			
EXP	+	+		-							-		+		-		+	
FCH						+				-					+			
MDS				-			+	-						+	+			
NRA							+					-						
RUS		-																-
SNP	+			-		+			+						+	+		

"+" – positive sign of coefficient in model, "-" – negative sign of coefficient in model.
Source: authors' calculations.

[18] Explanation of the CRAs' abbreviation can be seen in Appendix D.

Table 5. Accuracy of ratings prediction (by CRA).

	AKM	EXP	FCH	MDS	NRA	RUS	SNP
MacFadden's R^2	68.53%	42.64%	33.21%	30%	47.20%	47.30%	31.95%
τ_x of proxy and original	78.54%	63.98%	66.79%	64.71%	73.69%	76.02%	62.06%
Share of exact matches	84.69%	67.64%	28.41%	37.45%	60.45%	52.44%	28.18%
Share of +Δ1 rang	6.70%	14.88%	15.11%	14.26%	16.91%	13.41%	24.07%
Share of -Δ1 rang	8.61%	15.34%	13.14%	13.99%	17.86%	15.04%	20.74%
Share of ±Δ2 rang	0.00%	1.99%	13.14%	14.44%	3.03%	6.91%	16.44%
Share of ±Δ3 rang	0.00%	0.15%	11.00%	9.12%	1.12%	12.20%	7.05%
Total quantity	209	652	609	1108	627	246	511

Source: authors' calculations.

Although τ_x is high, the consensus of the modeled ratings have much lower discriminatory power (0.228) than the consensus of real ratings. This happens because the econometric models significantly misclassify the eventually defaulted banks implying their ratings to be much higher than actual ones. As a result, the banks defaulted over one-year horizon are rated higher in the proxy consensus than in the original consensus. As those observations correspond to low (speculative grade) real ratings, it is worth validating the result on a subsample of banks having investment grade ratings

We also observe moderate accuracy of the econometric models for international CRA (see Table 5), so it is reasonable to examine the agreement of the consensus of the modeled and real ratings on subsamples of national CRA.

4.2.1 Investment grade subsample

Here we check whether CRA investment grade ratings are more accurately predicted by financial data and therefore the consensus of their modeled values better agrees with the consensus of their real values (see Appendix C, model I and model V).

The ratings are investment grade, which reflects the high level of financial sustainability of a company, sovereign debt or securities. CRA usually publish their definition of the investment grade class for international scales. However, this is not a general rule for national ratings. For example, NRA divides its rating into three classes: investment, tolerable, speculative; other CRA have no investment\speculative classification for national scales. Therefore, we carry out this classification based on the similarity of the interpretations of the rating categories included in the investment grade of international agencies (Fitch and Moody's) and national agencies.

Thus, the first two ratings of AK&M, the first three ratings of RA "Expert", the first six ratings of RusRating, the first seven ratings of NRA, S&P, Fitch, and Moody's were taken. As a result the "investment" sample has 1,100 observation. The consensus rating is also reconstructed, as a narrower sample is used.

Table 6 shows the independent variable of the regressions of some CRA have changed – the number of explanatory variables decreased. Moreover, the share of exact matches and matches within ± 1 rating grade significantly increase (see Table 7 and Table 8). A particularly significant increase in prediction accuracy has occurred in international CRA[19].

Therefore, the ratings for investment grade and speculative grade should be assessed separately. In addition, the consensus rating constructed on these estimates agrees well with the original consensus rating: τ_x is 0.73.

Such a result may indicate that investment grade ratings introduce fewer contradictions to the proxy consensus rating than the speculative grade ratings. Ideally, to test this assumption, it would be worthwhile constructing a proxy consensus rating only for NCRA investment grades. This, however, is not yet possible because of the lack of data on NCRA ratings. Nevertheless, the models can be further improved in terms of their predictive power. The improvement of this can be facilitated by using more suitable financial or non-financial indicators for banks. This is especially important in forecasting international CRA ratings.

[19] However, it is worth noting that the consideration of state or foreign support factors could introduce distortions between real and model estimates.

Table 6. Explanatory factors for regression (by CRA, investment grades).

Factor	H1	H2	H3	H4	H7	H9.1	H10.1	H12	H1_i	H2_g	H3_g	H3_i	PA3	PL5	PK3	Di	Size	e_f
AKM							-						+					
EXP	-		+	+		+	-	+					+	+				-
FCH					+				-	-			+				-	
MDS	+	+			+	+						-				-	-	-
NRA	-							-					+		+	-	-	-
RUS															-		-	
SNP	+			-	+	+					x		+	-		-		-

Factor	Tr ratio	L_s _t	Non f	Res	D/L	CS	F	T_L	Tr ratio _g	S_i	Di_g	Res_ i	CS_g	Res_ g	PK3 _g	PL5 _g
AKM		+	+		-											
EXP	+	+		-				-					+		+	
FCH						+						-				
MDS										-						+
NRA				-					+					-		
RUS		-					-	-								
SNP	+			-		+					+					

"+" – positive sign of coefficient in model, "-" – negative sign of coefficient in model.
Source: author's calculations.

Table 7. Accuracy of ratings prediction (by CRA, investment grades)

	AKM	EXP	FCH	MDS	NRA	RUS	SNP
MacFadden's R^2	78.80%	44.09%	36.93%	29%	50.15%	46.83%	40.96%
Share of exact matches	95.32%	70.64%	51.00%	45.05%	63.47%	63.54%	49.63%
Share of +Δ1 rang	1.17%	14.44%	16.70%	13.66%	16.69%	11.46%	20.84%
Share of -Δ1 rang	3.51%	14.44%	17.37%	13.34%	17.52%	5.21%	19.60%
Share of ±Δ2 rang	0.00%	0.47%	10.02%	15.38%	2.31%	8.33%	7.20%
Share of ±Δ3 rang	0.00%	0.00%	4.23%	9.58%	0.00%	11.46%	2.73%
Total quantity	171	637	449	637	605	192	403

Source: author's calculations.

Table 8. Difference in the accuracy of predictions[20].

	AKM	EXP	FCH	MDS	NRA	RUS	SNP
Consensus Sample	100.00%	97.86%	56.66%	65.70%	95.22%	80.89%	72.99%
Investment Grade Subsample	100.00%	99.52%	85.07%	72.05%	97.68%	80.21%	90.07%
Δ precision	0.00%	1.66%	28.41%	6.35%	2.46%	-0.68%	17.08%

Source: author's calculations.

4.2 National CRA subsample

It is reasonable to assume that more precise rating estimates will give a proxy consensus rating which is more similar to the consensus rating. That is to say, the proxy is constructed from model estimates of national CRA ratings (see Appendix C, model I and model VI).

The same observations as in the previous part were used to predict NCRA ratings. Note that the consensus rating is determined for banks with at least two ratings by different CRAs. Therefore, only 394 observations were used in the comparison of the proxy and the original consensus ratings based on ratings from NCRA.

[20] Comparison of share of exact matches plus share of ±Δ1 rang of two samples.

Table 9. Explanatory factors for regression (by National CRA).

Factor	H1	H3	H4	H7	H9.1	H10.1	H12	PA3	PL5	PK3	Di	Size	e_f
AKM		+		+		-		+					
RUS								+		+		-	-
EXP	-	+	+		+	+	+	+	+				-
NRA	+	-	+				-	+		+	-	-	-

Factor	Tr ratio	L_S_T	NonF	Res	F	T_L	Tr ratio_g	CS_g	Res_g	Or	State	PK3_g
AKM	+	+	+									
RUS	-	-			-	-				+	+	
EXP	+	+		-		-		-				-
NRA				-			+		-			

"+" – positive sign of coefficient in model, "-" – negative sign of coefficient in model.
Source: authors' calculations.

In this sample, τ_x between the proxy and original consensus is 0.622. Note that this value is slightly smaller than the values for the previous samples despite the fact that rating prediction accuracy for this subsample is higher than for the Consensus Sample (Table 10.2 and Table 10.1 respectively).

Table 10.1 Accuracy of ratings prediction (by National CRA, full sample).

	AKM	EXP	NRA	RUS
MacFadden's R^2	68.53%	42.64%	47.20%	47.30%
Share of exact matches	84.69%	67.64%	60.45%	52.44%
Share of +Δ1 rang	6.70%	14.88%	16.91%	13.41%
Share of -Δ1 rang	8.61%	15.34%	17.86%	15.04%
Share of ±Δ2 rang	0.00%	1.99%	3.03%	6.91%
Share of ±Δ3 rang	0.00%	0.15%	1.12%	12.20%
Total quantity	209	652	627	246

Source: author's calculations.

Table 10.2 Accuracy of ratings prediction (by National CRA, paired ratings).

	AKM	EXP	NRA	RUS
Share of exact matches	92.12%	84.67%	76.54%	83.33%
Share of +Δ1 rang	7.88%	14.67%	22.22%	12.28%
Share of -Δ1 rang	0.00%	0.00%	0.00%	0.00%
Share of ±Δ2 rang	0.00%	0.67%	1.23%	1.75%
Share of ±Δ3 rang	0.00%	0.00%	0.00%	2.63%
Total quantity	165	300	243	114

Source: author's calculations.

First, the smaller agreement between the proxy and original consensus rating can be explained by the smaller sample (fewer observations and CRAs). The fewer the observations of CRAs in a sample, the more sensitive to a single mismatch the consensus rating becomes. Moreover, national CRAs agreed less on the ordering of banks according to their credit quality, so they are more likely to face the problem described above in the subsection on consensus of financial indicators.

Second, the disagreement is aggravated by the presence of speculative class ratings. Empirical calculations in Karminsky, Peresetsky (2008), and Hung, Cheng (2013) confirm these considerations: the largest errors in rating predictions correspond to the last investment grade and speculative grades. It may be a sign that another distinct model is required to accurately describe that group of ratings.

However, the obtained level of agreement is still high enough from practical point of view.

5 Conclusion

In this paper, we propose a two-step methodology for studying (1) if the aggregate rating performs as well as the best-practice credit risk models in its own domain of rated companies and (2) if an aggregate rating can be accurately modeled and, therefore, extrapolated to non-rated companies without loss of competitiveness. We argue that an aggregate rating that successfully passes these two steps suits credit analysis purposes and can be used in building and validating credit models.

We apply this methodology to the aggregated rating of Russian banks constructed as a consensus of ratings assigned by different rating agencies. We examine whether the aggregate rating constructed as a consensus of individual credit ratings can be accurately predicted by publicly available non-rating information.

We show that an aggregate (consensus) rating is comparable to a financial-data-based econometric default model in term of discriminatory power; however, the corresponding orderings have a low agreement. We also found that using models for predicting initial credit ratings allows the building of a proxy that has high agreement with the original aggregate rating, but the original aggregate rating outperforms the proxy in terms of discriminatory power. It was also found that greater agreement between the original aggregated rating and the proxy can be achieved on the

subsample of investment grade ratings. Nonetheless, this does not mean that a consensus is more suitable for investment grade ratings than for speculative.

Therefore, our approach shows that consensus rating performs well in its own domain and can be used in it as a factor for credit risk analysis and validation of internal models. However, we could not replicate the consensus rating with the available information and econometric models well enough to extrapolate outside its domain or to use it as a universal credit risk indicator.

Our approach can be applied to arbitrary aggregate ratings constructed with any rating inputs and aggregation methodology. The authors intend to exploit this in further research.

References:

Agarwal, V., and Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. Journal of Banking & Finance, 32(8), 1541-1551.

Aivazian, S., Golovan, S., Karminsky, A. and Peresetsky, A. (2011) O podhodah k sopostavleniju rejtingovyh shkal [The approaches to the comparison of rating scales], Applied Econometrics, 3(23), 13–40 (in Russian).

BCBS (2005). Studies on the validation of internal rating systems. Working Paper № 14. URL: https://www.bis.org/publ/bcbs_wp14.htm

Brandt, F., Conitzer, V., Endriss, U., Procaccia, A. D. and Lang, J. (Eds.). (2016). Handbook of computational social choice. Cambridge University Press.

Buzdalin A.V., Zanochkin A.Yu., Kurbangaleev M.Z., Smirnov S.N. (2017) Agregatsiya kreditnyh reytingov kak zadacha postroeniya konsensusa v sisteme ekspertnyh otsenok [Aggregation of credit ratings as a task of building consensus in the system of expert assessments] Global markets and Financial Engineering. 4(3). 181-207 (in Russian).

Campbell, J. Y., Hilscher, J. and Szilagyi, J. (2008). In search of distress risk. The Journal of Finance, 63(6), 2899-2939.

Eisl, A., Elendner, H. and Lingo, M. (2013). Re-Mapping credit ratings.URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1836877

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. Journal of Multi-Criteria Decision Analysis, 11(1), 17-28.

Fitch Ratings (2014), Definitions of Ratings and Other Forms of Opinion. URL: https://www.fitchratings.com/web_content/ratings/fitch_ratings_definitions_and_scales.pdf

Grün, B., Hofmarcher, P., Hornik, K., Leitner, C. and Pichler, S. (2013). Deriving consensus ratings of the big three rating agencies. The Journal of Credit Risk, 9(1), 75-98.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Hofmarcher, P., Kerbl, S., Grün, B., Sigmund, M. and Hornik, K. (2014). Model uncertainty and aggregated default probabilities: new evidence from Austria. Applied Economics, 46(8), 871-879.

Hornik, K., Jankowitsch, R., Leitner, C., Lingo, M., Pichler, S. and Winkler, G. (2008). A latent variable approach to validate credit rating systems. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1269306

Hung, K., Cheng, H. W., Chen, S. S. and Huang, Y. C. (2013). Factors that affect credit rating: An application of ordered probit models. Romanian Journal of Economic Forecasting, 16(4), 94-108.

Karminsky, A. M., Hainsworth, R. N. and Solodkov, V. M. (2013). Arm's length method for comparing rating scales. Eurasian Economic Review, 3(2), 114 -135.

Kavussanos, M.G. and Tsouknidis, D.A. (2016). Default risk drivers in shipping bank loans. Transportation Research Part E: Logistics and Transportation Review, 94, 71-94.

Lehmann, C. and Tillich, D. (2015). Applied Consensus Information and Consensus Rating. Dresdner Beiträge zu Quantitativen Verfahren Nr. 61/15. URL: https://tu-dresden.de/bu/wirtschaft/ressourcen/dateien/forschung_wiwi/publikationen/qv/ddpqv2015_61.pdf?lang=en

Lehmann, C. and Tillich, D. (2016). Consensus Information and Consensus Rating. In Operations Research Proceedings 2014. Springer International Publishing, 357-362.

Peresetsky, A. and Karminsky, A. (2008). Models for Moody's bank ratings. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1304590

Pomazanov, M. (2016). Upravlenie kreditnym riskom v banke: Podhod vnutrennix rejtingov (PVR) [Credit Risk Management in the Bank: Internal Ratings-Based Approach (IRB)]. Moscow: Urait. (In Russian).

Standard&Poors (2016). S&P Global Ratings Definitions. URL: https://-www. standardandpoors. com/en_US/web/guest/article/-/view/sourceId/504352.

Tasche, D. (2010). Estimating discriminatory power and PD curves when the number of defaults is small. arXiv:0905.3928. URL: https://arxiv.org/pdf/0905.3928v2.pdf

Appendix A. Variables description.

Variable	**Interpretation**
"Variable name"_index	A variable created by multiplying a financial indicator by a dummy variable State(index g) or Foreign(index i)
Cash	Cash
CS	(Bank capital/Total assets)
D/L	ln(Total deposits/Long term loans)
Da	Deposits of legal entitiess
Db	Interbank deposit
Dc	Deposits with the Bank of Russia
Df	Deposits of foreign entities
Dg	State bodies deposits
Dh	Retail deposits
Di	Deposit of financial companies
Di	Deposit of financial companies
e_f	(Deposits of legal entities)/(Bank capital)
F	Dummy, foreign bank
H1	Capital adequacy ratio
H10.1	Ratio of sum of loan to bank insiders to bank capital
H12	Ratio of sum of investments in shares of other legal entities
H2	Instant liquidity ratio
H3	Normative indicator of current liquidity
H4	Normative indicator of long-term liquidity
H9.1	Ratio of sum of loans to the bank owners to bank capital
L_short_tot (L_S_T)	Total short-term loans
La	Legal entities loans
Lb	Interbank credit
Lc	Loans issued by the Bank of Russia
Lf	Loans to foreign entities
Lg	State bodies credits
Lh	Retail loan
Li	Loans to financial companies
ln(dep/st_loans) D_S	ln(Total deposits/Total Short-term loans)
NonF	Ratio of nonfinancial entities loans to total loans
Or	Borrowed reserves
PA3	Indicator ofoverdue loans share
PA5 (H7)	Maximum size of major credit risks
PK3	Indicator of capital quality assessment
PL5	(Interbank deposits- Inrebank Loans)/(Total deposits)
Res	Excess reserves

Secur	Purchased securities
State	Dummy, state-controlled bank
Total loans (T_L)	Bank account loans
tr ratio (T_R)	(Total reserves)/(Bank capital)

Appendix B. Model quality in terms of discriminatory power.

AR interval, 1 years risk horizon	Model quality	Significance for risk-management
80% and more	Advanced	The rating system allows to automate decision-making process for credit transactions, loan reserves and capital allocation
60-80%	Very good	
40-60%	Good	The rating result should be of great weight in decision-making process for credit transactions
20-40%	Medium	Rating can only be regarded as informative (referential). Reservation and allocation of capital should be implemented with standardized criteria (Bank of Russia or Standardized Approach of Basel II)
20% and less	Insufficient	The rating result have not be taken into account in decision-making process for credit transactions Reservation and allocation of capital must be implemented with standardized criteria (Bank of Russia or Standardized Approach of Basel II)

Source: Pomazanov (2016), p.54

Appendix C. Components of consensus ratings and its alternative.

№ model	**Estimated parameter**	**Basic data**	**Method**
I	Consensus	Ratings	Algorithm
II	Scoring logit model	Public information	Logit model
III	Consensus of financial indicators	Financial indicators	Algorithm + logit model
IV	Consensus of modeled rating (a)	Public information, ratings	Algorithm + ordered logit model
V	Consensus of modeled rating (b)	Public information, investment class ratings	Algorithm + ordered logit model
VI	Consensus of modeled rating (c)	Public information, NCRA ratings	Algorithm + ordered logit model

Appendix D. Explanation of the CRAs' abbreviation.

AKM	Rating Agency AK&M
EXP	Rating Agency RAEX («Expert RA»)
FCH	FitchRatings
MDS	Moody's Analytics
NRA	«National Rating Agency»
RUS	RusRating
SNP	Standard&Poor's

Appendix E. Correlation matrix for models variable.

In research studies on the credit ratings aggregation, the VIF (variance inflation factor) or the correlation matrix of explanatory variables is used to measure multicollinearity, if the question of testing multicollinearity is discussed. All are unanimous that the VIF values are 10 or more, the correlation coefficient is 0.75 in modulus and more precisely indicates that the use of a such pair of variables simultaneously leads to multicollinearity. In this paper, the authors hold the same opinion on the correlations.

Model II

	h3	h4	pa5	pk3	size	D_S	T_L	e_f	Tr_ratio
h3	1.000								
h4	-0.158	1.000							
pa5	-0.281	0.032	1.000						
pk3	0.074	-0.039	-0.149	1.000					
size	-0.010	0.230	-0.230	0.275	1.000				
D_S	-0.068	0.265	0.015	0.040	0.449	1.000			
T_L	-0.028	0.127	-0.093	0.249	0.505	0.263	1.000		
e_f	0.044	0.041	-0.066	-0.168	0.072	0.046	0.018	1.000	
Tr_ratio	0.049	-0.114	-0.078	-0.158	-0.325	-0.184	-0.115	-0.009	1.000

Source: authors' calculations.

Model IV

AKM

	h10.1	h3	h7	L_S_T	nonf	pa3	D/L	orcha
h10.1	1.000							
h3	-0.003	1.000						
h7	0.062	-0.280	1.000					
L_S_T	-0.127	0.103	-0.190	1.000				
nonf	0.148	-0.372	0.365	-0.354	1.000			
pa3	-0.074	0.167	-0.322	0.102	-0.155	1.000		
D/L	-0.203	0.012	-0.221	0.209	-0.149	0.045	1.000	
orcha	0.060	0.044	-0.077	0.036	-0.031	0.182	-0.295	1.000

Source: authors' calculations.

FCH

	h101	h7	size	pl5	pa3	CS	e_f	res_i	h1_i	size_i	h2_g
h101	1.000										
h7	0.067	1.000									
size	-0.188	-0.216	1.000								
pl5	-0.073	0.096	-0.078	1.000							
pa3	-0.074	-0.321	0.065	-0.019	1.000						
CS	-0.117	-0.156	0.076	-0.028	0.094	1.000					
e_f	-0.106	-0.076	0.062	0.017	-0.100	-0.062	1.000				
res_i	-0.202	-0.322	0.232	-0.027	0.080	0.105	0.161	1.000			
h1_i	-0.135	-0.190	0.128	-0.019	0.021	0.110	0.079	0.572	1.000		
size_i	-0.202	-0.326	0.229	-0.024	0.078	0.106	0.166	0.499	0.580	1.000	
h2g	-0.065	-0.016	0.349	-0.024	0.050	-0.146	0.070	-0.079	-0.046	-0.079	1.000

Source: authors' calculations.

RUS

	L_S_T	size	T_L	pa3	pk3	e_f	T_R	F
L_S_T	1.000							
size	0.040	1.000						
T_L	-0.022	0.500	1.000					
pa3	0.094	0.065	-0.036	1.000				
pk3	0.117	0.042	0.095	-0.006	1.000			
e_f	0.093	0.062	0.019	-0.100	0.052	1.000		
T_R	0.035	-0.339	-0.115	0.182	0.002	-0.023	1.000	
F	0.056	0.216	0.006	0.079	0.025	0.178	-0.080	1.000

Source: authors' calculations.

	h1	h101	h12	h3	h4	h91	L_S_T	pl5	T_L	pa3	res	e_f	T_R	pk3_g	CS_g	e_f_g
h1	1.00															
h101	-0.03	1.00														
h12	0.08	-0.09	1.00													
h3	0.18	-0.01	-0.01	1.00												
h4	-0.09	0.15	-0.02	-0.15	1.00											
h91	0.01	0.13	0.04	0.00	0.01	1.00										
L_S_T	0.00	-0.14	-0.03	0.11	-0.28	-0.01	1.00									
pl5	-0.05	-0.07	-0.01	-0.12	-0.04	-0.08	-0.22	1.00								
T_L	0.04	-0.05	0.17	-0.03	0.13	-0.02	-0.02	-0.02	1.00							
pa3	0.04	-0.07	0.00	0.17	0.02	0.00	0.09	-0.02	-0.04	1.00						
res	0.09	-0.14	0.28	0.04	0.19	-0.03	0.07	-0.03	0.42	0.24	1.00					
e_f	-0.10	-0.11	0.02	0.03	0.05	0.00	0.09	0.02	0.02	-0.10	0.04	1.00				
T_R	0.03	0.06	-0.05	0.04	-0.15	0.04	0.04	0.00	-0.12	0.18	0.01	-0.02	1.00			
pk3_g	0.06	-0.02	0.17	-0.01	0.07	-0.02	-0.01	-0.05	0.74	-0.02	0.26	-0.03	-0.08	1.00		
CS_g	0.01	-0.10	0.33	-0.04	0.15	-0.03	-0.05	-0.01	0.40	0.04	0.34	0.04	-0.10	0.59	1.00	
e_f_g	-0.04	-0.07	0.30	-0.04	0.16	-0.04	-0.06	0.04	0.25	0.04	0.33	0.21	-0.06	0.25	0.73	1.00

Source: authors' calculations.

	h1	h10.1	h2	h7	h9.1	di	size	res	e_f	pl5_g	h3_i	size_i	T_R_g
h1	1.000												
h10.1	-0.029	1.000											
h2	0.084	-0.012	1.000										
h7	-0.103	0.067	-0.289	1.000									
h9.1	0.012	0.125	0.000	0.052	1.000								
di	0.043	-0.104	-0.028	-0.098	-0.014	1.000							
size	0.065	-0.188	-0.021	-0.216	-0.060	0.649	1.000						
res	0.048	-0.091	-0.015	-0.117	-0.021	0.711	0.616	1.000					
e_f	-0.095	-0.106	0.026	-0.076	0.003	0.131	0.062	0.041	1.000				
pl5_g	0.005	-0.123	-0.019	0.110	0.001	0.161	0.101	0.127	0.092	1.000			
h3_i	0.064	-0.094	0.147	-0.257	-0.032	-0.006	0.105	0.004	0.218	-0.010	1.000		
size_i	0.047	-0.202	0.007	-0.326	-0.044	0.036	0.229	0.054	0.166	-0.015	0.658	1.000	
T_R_g	-0.037	-0.041	-0.049	0.048	-0.040	0.326	0.168	0.215	0.064	-0.011	-0.050	-0.074	1.000

Source: authors' calculations.

	h1	h101	h4	h7	h91	di	pl5	pa3	CS	res	e_f	T_R	h3_g	LST_g	Di_g	e_f_i
h1	1.00															
h101	-0.03	1.00														
h4	-0.09	0.15	1.00													
h7	-0.10	0.07	0.04	1.00												
h91	0.01	0.13	0.01	0.05	1.00											
di	0.04	-0.10	0.17	-0.10	-0.01	1.00										
pl5	-0.05	-0.07	-0.04	0.10	-0.08	0.01	1.00									
pa3	0.04	-0.07	0.02	-0.32	0.00	0.01	-0.02	1.00								
CS	0.27	-0.12	-0.13	-0.16	-0.01	-0.08	-0.03	0.09	1.00							
res	0.09	-0.14	0.19	-0.31	-0.03	0.57	-0.03	0.24	0.11	1.00						
e_f	-0.10	-0.11	0.05	-0.08	0.00	0.13	0.02	-0.10	-0.06	0.04	1.00					
T_R	0.03	0.06	-0.15	-0.08	0.04	-0.17	0.00	0.18	-0.01	0.01	-0.02	1.00				
h3_g	-0.01	-0.08	0.16	0.00	-0.04	0.54	-0.03	0.08	-0.14	0.31	0.06	-0.06	1.00			
LST_g	0.01	-0.02	0.05	0.03	-0.03	0.33	-0.06	0.03	-0.08	0.16	0.03	0.02	0.68	1.00		
di_g	0.03	-0.04	0.12	-0.10	-0.03	0.88	0.01	0.02	-0.07	0.41	0.10	-0.12	0.64	0.40	1.00	
e_f_i	0.03	-0.16	0.11	-0.29	-0.03	0.00	-0.01	-0.01	0.09	0.14	0.35	-0.07	-0.07	-0.04	-0.05	1.00

Source: authors' calculations.

	h1	h12	h3	h4	di	size	pa3	res	pk3	e_f	res_g	T_R_g
h1	1.000											
h12	0.079	1.000										
h3	0.177	-0.006	1.000									
h4	-0.091	-0.024	-0.155	1.000								
di	0.043	0.333	-0.041	0.168	1.000							
size	0.065	0.292	-0.004	0.259	0.550	1.000						
pa3	0.039	-0.001	0.167	0.019	0.009	0.065	1.000					
res	0.092	0.276	0.041	0.192	0.569	0.570	0.236	1.000				
pk3	0.105	0.000	0.088	-0.026	-0.015	0.042	-0.006	0.076	1.000			
e_f	-0.095	0.018	0.030	0.049	0.130	0.062	-0.100	0.038	0.052	1.000		
res_g	-0.025	0.297	-0.053	0.196	0.512	0.416	0.048	0.375	-0.075	0.071	1.000	
T_R_g	-0.037	0.155	-0.027	0.147	0.326	0.168	0.074	0.194	-0.059	0.064	0.615	1.000

Source: authors' calculations.

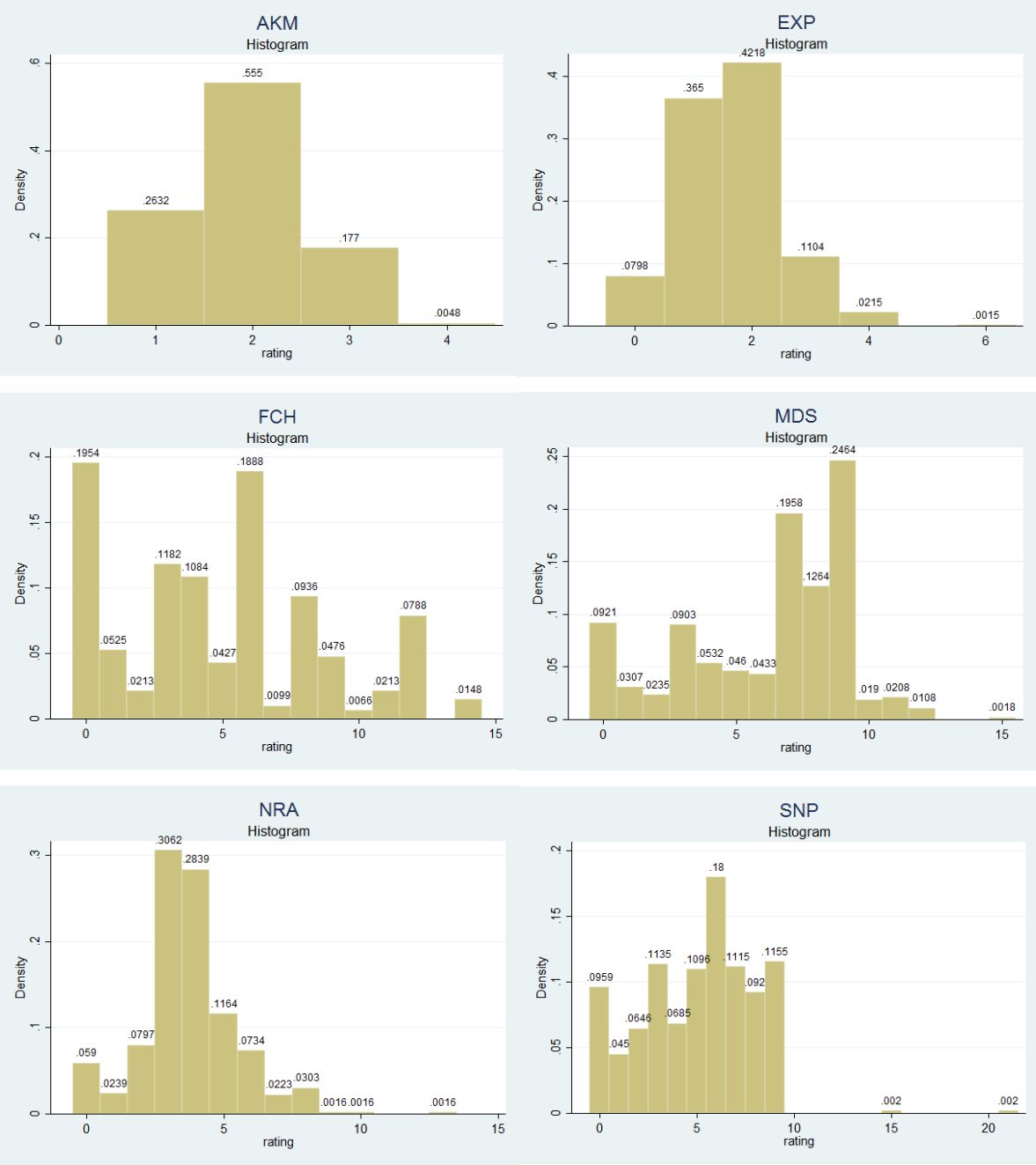
Appendix F. Descriptive statistics of explanatory variables.

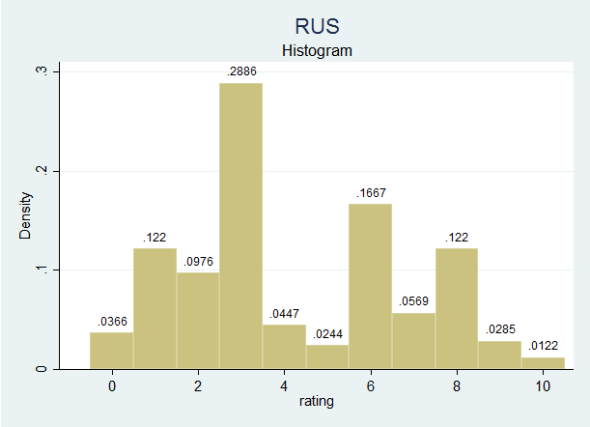
Variable	Obs	Mean	Std. Dev.	Min	Max
cs	1,700	0.29721	0.137209	0.00986	0.800349
cs_g	1,700	0.021651	0.074974	0	0.543285
D/L	1,700	17.79716	1.514874	12.88666	23.50904
D_S	1,700	21.45275	2.388049	14.88388	32.77662
di	1,700	8839488	2.43E+07	0	2.36E+08
di_g	1,700	4734544	2.23E+07	0	2.36E+08
e_f	1,700	0.092	0.065688	0.000756	0.466124
e_f_g	1,700	0.009814	0.036305	0	0.246516
e_f_i	1,700	0.009027	0.041216	0	0.466124
f	1,700	0.066471	0.249176	0	1
h1	1,700	7.551924	11.70625	0	77.21
h1_i	1,700	0.638847	4.127527	0	44.61
h101	1,700	0.928959	0.726141	0	6.93
h12	1,700	1.799224	4.395179	0	24.91
h2	1,700	71.89571	61.07142	10.84	898.6
h2g	1,700	5.618647	18.87149	0	168.86
h3	1,700	101.1857	63.59263	18.44	1014.37
h3_g	1,700	8.535929	27.94024	0	200.37
h3_i	1,700	8.725247	48.66564	0	926.75
h4	1,700	69.06213	27.90633	0	130.98
h7	1,700	279.3108	151.2251	0	866.14
h91	1,700	2.085876	9.917066	0	368.71
L_S_T	1,700	0.06672	0.121433	0	0.934414
LST_g	1,700	0.00398	0.019688	0	0.281264
nonf	1,700	0.825172	0.197842	0	1
or	1,700	2444386	1.14E+07	7474	1.59E+08
pa3	1,700	0.047723	0.046182	0	0.412712
pa5	1,700	279.3108	151.2251	0	866.14
pk3	1,700	4.975535	6.181481	-0.79566	67.44503
pk3g	1,700	0.348523	2.549823	-0.35753	67.44503
pl5	1,700	0.012728	0.050668	-0.45693	0.409203
pl5_g	1,700	0.000952	0.01689	-0.1744	0.134035
re_i	1,700	0.97307	3.663599	0	16.25949
res	1,700	13.15813	1.698182	7.885705	18.50355
res_g	1,700	1.401982	4.391423	0	18.50355
size	1,700	18.62496	1.550322	14.9372	25.45365
size_i	1,700	1.321562	4.96404	0	22.5007
state	1,700	0.094118	0.292078	0	1

T_R	1,700	0.011994	0.007954	0.001833	0.090487
T_R_g	1,700	0.000937	0.003393	0	0.034361
Total loan	1,700	2.33E+08	1.14E+09	426211	1.75E+10

Source: authors' calculations.

Appendix G. Ratings' histograms.





Source: authors' calculations.

Authors:

1) Marat Kurbangaleev, National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab.
   E-mail: mkurbangaleev@hse.ru
2) Victor Lapshin, National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab; PhD in Mathematical Modelling, Numerical Methods and Software Complexes.
   E-mail: vlapshin@hse.ru
3) Zinaida Seleznyova, National Research University Higher School of Economics (Moscow, Russia). Financial Engineering & Risk Management Lab.
   E-mail: zseleznyova@hse.ru