Low Default Portfolios in Basel II and Basel III as a Special Case of Significantly Unbalanced Classes in Binary Choice Models¹

Henry Penikas, *Bank of Russia; National Research University Higher School of Economics; Lebedev Physical Institute of the Russian Academy of Sciences*
penikas@hse.ru

In contemporary world, binary choice models are used in many areas. However, for all such areas, a problem arises when the share of one of the classes in the data sample is small. If this share is significantly small, this class is referred to as low default class. The purpose of this paper is to examine the definitions of such a portfolio and the approaches to building models on its basis. Although various methods exist for obtaining results, this paper shows that distinguishing a low default portfolio class, on the one hand, benefits banks, as does any more detailed segmentation, but, on the other hand, it deteriorates the statistical properties of the models for the probability of default. It is therefore justified that for the internal rating-based approach in the framework of Basel II and Basel III the regulator should require that banks build their models based on combined data sets discouraging them from setting excessive low default portfolio classes.

Keywords: *Basel II, IRB, low default portfolio, LDP, probability of default, unbalanced classes, binary choice (response) model*

JEL Codes: *G21, G28*

Citation: Penikas, H. (2020). Low Default Portfolios in Basel II and Basel III as a Special Case of Significantly Unbalanced Classes in Binary Choice Models. *Russian Journal of Money and Finance*, 79(2), pp. 101–128.

doi: 10.31477/rjmf.202002.101

1. Binary choice models and the problem of unbalanced classes

Binary choice (response) models have become widespread in practically every area of human activities. In finance, they make it possible to build a forecast as to whether or not a loan will be repaid, or, which is effectively the same, whether or not

¹ The author expresses his gratitude to Sergey Aliyev, Mikhail Bukhtin, Nadezhda Ivanova, Vladimir Korovin, Alexey Lobanov (Bank of Russia), Vladimir Babikov (Business Systems Consult), Vladimir Kozlov (Raisk.ru), Dmitry Malakhov (HSE), Igor Farrakhov (Riskfin), and two anonymous reviewers for their comments. The author's opinion may differ from the standpoints of affiliated institutions.

a borrower will end up in default (Lopez, 2004); in medicine, whether or not a patient will recover (Bakbergenuly et al., 2016); and in production, whether or not equipment will fail (Zaigraev and Kaniovski, 2013). Thus, the primary value of binary choice models is in categorising observations into one of two classes, or, in other words, for dividing into classes or discriminating between classes. The best model is the one that ensures the most accurate classification, division, or discrimination.

Despite the benefits of binary choice models, their extensive use has begun to cause worries among researchers. Some even started to accuse these models of being unethical – namely, that the models are biased against people. Let us consider two cases. In case one models for assessing the suitability of a candidate for a job vacancy may, all else being equal, be less likely to offer the position to women. Case two relates to loan approval models that may, all else being equal, be less likely to suggest a decision to grant a loan to certain population categories (for example, Hispanics or people of colour in the US, see Fuster et al., 2018). The researchers even complain that the models were not given input attributes for discrimination (for example, gender or race, as in the above example), but still discriminated the said categories when forming recommendations. On the one hand, the models are not to blame for the discrimination that they are intended to ensure, as often such attributes as gender or race are correlated with factors such as income, education level and number of children. Therefore, even if there are no gender or race factors in the primary data array, the models will give a discriminating recommendation if the primary array contains related factors (for example, income level). On the other hand, there is another reason for such discriminatory behaviour of models which I would like to cover in more detail in this paper.

The fact is that the second reason for discriminatory recommendations is not just the input of factors related to the discriminatory attributes of gender or race but the input of disproportionate, unbalanced, unmatched (Newby et al., 2013), sparse (McCullah and Nelder, 1989, p. 120), or rare (King and Zeng, 2001) classes. In the above case one, the primary dataset has higher share of men, so, all else being equal, the *a priori* probability (not an *a posteriori* model-implied one) of employing women is lower (consider an extreme case that the dataset has no recordings about women being employed; then a natural expectation originating from such a dataset is that men only are employed); similarly, the share of white people is greater among those who regularly repay loans, so they are more likely to be approved for loans than others. Despite the clear logic of data processing in this case – namely, the prediction of a higher probability for any new data point to be classified as a member of a more represented class (a class having a larger rate of observations in the total sample), the question arises: With what size (absolute or relative) can the forecast of this model be trusted, and, consequently, if it cannot be trusted, what must be done? After all, the need for a recommendation based on the data and statistics as to whether or not to make a decision (on hiring or issuing a loan, for instance) still remains.

An extreme case where the problem of unbalanced classes occurs in finance is the concept of the low default portfolio (LDP). It describes a situation where

the class of non-performing loans (defaults) has a significantly low share in the sample for building a binary choice model – namely, the forecast model for loan repayment (non-repayment). The matter of LDPs is especially relevant for banks using the internal rating-based (IRB) approach, which was originally introduced by Basel II (Basel Committee on Banking Supervision, 2006) and retained in Basel III (Basel Committee on Banking Supervision, 2017, pp. 138–182).

To understand the specifics of LDP modelling, including for building binary choice models with significantly unbalanced classes, the paper will be structured as follows. In Section 2, we will introduce the basic concepts used in the IRB approach and their relation to the LDP. In Section 3, we will consider definitions of LDP that include both absolute and relative metrics. In Section 4 we will consider development of probability of default models for LDPs. Section 5 will examine a hypothetical example demonstrating how the allocation of an LDP may lead to substantial benefits for a bank to the detriment of the statistical properties of mathematical models. In Section 6, we will summarise and draw specific conclusions useful to the regulator, bank risk managers and to researchers working with binary choice models.

2. Basic terms used in the IRB approach

The logic of the Basel II agreement is to enable banks to calculate the denominator of the capital adequacy ratio (H1 and H20 in Russia) based on their own statistics on non-performing loans. The denominator is understood to be the value of risk-weighted assets (RWA) as it pertains to credit risk in monetary (value) terms, or the risk weight (RW) as a relative measure of credit risk. To that end, the bank must estimate the probability of default (PD) for practically every balance sheet and off-balance sheet item exposed to credit risk. This parameter is key in the IRB approach as it serves as the input parameter for the regulatory formula (Bank of Russia, 2015) that corresponds to the credit portfolio loss distribution model described by the Czech mathematician Oldrich Vasicek (Vasicek, 1987). Here, we will not dwell on the properties or disadvantages of Vasicek's model. We will only mention that, as with Value-at-Risk (the maximum possible loss after we exclude all the worst outcomes whose predefined combined probability is at most p), the model involves evaluating the worst losses in the event of the materialisation of two factors: a systematic risk factor and an idiosyncratic risk factor. The effect of the latter depends on the estimated probability of default for the exposure, while the effect of the former is set by the regulator in advance on the assumption of 0.1% of its worst materialisation. One minus the latter value of 0.1% (namely 99.9%) is called the significance level for the materialisation of the systematic risk factor.

If the bank needs to estimate the probability of default for all categories of the bank's assets, the question arises as to whether the bank should build a single model or several models to do this.

First, the regulator requires that such a parameter of Vasicek's model as asset correlation (R) be different for different classes of assets (exposures). For example, for corporate exposures, R should vary from 12% to 24%, while for exposures of global systemically important financial institutions (SIFI; in Russia, there are only subsidiaries of SIFIs) R should take a value from 15% to 30%. As a result, RWAs for SIFIs, all else being equal, will be higher than those required from corporates.

At the same time, the regulator does not prohibit banks from singling out segments of exposures within classes. Thus, banks have a variety of options – from building a single model (if the bank can prove that its exposures are homogeneous, which is beyond the scope of this paper) to building a number of models equal to the number of segments.

The problem of PD estimation can be addressed in two primary ways: non-parametric method (historical) or parametric method (model). The non-parametric method was historically considered to be the simplest one. It involves assigning exposures the historically observable default rate as an estimate of probability of default for them. It is obvious that this approach would reflect the 'average temperature for the hospital' (misleading average) as it would assign the same value to both good and bad borrowers. Therefore, Basel II requires the identification of at least seven non-default groups (ratings, rating categories, grades). The most common practice is to divide the unit scale of probability of default into segments equal in terms of the growth rate of the probability of default in them on a logarithmic scale (Ozdemir and Miu, 2009, p. 136). For example, below is the rating scale of the international rating agency Standard & Poor's. It has seven grades (ratings); the eighth grade (rating) D, not given, corresponds to defaults. For example, the average probability of default between the grades increases by 40%, or is 1.4 times higher on average on the logarithmic scale. Based on the grades, a rating scale (Master Scale) is built, in which every grade is assigned corresponding minimum and maximum thresholds of probability of default.

Figure 1. Example of a rating scale based on Standard & Poor's data (1981–2018)

Source: Standard and Poor's, author calculations

An exposure can be categorised into a particular grade by experts (for example, this is the only method advised for the LDP segment by Ozdemir and Miu, 2009, p. 2, and Gordy and Heitfield, 2010, p. 60) or based on a mathematical (statistical or econometric) model. The latter can employ two options. In the first one, the model gives a grade forecast, and the borrower is assigned the average historically observable default rate for the entire grade as a PD estimate. Alternatively, the model can give a point-in-time PD forecast. This forecast is then compared to the grade boundaries and an exposure is assigned into the grade within the boundaries of which the PD estimate falls. As a result, the exposure is assigned the average historical default rate for this grade as an estimate of the probability of its default. It is natural to expect that the point-in-time PD estimate and the average historical default rate will differ. At the level of the entire portfolio, the author did not identify any systematic or material benefits or losses for the bank from such a discrepancy. Therefore, for the purpose of this paper we will not dwell on this discrepancy.

If a bank opts for the parametric approach – namely, building a mathematical (statistical) model of the PD forecast – it needs to form a training sample that includes cases of defaults and non-defaults. This is where the concept of the LDP appears. Although the meaning of a LDP must be clear based on the above logic, in practice its definition is not clear and unambiguous, as will be shown below.

The most vivid examples of LDPs are the following two types: exposures to the largest borrowers by organisation size (Pluto and Tasche, 2006) or exposures to new recently created products (usually retail, see Sabato, 2006). In the first example, defaults are naturally rare, while in the second example they might have not accumulated yet. Thus, the problem of LDPs in retail is relevant for the applicative segments characterising, for example, the first three months of service of a retail client by the bank, and not for the behavioural ones.

When adopting Basel II, the Basel Committee realised the relevance of building IRB models for the LDP segment. Then, they formulated a key guideline: to try, wherever possible, to combine data – that is, in the above terms this should be read as a recommendation to single out fewer segments within a class so as not to reduce the number of defaults in the segments (Basel Committee on Banking Supervision, 2005, p. 5). After the crisis of 2007–2009, the Basel Committee returned to the question of what to do with LDP portfolios (Basel Committee on Banking Supervision, 2016). They even discussed the prospect of banning the development of IRB models for them, but ultimately this suggestion was not approved in Basel III.

To try to understand the logic of the Basel Committee and (or perhaps, or) of banks, it is worth taking a few steps back and asking the following questions:

- 1) How natural is the origin of LDP segments? It is not clear whether the LDP is the nature of individual assets (for instance, Pomazanov (2016, p. 68) notes that international banks are placed in LDPs, while in Russia banks are not in the LDP segment as there are extensive statistics on

their defaults, bankruptcies and license revocations), or whether it is the result of certain interests of banks in selecting such portfolios.

- 2) If a bank has such an interest, how might it manifest itself? For example, if we are talking about a possible release of capital because of using the IRB approach, can the bank maximise the volume of such a release by using an appropriate LDP?

To answer these questions, let us start by defining the LDP and consider the accumulated experience of building models for estimating the probability of default based on them in particular and on data with unbalanced classes in general.

3. Defining the LDP segment

When speaking of LDPs, we should consider three groups of applied research papers:

1. **Non-financial:** Carvalho et al. (1999), Brown et al. (2001), Yelisseyeva and Yuzbashev (2002, p. 190, Section 7.7 'Small sample'), Kulinskaya et al. (2008), Lachin (2011), Song et al. (2011), Newby et al. (2013), Xue and Hall (2015);
2. **Financial:** Wehrspohn (2004), Benjamin et al. (2006), Keifer (2009), Dey et al. (2011), Pluto and Tasche (2006), Dzidzevičiūtė (2012), Surzhko (2014), Prorokowski (2016);
3. **Regulatory:** Basel Committee on Banking Supervision (2005, 2016), Financial Conduct Authority (2018).

Let us talk more specifically about the conclusions and recommendations made in these papers. The book by Yelisseyeva and Yuzbashev (2002, p. 190) says that estimates of statistical parameters for samples of less than 30 observations ('small sample') are unstable and require additional adjustments. In terms of the IRB approach, this can be interpreted in two ways.

On the one hand, there must be at least 30 borrowers ($n > 30$, where n is the number of observations) in any given segment for statistical models to be applicable. Similarly to Yelisseeva and Yuzbashev (2002), the paper by Brown et al. (2001) introduces the concept of a 'small sample' where the total number of observations is proposed to include 40 observations – that is, all borrowers in our case, not limited to defaulting borrowers. For comparison, Wehrspohn (2004) does not consider LDPs, but he claims that to validate models of probability of default there must be 250 observations. Such a 'small sample' requirement can be referred to for our paper purpose in the following way. One may expect that in case there are few observations (borrowers) within a segment (portfolio), then no matter the number of default cases, the model cannot be reliably estimated. That is, we may define LDP as a special dataset on which we are either unable to develop a model or model estimates are not efficient. Therefore, a segment with a few number of borrowers (less than 30) can be also classified as LDP.

On the other hand, the limit of 30 observations out of Yelisseyeva and Yuzbashev (2002) can also be interpreted to mean that, if there are fewer than 30 defaults

($n \cdot p \leq 30$, where p is the default rate), the portfolio should be called a LDP. This statement comes from the fact that given one class is fully unrepresented (i.e. there are zero defaults), we cannot estimate a model. Remember, when Yeliseeva and Yuzbashev (2002) and Brown et al. (2001) discuss ‘small sample’, they assume that each datapoint has, at least, two parameter values: one – for the independent X factor; another one – for the dependent Y factor. Implicitly they assume that the Y driver is a random variable with infinite or finite domain and all the Y values do not equal or are not statistically equal to one another. The more complicated situation arises when one deals with the Bernoulli random variable being a dependent variable. In such a situation the variability of Y is significantly reduced as it may only take values of one or zero. That is why we deduce that each of two classes should have sufficient number of observations to estimate the binary choice model. This second approach is more consistent with the world’s only explicit indication of a regulator (Financial Conduct Authority, 2018, par. 4.6.31) known to the author that a low default is a portfolio with up to 20 defaults. Now we can draw the following correspondence between Financial Conduct Authority (2018) definition and Brown et al. (2001) ‘small sample’ definition when extending the logic above described for the Yeliseeva and Yuzbashev (2002). In case each of two classes has 20 observations at least (the Financial Conduct Authority (2018) requirement), we avoid the ‘small sample’ situation, as the total set exceeds 40 observations (Brown et al. (2001) definition). In the works by Dzidzevičiūtė (2012, p. 153) and Pomazanov (2016, p. 67) as well as in Financial Conduct Authority (2018), it is suggested that a portfolio be considered low default if it includes fewer than 20 defaults. It can be said that this threshold as a definition of an LDP is also supported in the works by Keifer (2009), Tasche (2013), Kruger (2015), and Prorokowski (2016). However, while Keifer (2009, p. 169) gives a threshold of 10 defaults as one of the examples and claims that it is not unrealistic, Kruger (2015, p. 7) and Tasche (2013, p. 320) consider the example of data from Moody’s where the maximum number is 14 defaults for one year (in 2002 and 2008).

The English regulator’s requirement that there be at least 20 defaults for developing an IRB model is substantiated in the paper by Benjamin et al. (2006, pp. 24–25) where the author shows that with a smaller absolute number of defaults the estimated probability of default used for calculating the RWA should be adjusted. For example, for a model with a significance level of 75% (instead of the regulatory 99.9%) and a correlation level (R) of 12% with a probability of default estimated by the bank for one year equal to $PD = 1\%$, the calculation of RWAs according to Vasicek’s model should include a value of about $PD = 3\%$ and not the calculation result ($PD = 1\%$).

Interestingly, while Dzidzevičiūtė (2012, p. 153) recommends defining LDP by the number of defaults without taking into account the number of borrowers in the portfolio, in Kruger (2015, p. 1), on the contrary, a LDP refers to a portfolio with a small total number of borrowers, but the author does not state the exact threshold.

Even if there is an absolute criterion for defining a LDP, it may be of little use for numerous but low-value exposures (for example, in consumer lending or microfinance). In this case, it is expected that a relative criterion must be used to define the low default portfolio. Indirectly, this criterion can be found in the Basel Committee's guidelines (Basel Committee on Banking Supervision, 2016, p. 6) where the use of the lower threshold in the assessment of probability of default is suggested for the LDP segment (for example, at least 5% for corporate exposures). This means that a portfolio with a default rate of less than 5% of all borrowers may be considered a low default portfolio.

Table 1. Summary of proposals as to the maximum number of defaults to be included in a LDP

No	Source	Type of indication	Number of defaults	Share of defaults, %	Number of observations
1	Bull et al., 2002, p. 63	not explicit ²			10
2	Forrest, 2005, p. 5	explicit	10		
3	Keifer, 2009, p. 169	explicit	10		
4	Prorokowski, 2016, p. 235	explicit	10		
5	Pomazanov, 2016, p. 67	explicit	10–20		
6	Dzidzevičūtė, 2012, p. 143	explicit	20		no
7	Financial Conduct Authority, 2018, par. 4.6.31	explicit	20		
8	Cochran et al., 2006, p. 1194	not explicit	26		
9	Yelisseyeva and Yuzbashev, 2002, p. 190	not explicit	30		30
10	Heinze and Schemper, 2002, p. 2410	not explicit	30		
11	Deakin, 1972, p. 168	not explicit	32		
12	Altman, 1968, p. 599	not explicit	33		
13	Brown et al., 2001, p. 115	not explicit			40
14	King and Zeng, 2001, p. 138	not explicit		0.34	
15	Löffler et al., 2004, p. 5	not explicit		0.8	
16	Mogilat, 2019, p. 108	not explicit		0.8	
17	Schuermann and Hanson, 2004, p. 8	not explicit		1.66	
18	Sabato, 2006, pp. 15–16	not explicit		2.4–4.2	
19	Carvalho et al., 1999, p. 2	not explicit		5–7	
20	Song et al., 2011, p. 1587	explicit		10	
21	Kruger, 2015, p. 7	not explicit			+
22	Ohlson, 1980, p. 110	not explicit	105		
23	Wehrspohn, 2004, p. 7	not explicit			250
Average			28.4	2.7	82.5

Although the work by Kruger (2015) has already been mentioned as containing an example of an absolute criterion for defining an LDP, it can also be considered as a confirmation of the relative definition of the LDP segment. Though the maximum number of defaults in a given year does not exceed 14 for

² Bull et al. (2002) indicate that at least 10 observations for each model parameter are needed to obtain the unbiased estimates of parameters of a binary choice logistic model. Therefore, the minimum requirement will be 10 observations for one factor.

2000–2014, their total number is 55, while the total number of borrowers varies from 2,481 to 2,777 (Kruger, 2015, p. 7), which in relative terms is equivalent to a maximum default rate (DR; central tendency, CT) of 0.14%.

As an example, to understand the definition of a LDP, let us consider a work on forecasting the defaults of car companies in Italy (Micheli, 2015). The paper covers 20 observations, 10 of which are defaults, and considers four default factors. Thus, we believe it is an LDP study based on the criterion of a small minimum number of defaults. For example, Sabato (2006, p. 13) classified as LDP studies the following earlier research papers which were comparable by the number of defaults and did not explicitly introduce the concept of LDP: 33 in Altman (1968, p. 599); 32 in Deakin (1972, p. 168); and 105 in Ohlson (1980, p. 110).

It is, however, important that for the purpose of determining a LDP it is not the number of degrees of freedom that is taken into account (in this case $20 - 4 = 16$) but the number of defaults (10) and the total number of observations (20). The question of the statistical significance of factors in the LDP model remains open, but this must be considered separately outside the scope of this paper.

A summary of LDP definitions is provided in Table 1 where a special column with the type of indication makes it clear whether the author of the given paper explicitly sets such a threshold for LDPs or just mentions this value, and thus we interpret it implicitly.

4. Development of probability of default models for LDPs

When the developer of a binary choice model encounters clearly unbalanced classes in a sample of primary data, or a LDP in our case, there are several options:

- 1) Reduce the number of observations from the over-represented (non-default) class to make the ratio 50:50;
- 2) Increase the under-represented (default) class to equal proportions of 50:50;
- 3) Increase the number of occurrences viewed as defaults so that the share of the under-represented class grows as compared to the initial ratio but generally not to equal shares;
- 4) Keep the representation of classes unbalanced but apply adjustments to the estimation of the probability of observation of the under-represented class (the probability of default);
- 5) Combine the sample with another one to increase the share of the under-represented class.

Options 1 and 2 are common in the academic literature (Altman, 1968; Platt and Platt, 1990; McGurr and DeVaney, 1998; Newby et al., 2013). Their disadvantage is that by rebalancing them (removing over-represented observations or multiplying the under-represented ones) you can get distorted (biased) estimates of parameters of the PD prediction model. However, Newby et al. (2013) support this approach as in his view it increases the accuracy of the probability forecast.

A sub-type of option 2 is the adjustment of the maximum likelihood function to work with an equal ratio of classes, as, for example, in King and Zeng (2001) and Mogilat (2019). This approach can be useful to avoid biased estimates of model parameters (the determinants of probability of default). However, for the regulatory purpose of assessing the probability of default, this approach will be of little use.

It is worth mentioning the paper by Xue and Hall (2015, p. 1111), in which the authors prove that such a rebalancing increases the discriminatory ability of the model only if the data in classes are distributed according to the multidimensional Gaussian law. Otherwise, in the case of non-Gaussian distributions, which is the case in the overwhelming majority of applications, including the financial sphere, the results of rebalancing will fall short of expectations.

Then, option 3 should be considered where the number of defaults is increased. Here, the concept of semi-defaults (or quasi-defaults) can be introduced (Karminskiy, 2015; Lozinskaia et al., 2017), or a forecast of variable shocks, which according to the expectations of researchers may be associated with the probability of default, can be built. For example, such shocks may include loss of job or divorce for retail borrowers (Sabato, 2006). However, for the purpose of the IRB approach, instead of options 1–3, banks prefer the fourth option where the obtained estimate of probability of default is adjusted.

The easiest way within option 4 is the approach of Lachin (2011, p. 19). The upper estimate for the probability of default will be the value following from the definition of the binominal distribution:

$$PD = 1 - (\alpha)^{\frac{1}{n}},$$

where α is the level of significance at which the upper estimate of the probability of default is to be obtained; n is the total number of borrowers.

However, according to the author's observations, this approach is not widely used by banks in their risk management practices. On the one hand, it gives the most conservative PD estimate for the class that had no defaults, which is good for the regulator but not good for banks. On the other hand, the management of banks' risk management divisions may be unprepared to assign zero default probabilities for the purpose of IRB to exposures from a class which has not had any defaults.

Therefore, in their practice banks prefer the approaches from the papers Lando and Skodeberg (2002), Forrest (2005), Pluto and Tasche (2006) и Keifer (2009). Let us look at these methods in more detail.

In Lando and Skodeberg (2002) the authors propose evaluating the transition matrices (migration matrices) between the ratings and based on their extrapolation obtain estimates of 'transitions' to default status for highly rated portfolios, including LDPs. The authors did their research using a sample of almost 7,000 primarily American companies over the 18 years from 1981 to 1997. The PD forecast based on the transition matrices was later used in Schuermann and

Hanson (2004) where the class of defaults accounted for 1.66% of all observations, or 842 cases of non-repayment.

Forrest (2005) proposes performing modelling while taking expert opinion into account. This approach was later developed into a Bayesian adjustment in the papers by Keifer (2009) and Tasche (2013). The strong point in the research by Forrest (2005) is the substantiation of what cut-off level to take for the likelihood ratio (LR) depending on the number of defaults in the sample (see Figure 2). This cut-off level makes it possible to define the confidence interval for the assessment of the probability of default whose upper limit is recommended for the purpose of the IRB approach. The example provided by Figure 2 shows that for a smaller number of defaults the assessment of probability of default should be more conservative, or, which is effectively the same, the adjustment coefficient should be higher.

Figure 2. The cut-off levels for the likelihood ratio

Source: Forrest, 2005

Pluto and Tasche (2006) suggest first assessing the default rate in the rating scale categories where they were observed. On the basis of default ratios obtained (often averaged over time), a trend (extrapolation) is built for the categories which had no defaults. This part of the approach is reflected in a simplified manner at the beginning of this paper in Figure 1 for grade AAA.

Taking into account the deviations ('noise') arising when building the PD trend for the grades, the upper PD estimate is built for grades that had no defaults – that is, for the LDP segment. To build the estimate, the level of significance is selected. In the paper by Dey et al. (2011, p. 6–8), it is shown that for the significance level of 99.9% established in Basel II (Bank of Russia, 2015) the PD estimate for a low default portfolio – that is, a portfolio that had no defaults, can approach 1, or, more precisely, be equal to 81%, which is significantly higher than zero expected of a low default portfolio.

While in the paper by Pluto and Tasche (2006) it is suggested that the regulator set the level of significance, in Keifer (2009) and Tasche (2013) the authors suggest a Bayesian adjustment approach to assessing the probability of default, which according to the authors will not require choosing a significance level. It should

be noted that the first attempt to apply a Bayesian adjustment to the PD forecast was made in Löffler et al. (2004, p. 5), where the default rate was 0.8%.

The difference between Keifer (2009) and Tasche (2013) is that the former provides a non-reproducible method of work with one expert, chosen by the author, and the parameters of the Bayesian adjustment were selected taking into account the opinion of this expert. In Tasche (2013), the best format of such a Bayesian adjustment is chosen out of several approaches.

In Dzidzevičiūtė (2012), based on the example of data on loans issued to Lithuanian companies in 2005–2008, the author compares the accuracy of the predicted PD for LDP segments using the methods considered. Dzidzevičiūtė (2012) concludes that the optimum approach is the approach of Forrest (2005), which ensures a monotonous ranging of borrowers by the level of their creditworthiness, which is not accomplished in Pluto and Tasche (2006) and Keifer (2009), though Pluto and Tasche (2006) method is the easiest one for banks to implement. The method of Lando and Skodeberg (2002) as well as the approaches of Van der Burgt (2007) and Tasche (2009) according to Dzidzevičiūtė (2012) do not fall within the said 'top three methods' for the development of PD assessment models for a LDP. The conclusion about the advantages of methods may require adjustments if we consider the correlations of defaults excluded from Dzidzevičiūtė (2012).

Interestingly, while Pluto and Tasche (2006) proposes using the PD modelling methods for the entire rating system, in Dzidzevičiūtė (2012, p. 153) the author suggests using these methods only at the level of individual grades.

In Tasche (2013), it is also noted that if a standard method was used for choosing the upper limit of the confidence interval to assess the probability of default, then with a low (18%) correlation of assets (R), the optimum Bayesian adjustment corresponds to the significance level of 50–75%, while with a higher ($R = 24\%$) correlation it corresponds to a significance level of 75–90%. This is an important conclusion for practical LDP modelling, which suggests that if a LDP consists of retail credits (they are expected to have little correlation with the system factor), the conservative significance level will be 75%. If a LDP is a non-retail segment (class), which correlates more with the system risk factor, the conservative significance level must be 90%.

Pluto and Tasche (2006) additionally specify when an IRB model based on the LDP segment can be considered for regulatory purposes. Namely, the following condition must be met: when building an IRB model for the LDP segment, the bank expressly defines the segment – that is, establishes the level (rate or number) of defaults at which a different (standard) method of building the PD assessment model without an adjustment will be used.

A useful, intuitive but not officially established recommendation is given in Dzidzevičiūtė (2012, p. 153), which suggests that if a class (segment) becomes an LDP in one year only, this class (segment) should be treated as a LDP only in this year and not permanently.

In Surzhko (2014), the author does not touch upon the problems of building the IRB models in the LDP segment but only claims that using CDS statistics can

help obtain estimates of the probability of default for segments of exposures and rating scale categories which did not have any defaults. Although it is mentioned as an advantage of the method that the obtained estimates of probability of default vary in time, it is not verified in the research paper. This is only to be expected due to the lack of default statistics for such segments. It is, therefore, impossible to assess whether the proposed approach is correct or confirm the possibility of using this approach for regulatory purposes (Bank of Russia, 2015).

Among the research dedicated to LDP modelling, of special mention are papers by Wei and Yuan (2016) and Shi et al. (2017) that offer approaches to building limit distributions for the share of loss given default (LGD) and cumulative loan losses for the LDP, respectively. Their disadvantage is the lack of comparisons between the proposed approaches and other (now traditional) methods described in previous research and mentioned above. Therefore, the practical applicability of the conclusions in Wei and Yuan (2016) and Shi et al. (2017) is questionable.

Having considered the approaches to defining a LDP and the methods of building PD assessment models based on it, let us try to answer the questions raised at the beginning of this paper: can we assess a specific interest for banks in allocating a particular LDP, or is it just a historically established practice of having one?

5. An example of the economic effect of LDP allocation

As a starting point, let us consider the work by Kaltofen et al. (2006, p. 19), where the authors suggest that a more detailed segmentation enables a bank to benefit from a smaller capital requirement for the same portfolio of exposures (see Figure 3). For example, when moving from a common (zero) segmentation level to the third level a bank can reduce the risk weight by 18%, which is equivalent to an ability to increase the volume of loans issued without increasing its equity capital on a similar scale.

Figure 3. Benefit from the decrease in capital requirement for the bank given rise in segmentation granularity, %

Source: Kaltofen et al. (2006, p. 19)

It seems that the benefit in terms of weight risk and, as a result, the capital adequacy ratio can be obtained by higher segmentation. Then, the question arises as to whether a similar effect can be obtained by selecting a LDP.

To answer this question, let us consider a simple example. Let us assume we have two factors X_1 and X_2 for a single homogeneous class of one thousand exposures. These factors are associated with default status (either linearly or non-linearly). Their weighted sum with some 'noise'³ determines the value of the abstract (unobservable) indicator of the borrower's financial stability. By externally assigning a threshold for this indicator, we can determine the number of defaults that will occur. See Appendix for a more detailed description of the technical implementation of data generation.

Let us assume that a bank builds segmentation based on one of the factors. For example, on X_1 , which reflects the level of debt burden. Then, companies with a high debt burden will fall into one class, and those with a low burden into the other. The bank may wonder how many defaults to include in the LDP segment (for example, for companies with a low debt burden); build three models for assessing the probability of default (one based on a data array for the entire class (pooled) and one for each of the two sub-classes: LDP and class 2) and compare the cumulative capital requirements (the sum of the value of expected loss deductible from the capital in the numerator of the capital adequacy ratio and the value of unexpected losses in the denominator, which are equal to RWAs, divided by 12.5, see Appendix for details) for the cases of one model for a pooled class or two models for two classes.

Figure 4. The number of defaults in a class and bank's benefits, case of linear relationships

Note: 173, 309 and 458 are the total number of defaults in two classes; horizontally, the number of defaults in the first (LDP) class out of the specified number are plotted; NO WN = no 'noise' (random walk, white noise, WN) in the model; linear = a linear relationship between the X factor and the status of default. For more information on building a numerical example, see Appendix.

³ A special feature and to a certain extent a restriction in the validation of IRB models on the part of the regulator is that it confirms a model of adequate quality, where quality is understood as the excess of statistical indicators over the set threshold. It is logical to assume that a bank can have arbitrage (whether knowingly or not) from having a model of sufficient but not perfect quality. For example, the AUC for the ROC equals 40%, which is acceptable (Pomazanov, 2016, p. 54, Table 2.6), but not 100%, which is ideal. In this case, the assessment of capital requirements may not just be distorted but may be reduced to the benefit of the bank. For more information on how to identify such problems and rectify the identified areas of potential arbitrage, see Ermolova et al. (2019).

Figure 5. The number of defaults in a class and bank's benefits, case of non-linear relationships

Note: 173, 312 and 458 are the total number of defaults in two classes; horizontally, the number of defaults in the first (LDP) class out of the specified number are plotted; NO WN = no 'noise' in the model (white noise, WN); LN = a non-linear (logarithmic) relationship between the X factors and the status of default. For more information on building a numerical example, see Appendix.

Below, we will outline interesting conclusions. In particular:

- 1) The fewer defaults are included in the LDP segment, the higher the benefit for the bank in terms of cumulative capital requirements (see Figure 4, Figure 5).
- 2) The benefit in terms of capital release can reach 30%, if only 5 defaults are included in the case of a linear relationship between the factors and the status of default (see Figure 4).
 - The above positive effect (benefit) for the bank from distinguishing a LDP is not obvious as the author would expect that the reduction in the risk estimates in one (LDP) segment will be compensated by the increase in the risk estimates in the other (second, non-LDP) segment. However, as it will be shown below, this effect of the mutual compensation of risk estimates is only observed in the second class (non-LDP). Therefore, although the above benefits may seem obvious to the readers, they have not been demonstrated before. Thus, the aforementioned paper of Kaltofen et al. (2006) contains an example of benefits from the identification of clusters, but, unlike this work, it does not provide the effect of the inclusion of a various number of defaults in one of the classes.
- 3) If the relationship between X factors and the state of default is linear, the benefit for the bank is higher (compare Figure 4 for a linear relationship and Figure 5 for a non-linear relationship).
- 4) The more defaults a class contains, the less the benefit of LDP allocation is (see Figure 4, Figure 5).
 - Initially, we set the level of defaults for the entire portfolio (the entire sample, for the sum of defaults in two classes) as 17%. Then, we

considered levels about two and three times higher – namely, 31% and 46%, which correspond to approximately 170, 310 and 460 defaults in total (out of 1,000 observations). These figures demonstrate the effect of allocating 5 to 173 defaults out of the total indicated number into a LDP segment. Thus, it is 5, 10, etc defaults that are a sign of a LDP, as we have seen in Section 4 above, and not the total default rate in both segments (naturally, for 1,000 observations at minimum 170 defaults are an unbalanced class but not a LDP).

- 5) The ‘noisier’ is the model, the lower is the benefit to the bank (see Figure 6, Figure 7).

Figure 6. The number of defaults in a class, bank’s benefits and noisiness of the linear model

Note: 173 = the total number of defaults in two classes; NO WN = no ‘noise’ in the model (white noise, WN); linear = linear relationship between the X factors and the status of default; 10 and 20 = the multipliers increasing the ‘noise’ effect. For more information on building a numerical example, see Appendix.

Figure 7. The number of defaults in a class, bank’s benefits and noisiness of the non-linear model

Note: 173 = the total number of defaults in two classes; NO WN = no ‘noise’ in the model (white noise, WN); LN = non-linear (logarithmic) relationship between the X factors and the status of default; 10 and 100 = the multipliers increasing the ‘noise’ effect. For more information on building a numerical example, see Appendix.

After considering the cases where allocating a LDP benefits the bank, the question arises of why it happens. For that, let us consider the most extreme case, when there is a linear dependence of factors and no ‘noise’, the LDP segment has 10 defaults, and the bank’s total benefit is 30% of the total capital requirements (expected loss (EL) and unexpected loss (UL)).

First, we see in Figure 8 that the allocation of two classes will predictably lead to a distortion in the estimates of the probability of default. However, while in an ordinary (non-LDP) segment some estimates become lower, and others increase proportionately (see the Z-shaped orange line in Figure 8), the estimates of the probability of default for the LDP segment are significantly understated. Let us describe Figure 8 in more detail. The same descriptive logic is then applicable to Figure 9. Each dot on the graph stands for a borrower from our hypothetical example (remember the total number of borrowers equals to 1,000). A part of borrowers is assigned to class 1 (LDP; blue dots); others – to class 2 (non-LDP; orange dots). Black line is the bisector, introduced for visual comfort to easier identify the areas where the PD values are underestimated or overestimated. In the current example we see that the pooled PD estimated falls within [0%; 80%] range. However, when two classes are subdivided, we see the estimates deviation. For low risk borrowers (see the lower left segment of the chart) the PD estimates approach to or are equal to zero. Interestingly, the PD estimates for LDP borrowers (blue dots) happen to be higher than those for non-LDP ones (orange dots), though prior to segmentation both had equally low PDs. At the same time, for the non-LDP borrowers (orange dots) we notice that PDs are both underestimated for low risk ones and overestimated for high risk ones. However, for the LDP-borrowers (blue dots) we see only the PD underestimation. This fact to a large extent contributes to the capital gain (regulatory arbitrage) arising from LDP segmentation for a bank.

In this case, we do not adjust the estimates according to the above approaches to demonstrate to our readers the ‘clean’ effect – that is, without adjustments. A natural step expected of the regulator in this case, similarly to Benjamin et al. (2006), would be to require an increase in the estimates of the probability in the LDP segment (the blue line in Figure 8) to the level of the estimated probability of default in the case of combined data (the black line). Thus, instead of the value of $PD = 1\%$ on the horizontal axis (2 classes) (it corresponds to PD estimate for LDP (blue) borrowers originating from LDP-based model), the regulator should require the use of the estimate of 2.8% (this is the PD estimate on the vertical axis (Pooled) corresponding to that very same LDP borrowers if PD was taken from the pooled model), which is comparable to the calculations of Benjamin et al. (2006) resulting in about 3%. In other words, one may notice that looking at the horizontal axis the A point corresponding to PD estimated from the pooled model is to the right from the B point standing for the respective estimate from the LDP-based model.

Due to the non-linear nature of capital requirement dependence on the probability of default, the effect on their changes when a LDP segment is selected is more evident for them (see Figure 9) than the effect on PD estimates (Figure 8). As can be seen from Figure 8, capital requirements for exposures with a substantially higher PD estimate in the second non-LDP class grow insignificantly. The reason for this is Vasicek's model, according to which the maximum risk weight (unexpected loss) is achieved with the PD value of 30%. Therefore, if the probability of default is significantly higher, the aggregate capital requirements approach the estimated expected loss. Thus, the differentiation of PD estimates obtained as a result of the segmentation leads to a reduction in capital requirements practically for each exposure and to the overall benefit of the bank, as stated above.

Then, the following logical question arises: What is the reverse side of such benefits for the bank, following simply from the optimised solution of the classification (clustering) problem? In Table 2 the author provides the estimated ratios of three simplified models examined for the above case where the LDP has 10 defaults. For simplicity, the estimates used ordinary least squares (OLS). Now, we are interested not so much in the values of the estimated model parameters since we have seen the result of their distortion above in Figure 8 in the differentiation of resulting PD estimates. Of more importance are the significance characteristics of these parameters (coefficients of models), as can be seen by the t-statistics indicator, and the discriminatory ability, which is measured by the AUC (area under the cumulative accuracy profile (CAP) curve).

Figure 8. Comparison of the probability of default forecasts at the level of exposures, %

Note: in all axes, PD values are set for individual exposures for a single model (Pooled; vertically) and for two models (2 classes; horizontally); 1:1 = a bisector which, for illustration purposes, shows the ratio of equal PD estimates. For more information on building a numerical example, see Appendix.

Figure 9. Comparison of the aggregate capital requirements (conventional monetary units) at the level of exposures

Note: aggregate capital requirements (EL+UL) are set along all axes for individual exposures for a single model (Pooled; vertically) and for two models (2 classes; horizontally); 1:1 = a bisector which, for illustration purposes, shows the ratio of equal estimates of capital requirements. For more information on building a numerical example, see Appendix.

Table 2. Estimation of parameters of PD linear assessment models
(see Appendix, Section 2)

Pooled	Coefficient <i>b</i>	s.e.	t-statistics	P-value	AUC		df
Intercept	0.16	0.02	6.90	0.0000000		Regression	2
X1	0.65	0.03	21.23	0.0000000	87.46%	Remainder	997
X2	-0.63	0.03	-20.39	0.0000000	13.76%	Total	999

LDP	Coefficient <i>b</i>	s.e.	t-statistics	P-value	AUC		df
Intercept	0.02	0.01	1.16	0.2484039		Regression	2
X1	0.16	0.03	4.99	0.0000008	54.52%	Remainder	578
X2	-0.09	0.02	-4.98	0.0000009	2.43%	Total	580

Class 2	Coefficient <i>b</i>	s.e.	t-statistics	P-value	AUC		df
Intercept	0.20	0.09	2.30	0.0220595		Regression	2
X1	1.09	0.11	10.00	0.0000000	19.42%	Remainder	416
X2	-1.34	0.05	-29.61	0.0000000	1.09%	Total	418

Note: coefficient *b* is the estimates of the model coefficients; s.e. is the standard error of the coefficients estimate; t-statistics = coefficient *b*/s.e. is the indicator of the coefficient's significance; P-value is the probability of assuming a zero hypothesis that coefficient *b* of the model is equal to zero (the lower the value, the better); AUC is the indicator of discriminatory ability (the higher, the better); df is the number of degrees of freedom for the estimated model, where Regression is the number of factors in the model (everywhere 2, apart from the constant); Total is the total number of observations minus one (corresponds to the constant); Remainder is the number of degrees of freedom itself (the higher, the better).

It can be seen that the coefficients for both factors X1 and X2 are statistically significant in all models, although the X2 factor has insufficient discriminatory ability (AUC is less than 14% in all models; it is desirable to have at least 67.5%, which is equivalent to the accuracy ratio AR = 35%; see Pomazanov, 2016, p. 54,

Table 2.6). However, it is important that with selected LDP segment the significance and discriminatory ability of the main factor X_1 dropped in both sub-class models as compared to the pooled model. At the same time, the significance of the X_2 factor with the worst discriminatory ability increased in the model for the second (non-LDP) class. We can draw a key conclusion that the adjustment of PD estimates alone is insufficient as it does not contribute to the improvement of the statistical characteristics of the models built. Therefore, the regulator should require that the bank combines data and builds the model using the largest possible array of data in terms of the number of defaults, which was mentioned in the first recommendations of the Basel Committee (Basel Committee on Banking Supervision, 2005) and in further research of Wei and Yuan (2016, p. 123).

To test the adequacy of the above approach, let us compare the estimates made using the OLS and Probit methods for two basic cases: the linear and non-linear model with the level of defaults of 17% across the entire sample with an added error ($10 * WN$). Without the latter, the Probit model becomes fully defined and the PD estimates degenerate into 0 and 1.

Let us explain why we did not choose the Probit model initially. When validating IRB models the regulator does not have the goal of building the best model in place of the bank. The regulator's task is to confirm that the model developed by the bank is of adequate quality. The above case shows that in all variations of the model estimated using the OLS method the coefficients are statistically significant, and in many cases they may have a sufficient discriminatory force. Therefore, we cannot exclude that the bank may propose such a model. In fact, as may be seen from Table 3, the bank's benefit from using the Probit model is lower than that from the OLS. Therefore, the bank may be interested in using a model that would give it a higher benefit, although generally lower capital requirements are set for the Probit model across the entire portfolio consisting of two classes. However, as shown below, in addition, applying the Probit model understates the PD estimate in the LDP segment. If we assume that the LDP may have exposures which are substantially higher in absolute terms, the benefit from the capital requirements will increase.

Table 3. Comparison of average capital requirements for different types of data and models

No		Linear	Non-linear
1	L_ols_pool	21.5	21.9
2	L_ols_model	20.0	19.7
3	L_probit_pool	11.3	17.4
4	L_probit_model	11.0	17.0
5	Benefit_ols = ((1) - (2))/(1)*100	7.0%	10.0%
6	Benefit_probit = ((3) - (4))/(3)*100	2.7%	2.3%

If we go into detailed comparison of the OLS and Probit models for the above two cases, Figure 10 shows the following:

Figure 10. The ratio of capital requirements (Losses) as per the OLS and Probit

Note: L (Losses) is the capital requirements; model = with a separate estimation of models for two classes; light blue dots = class 1 (LDP); violet dots = class 2 (ordinary); in graphs a–d horizontally = the values of capital requirements in the event of separate estimation of models for two classes; vertically = estimation based on a single (Pooled) model; in graphs e–f horizontally = the estimation of capital requirements according to Probit, vertically = according to OLS; both estimates were obtained for a sample divided into two classes: LDP and ordinary. The yellow line in the graphs is a linear trend between all the dots, the blue line is a quadratic trend; graphs g and h show the PD ratio for OLS (vertically) and for Probit (horizontally) for estimation on two classes.

- 1) When applying the OLS to the LDP class, the estimates of capital requirements (Losses) grow for borrowers with small exposures (TC_i , see Appendix, Section 3) (the light blue dotted line in Figures 10a and 10b on the bottom left goes horizontally, then grows).
- 2) The use of the Probit model reduces, albeit insignificantly, the PD estimates for the LDP class (the light blue dotted line in Figures 10c and 10d grows at the beginning, being above the bisector).
- 3) The estimates of capital requirements using the Probit model is lower for the LDP class (light blue dots) than with the OLS when models are estimated separately for two classes, Figures 10e and 10f. This is the result of the underestimated probability of default for the LDP class using Probit as compared to the OLS (see Figures 10g and 10h).

Above, we have illustrated an example that would allow a bank to achieve capital release when selecting an LDP segment and apply models separately. This example allows us to draw two important conclusions. Firstly, we cannot exclude that the bank may end up in a similar scenario. Secondly, this example contains a more important conclusion than just a specific value of capital gains. For example, the above example shows that when a bank presents a model based on the LDP segment to the regulator, the bank most commonly has a non-LDP segment which is homogeneous (somehow related) to the presented LDP segment. Thus, the regulator can assess the bank's benefits each time in a given case by re-estimating the model based on a pooled data (or instruct the bank to do so). In this case, the regulator will receive information on the extent of the bank's benefit. If it is substantially positive, the regulator should require that the bank pool the data, adjust the model based on these data and subsequently use this model in its processes.

6. Conclusion

We have shown above that difficulties of building binary choice models, in particular PD forecasts, arise in the case of unbalanced classes, especially with a significantly low share of one class. In finance, this is called a LDP. According to previous research, this portfolio should be defined if one of the following conditions is met (see the summary in the last line in Table 1):

- 1) the number of defaults is less than 25;
- 2) the default rate is less than 3% of total observations.

There are different approaches to addressing unbalanced classes. They include rebalancing and attempts to form additional statuses of defaults (semi-defaults or quasi-defaults), adjust the estimate of PD forecasts and combine data samples. Despite the feasibility of such approaches, for IRB purposes, the regulator has good reasons for not allowing banks to build models based on LDP segments, and if such models are built, it should require that banks rebuild them using an extended array

of combined segments, provided the following condition is met. If the regulator identifies a significant (e.g. in excess of plus 5% of the total capital requirements) benefit in the amount of capital arbitrage when comparing the capital requirements obtained from the assessment across the entire portfolio and after dividing it into classes, the regulator should require that the data are combined and the single model is estimated. For example, the benefit of 2.3–2.7% (see line 6 in Table 3) will not be material, and two models can be allowed.

By requiring banks to pool the data without distinguishing a LDP segment, the regulator will eventually obtain reliable models from banks, which will generally contribute to the improvement of financial stability in the banking system. As an objection, readers may refer to Mogilat (2019, pp. 108–109), where the sample, generalised to include all Russian companies, also has a 0.8% default ratio for a total of 350,000 to 1,000,000 observations. In this regard, we should mention a remark from the above cited paper (McCullah and Nelder, 1989, p. 120) where the authors claim that with sufficiently large samples, in terms of observations included, the estimates of binary choice model parameters will be sufficiently accurate.

An additional argument against pooling data may be the appearance of significant heterogeneity – namely, borrowers from industries with different risk determinants. It is worth remembering this effect; however, we can expect that it will not have significant distorting impact for two reasons. Firstly, when applying the IRB approach, banks must distinguish homogeneous classes of assets: for example, corporate loans, loans to sovereign borrowers, loans to financial institutions, retail loans and special-purpose lending (leasing, project financing, commodity financing, etc., see Bank of Russia, 2015, chapter 2). Within these classes, they can identify more segments to be used to assess models. Therefore, the regulator should separately verify the homogeneity of exposures.

Secondly, we can refer to the paper by Mogilat (2019, p. 110) as an example. Despite the omission of the above classes of assets, the author estimated one model based on a sample from substantially different industries (mining, electricity, manufacturing). It is important that as a result significant coefficients were obtained for the determinants of high probability of default (loan non-performance). Therefore, the expectation of heterogeneity may turn out not to be justified. A separate follow-up research can be done here (though it falls out of the current paper scope). One can compare change in coefficient estimates accuracy after breaking the set into separate segments (e.g. industry-wise) and change in the model's goodness-of-forecast. We cannot exclude the possible finding that the increase in the former might be associated with the decrease in the latter.

To conclude, let us examine two additional considerations to be taken into account and to think about in the future when dealing with LDP. First, it may seem that the question of allocating and modelling a LDP segment becomes irrelevant with the introduction of IFRS 9 (see Farrakhov, 2019), which regulates cash flows. However, if there are no or few defaults (non-payments) in a class

(segment, portfolio or product), a question similar to LDP-related issues arises regarding the accuracy of cash flow forecasts that will be built.

Second, the ambiguity or impossibility of correct PD assessments for LDPs may be the result of a high correlation of defaults. Let us consider the simplest case of 100% correlation of defaults. This means that defaults either occur simultaneously or do not occur at all. Then, the default rate in the sample will always be 0 or 100%. For example, Figure 11 shows two peaks on the default ratio (DR) distribution density graph, where the left peak (mode) for the case of no defaults has the frequency $(1 - PD)$, and the right peak for the case of defaults has the frequency PD . It follows from the properties of distribution of binary random variables that the average ratio of defaults will always be equal to the probability of defaults, as $DR = PD$.

Figure 11. LDP with a 100% default correlation

The above example shows that a low default portfolio may occur for any PD values, if due to its low depth, the sample has not yet accumulated observations corresponding to the right peak of DR distribution density. This problem is especially relevant for banks, which submit applications for the use of the IRB approach for the first time and have a minimum depth of data of five to seven years. For example, from the paper by Gordy and Heitfield (2010, pp. 44, 53) we can conclude that it is expedient to have historical data for at least 20 years.

Appendix is available at
[http://rjmf.econs.online/en;](http://rjmf.econs.online/en;dx.doi.org/10.31477/rjmf.202002.101)
dx.doi.org/10.31477/rjmf.202002.101

References

- Altman, E. I.** (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23(4), pp. 589–609.
- Bakbergenuly, I., Kulinskaya, E. and Morgenthaler, S.** (2016). Inference for Binomial Probability Based on Dependent Bernoulli Random Variables with Applications to Meta-Analysis and Group Level Studies. *Biometrical Journal*, 4(58), pp. 896–914.

- Bank of Russia** (2015). *Regulation on the Calculation of Risk Ratios Based on Internal Ratings*, N 483-P, 6 August.
- Basel Committee on Banking Supervision** (2005). *Validation of Low Default Portfolios in the Basel II Framework*. Available at: http://www.bis.org/publ/bcbs_n16.htm [accessed on 4 March 2020].
- Basel Committee on Banking Supervision** (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework – Comprehensive Version*. Available at: <http://bis.org/publ/bcbs128.htm> [accessed on 12 December 2018].
- Basel Committee on Banking Supervision** (2016). *Reducing Variation in Credit Risk-Weighted Assets – Constraints on the Use of Internal Model Approaches*. Available at: <https://www.bis.org/bcbs/publ/d362.pdf> [accessed on 4 March 2020].
- Benjamin, N., Cathcart, A. and Ryan, K.** (2006). *Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities*. UK: Financial Services Authority. Available at: https://webarchive.nationalarchives.gov.uk/20081231200612/http://www.fsa.gov.uk/pubs/international/default_probabilities.pdf [accessed on 4 March 2020].
- Brown, L., Cai, T. and DasGupta, A.** (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2), pp. 101–133.
- Bull, S., Mak, C. and Greenwood, C.** (2002). A Modified Score Function Estimator for Multinomial Logistic Regression in Small Samples. *Computational Statistics and Data Analysis*, 39(1), pp. 57–74. doi: 10.1016/S0167-9473(01)00048-2
- Carvalho, D. R., Ávila, B. C. and Freitas, A. A.** (1999). *A Hybrid Genetic Algorithm / Decision Tree Approach for Coping with Unbalanced Classes*. Available at: <https://sci2s.ugr.es/keel/pdf/specific/congreso/padd99.pdf> [accessed on 4 March 2020].
- Cochran, J. J., Darrat, A. F. and Elkhail, K.** (2006). On the Bankruptcy of Internet Companies: An Empirical Inquiry. *Journal of Business Research*, 59(10–11), pp. 1193–1200. doi: 10.1016/j.jbusres.2005.12.006
- Deakin, E. B.** (1972). A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, 10(1), pp. 167–179.
- Dey, A., Peterson, B. G. and Yadav, G.** (2011). *Modeling Low Default Credit Portfolio in R*. KPMG. Available at: <http://past.rinfinance.com/agenda/2011/Yadav+et+al.pdf> [accessed on 4 March 2020].
- Dzidzevičiūtė, L.** (2012). Estimation of Default Probability for Low Default Portfolios. *Ekonomika*, 91(1), pp. 132–156.
- Ermolova, M., Penikas, H. and Polyanskiy, Y.** (2019). Risk-Weighted Asset Add-On for Model Risk under the Internal Ratings-Based Approach. *Upravleniye Finansovymi Riskami*, 1, pp. 32–51. [In Russian].
- Farrakhov, I.** (2019). IFRS 9: Estimate of Expected Credit Losses ‘Without Excessive Costs or Effort’. *Risk-Menedzhment v Kreditnoy Organizatsii*, 2(34), pp. 4–13. [In Russian].
- Financial Conduct Authority** (2018). *Prudential Sourcebook for Investment Firms*. Available at: <https://www.handbook.fca.org.uk/handbook/IFPRU.pdf> [accessed on 4 March 2020].

- Forrest, A.** (2005). *Likelihood Approaches to Low Default Portfolio*. Available at: <https://crc.business-school.ed.ac.uk/wp-content/uploads/sites/55/2017/03/forest-alan-1.pdf> [accessed on 13 September 2019].
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A.** (2018). *Predictably Unequal? The Effects of Machine Learning on Credit Markets*. Available at: <https://ifin.usi.ch/assets/docs/seminars/a-fuster-338054.pdf> [accessed on 7 January 2020].
- Gordy, M. B. and Heitfield, E.** (2010). Small-Sample Estimation of Models of Portfolio Credit Risk. In: M. Kijima, C. Hara, Y. Muromachi and K. Tanaka, eds., *Recent Advances in Financial Engineering: The Proceedings of the KIER-TMU International Workshop on Financial Engineering 2009*. New Jersey: World Scientific, pp. 43–63.
- Heinze, G. and Schemper, M.** (2002). A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, 21(16), pp. 2409–2419.
- Kaltofen, D., Paul, S. and Stein, S.** (2006). *Retail Loans and Basel II. Using Portfolio Segmentation to Reduce Capital Requirements*. ECRI Research Report, N 8. Available at: <https://www.ceps.eu/wp-content/uploads/2009/08/1368.pdf> [accessed on 4 March 2020].
- Karminskiy, A.** (2015). *Credit Ratings and Their Modelling*. Moscow: HSE Publishing House.
- Keifer, N. M.** (2009). Default Estimation for Low default Portfolios. *Journal of Empirical Finance*, 16(1), pp. 164–173. doi: 10.1016/j.jempfin.2008.03.004
- King, G. and Zeng, L.** (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), pp. 137–163.
- Kruger, M.** (2015). Validation and Monitoring of PD Models for Low Default Portfolios Using PROC MCMC. In: *SAS Global Forum 2015*. Available at: <https://support.sas.com/resources/papers/proceedings15/1400-2015.pdf> [accessed on 4 March 2020].
- Kulinskaya, E., Morgenthaler, S. and Staudte, R. G.** (2008). *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. Chichester, UK: John Wiley and Sons.
- Lachin, J. M.** (2011). *Biostatistical Methods: The Assessment of Relative Risks*. 2nd ed. New Jersey: John Wiley and Sons.
- Lando, D. and Skodeberg, T.** (2002). Analyzing Rating Transitions and Rating Drift with Continuous Observations. *Journal of Banking and Finance*, 26(2–3), pp. 423–444.
- Löffler, G., Posch, P. N. and Schoene, C.** (2004). *Bayesian Methods for Improving Credit Scoring Models*. University Library of Munich, EconWPA: Finance, N 0505024. Available at: <https://econwpa.ub.uni-muenchen.de/econ-wp/fin/papers/0505/0505024.pdf> [accessed on 4 March 2020].
- Lopez, J. A.** (2004). The Empirical Relationship between Average Asset Correlation, Firm Probability of Default, and Asset Size. *Journal of Financial Intermediation*, 13(2), pp. 265–283. doi: 10.1016/S1042-9573(03)00045-7
- Lozinskaia, A., Merikas, A., Merika, A. and Penikas, H.** (2017). Determinants of the Probability of Default: The Case of the Internationally Listed Shipping Corporations. *Maritime Policy and Management*, 44(7), pp. 837–858. doi: 10.1080/03088839.2017.1345018
- McCullah, P. and Nelder, J.** (1989). *Generalized Linear Models*. London: Chapman and Hall.

- McGurr, P. T. and DeVaney, S. A.** (1998). Predicting Business Failure of Retail Firms: An Analysis Using Mixed Industry Models. *Journal of Business Research*, 43(3), pp. 169–176. doi: 10.1016/S0148-2963(97)00222-1
- Micheli, A. P.** (2015). The Default Risk: An Empirical Analysis on the Automotive Companies. *Management Studies*, 3(7–8), pp. 169–178. doi: 10.17265/2328-2185/2015.0708.001
- Mogilat, A.** (2019). Modelling Financial Distress of Russian Industrial Companies, or What Bankruptcy Analysis Can Tell. *Voprosy Ekonomiki*, 3, pp. 101–118. [In Russian]. doi: 10.32609/0042-8736-2019-3-101-118
- Newby, D., Freitas, A. A. and Ghafourian, T.** (2013). Coping with Unbalanced Class Data Sets in Oral Absorption Models. *Journal of Chemical Information Modeling*, 53(2), pp. 461–474. doi:10.1021/ci300348u
- Ohlson, J. A.** (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), pp. 109–131.
- Ozdemir, B. and Miu, P.** (2009). *Basel II Implementation: A Guide to Developing and Validating a Compliant Internal Risk Rating System*. New York: McGraw-Hill.
- Platt, H. and Platt, M.** (1990). Development of a Class of Stable Predictive Variables: The Case of Bankruptcy Prediction. *Journal of Business Finance and Accounting*, 17(1), pp. 31–51. doi: 10.1111/j.1468-5957.1990.tb00548.x
- Pluto, K. and Tasche, D.** (2006). Estimating Probabilities of Default for Low Default Portfolios. In: B. Engelmann and R. Rauhmeier, eds., *The Basel II Risk Parameters*. Berlin: Springer. doi: 10.1007/3-540-33087-9_5
- Pomazanov, M.** (2016). *Credit Risk Management in a Bank: The Internal Ratings-Based (IRB) Approach*, 1st ed. Moscow: Urait.
- Prorokowski, L.** (2016). Rank-Order Statistics for Validating Discriminative Power of Credit Risk Models. *Bank i Kredyt*, 47(3), pp. 227–250.
- Sabato, G.** (2006). *Managing Credit Risk for Retail Low Default Portfolios*. Available at: <https://ssrn.com/abstract=908373> [accessed on 4 March 2020].
- Schuermann, T. and Hanson, S.** (2004). *Estimating Probabilities of Default*. Federal Reserve Bank of New York Staff Report, N 190.
- Shi, X., Tang, Q. and Yuan, Z.** (2017). A Limit Distribution of Credit Portfolio Losses with Low Default Probabilities. *Insurance: Mathematics and Economics*, 73, pp. 156–167.
- Song, J., Lu, X., Liu, M. and Wu, X.** (2011). Stratified Normalization LogitBoost for Two-Class Unbalanced Data Classification. *Communications in Statistics – Simulation and Computation*, 40(10), pp. 1587–1593. doi: 10.1080/03610918.2011.589332
- Surzhko, D.** (2014). *Approach for Multi-Period PD Calibration for LDP Portfolios*. Available at: <http://vixra.org/pdf/1404.0001v1.pdf> [accessed on 21 November 2018]
- Tasche, D.** (2009). *Estimating Discriminatory Power and PD Curves when the Number of Defaults is Small*. arXiv:0905.3928. Available at: <https://arxiv.org/details/arxiv-0905.3928> [accessed on 13 September 2019].
- Tasche, D.** (2013). Bayesian Estimation of Probabilities of Default for Low Default Portfolios. *Journal of Risk Management in Financial Institutions*, 6(3), pp. 302–326.

- Van der Burgt, M. J.** (2007). Calibrating Low default Portfolios, Using the Cumulative Accuracy Profile. *Journal of Risk Model Validation*, 1(4), pp. 17–33.
doi: 10.21314/jrmv.2008.016
- Vasicek, O.** (1987). *Probability of Loss on Loan Portfolio*. San Francisco: KMV Corporation.
- Wehrspohn, U.** (2004). *Optimal Simultaneous Validation Tests of Default Probabilities, Dependencies, and Credit Risk Models*. Available at: <https://ssrn.com/abstract=591961> [accessed on 4 March 2020].
- Wei, L. and Yuan, Z.** (2016). The Loss Given Default of a Low default Portfolio with Weak Contagion. *Insurance: Mathematics and Economics*, 66, pp. 113–123.
doi: 10.1016/j.insmatheco.2015.10.005
- Xue, J.-H. and Hall, P.** (2015). Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5), pp. 1109–1112. doi: 10.1109/TPAMI.2014.2359660
- Yeliseyeva, I. and Yuzbashev, M.** (2002). *General Theory of Statistics*, 4th ed. Moscow: Finance and Statistics.
- Zaigraev, A. and Kaniovski, S.** (2013). A Note on the Probability of at Least k Successes in n Correlated Binary Trials. *Operations Research Letters*, 41(1), pp. 116–120.
doi: 10.1016/j.orl.2012.10.002