Journal of the Operational Research Society (2014) 65, 416–434

www.palgrave-journals.com/jors/

Modelling credit risk with scarce default data: on the suitability of cooperative bootstrapped strategies for small low-default portfolios

Raquel Florez-Lopez* and Juan Manuel Ramon-Jeronimo
University Pablo Olavide of Seville, Seville, Spain

Credit risk models are commonly based on large internal data sets to produce reliable estimates of the probability of default (PD) that should be validated with time. However, in the real world, a substantial portion of the exposures is included in low-default portfolios (LDPs) in which the number of defaulted loans is usually much lower than the number of non-default observations. Modelling of these imbalanced data sets is particularly problematic with small portfolios in which the absence of information increases the specification error. Sovereigns, banks, or specialised retail exposures are recent examples of post-crisis portfolios with insufficient data for PD estimates, which require specific tools for risk quantification and validation. This paper explores the suitability of cooperative strategies for managing such scarce LDPs. In addition to the use of statistical and machine-learning classifiers, this paper explores the suitability of cooperative models and bootstrapping strategies for default prediction and multi-grade PD setting using two real-world credit consumer data sets. The performance is assessed in terms of out-of-sample and out-of-time discriminatory power, PD calibration, and stability. The results indicate that combinational approaches based on correlation-adjusted strategies are promising techniques for managing sparse LDPs and providing accurate and well-calibrated credit risk estimates.

Journal of the Operational Research Society (2014) **65,** 416–434. doi:10.1057/jors.2013.119
Published online 16 October 2013

Keywords: risk; credit scoring; small data sets; cooperative models; class imbalance; specification error

1. Introduction

Heavy losses associated with the recent world financial crisis have raised interest in the models used to manage credit risks and have demanded new solutions to address the deteriorated portfolios that were previously considered as low-risk. Credit-scoring models (Thomas *et al*, 2001), which are based on classification of credit risk applicants into two classes (good and bad payers), must be adapted to manage portfolios that are characterised by both a reduced number of observations and a still limited but largely unexpected number of defaults.

Under the pre-crisis Basel II regulation, banks previously developed internal scoring models to estimate the probability of default (PD) by considering large historical data sets (BCBS, 2004, paragraph 449). The estimated PD should be regularly compared with actual default rates to demonstrate that the occurrences are within the expected range for each grade (BCBS, 2004). A sufficient number of observations should be used to guarantee meaningful PD quantification and validation at each grade level. These requirements made it difficult to extend internal credit risk modelling to portfolios with limited default experience (so-called 'low-default portfolios' or LDPs) (BCBS, 2005a).

The LDPs include different categories of portfolios with limited default experience (Benjamin *et al*, 2006), including portfolios that historically have experienced a low number of defaults and are generally considered as low-risk (eg, banks, sovereigns, insurers, highly rated corporations), portfolios with a low number of counterparties (eg, train operating companies, niche markets, public–private partnerships), portfolios with a lack of historical data (eg, a new entrant into a market), and portfolios that may have not incurred recent losses but that large historical experience suggest a greater PD than is captured in recent data (eg, retail mortgages in a number of jurisdictions). Although limited attempts have been made to obtain robust PD estimates in the presence of LDPs, the empirical work is constrained to portfolios with few recent losses but with large historical experience (Sabato, 2006; Van der Burgt, 2008). The presence of such large LDPs has permitted to address them using simple techniques based on assumptions on PD distributions, including confidence intervals (Forrest, 2005; FSA, 2005; Pluto and Tasche, 2005, 2006; Schuermann and Hanson, 2005), *a priori* PD averages (Schuermann and Hanson, 2005; Benjamin *et al*, 2006) or *ad hoc* prior information (Löffler *et al*, 2004).

However, the world financial tsunami initiated in 2007 has shown that the nature of LDPs should be revisited, so that most

**Correspondence: Raquel Florez-Lopez, Department of Accounting and Finance, University Pablo Olavide of Seville, Ctra. Utrera km. 1, 41013 Seville, Spain.*
E-mail: rflorez@upo.es

robust and sophisticated techniques must be developed to prevent major financial losses (BCBS, 2010). Examples such as the European sovereign debt crisis (Beirne and Fratzscher, 2013), the failure and recapitalisation problems of the OECD banks (Brei *et al*, 2013), the bankruptcies of oligopolistic firms and government agencies (Huang and Lee, 2013), and the increased default observed in specialised retail lending (Van der Burgt, 2008) have demonstrated that the nature of LDPs has noticeably changed in the post-crisis scenario, with major losses occurring as a result of small portfolios, a lack of historical default data, and a reduced number of counterparties (scarce LDPs).

The presence of scarce LDPs underscores two problems for credit risk modelling: the class imbalance problem that underlies the LDPs (Brown and Mues, 2012) and the specification error bias that emerges for small data sets (Beleites *et al*, 2005). Additionally, data scarcity precludes the use of models based on PD distributions to estimate the PD. Literature has largely forgotten the implications and particular characteristics of scarce defaults in small data sets, even if modelling of scarce retail LDPs is particularly critical because the PD must be based on quantitative data and statistical modelling under international regulations (BCBS, 2004, 2010).

The aim of this paper is to conduct a comparative study of cooperative models that may offer potential advantages for default modelling in the presence of scarce LDPs. Our approach provides evidence on recent concerns in the credit risk area, which remained unsolved in the literature: the effect of the actual number of observations on the LDP models (Brown and Mues, 2012), the potential for an ensemble of multiple techniques to examine imbalanced data sets (Brown and Mues, 2012), and the usefulness of bootstrapping approaches to obtain accurate and stable PD estimates over classical undersampling and oversampling approaches (Marques *et al*, 2013).

Cooperative models are constructed by combining individual classifiers such as linear discriminant analysis (LDA), logistic regression (LR), support vector machines (SVM), nearest-neighbour classifiers (*k*-nn), and supervised artificial neural networks (ANN). The models are assessed in terms of their discriminatory power using the area under the receiver operating characteristic curve (AUC), Type-I and Type-II errors, and the DeLong and Friedman test for analysis of the AUC differences between classifiers. In addition, we evaluate the usefulness of individual and cooperative classifiers for generating well-calibrated out-of-time PD estimates. Two real-life credit-scoring data sets were used in the empirical research. The initial populations of 1000 and 125 instances were drastically cut to generate experimental samples of 100 and 25 exposures, respectively, maintaining the out-of-sample data for validation purposes. A moderated imbalance ratio was considered (70/30) in line with the peculiarities of post-crisis LDPs.

The paper is structured as follows. Section 2 provides a critical review of the pre-crisis proposals for LDP management and deepens the topics of class imbalance and specification error as characteristics of the post-crisis scarce LDPs. Methodological issues are discussed in Section 3, including a brief explanation of the individual classifiers used to develop our cooperative proposal. In Section 4 the empirical results are presented and discussed. Section 5 summarises the conclusions and recommendations for further research work.

2. Literature review

The aim of credit scoring is to provide a score of the likelihood of credit applicants to make repayments based on the distinction between good (non-defaulters) and bad (defaulters) payers with consideration of historical default experience. However, because LDPs contain very limited default experience, the historical PD estimates may underestimate the real default rates (BCBS, 2005a; FSA, 2005). Additional drawbacks arise from the validation of the LDP models (BCBS, 2005b; Benjamin *et al*, 2006; Van der Burgt, 2008)[1]: the discriminatory power could be too low for the models to meaningfully differentiate among borrowers; sparse default data might lead to difficulty in statistical backtesting, that is, calibration of whether the PD estimates agree with the observed PD within a confidence interval could be difficult; small data sets could lead to model over-fitting, thus exhibiting low out-of-sample discriminatory power and the rating grades could exhibit a low stability if the models include spurious dependencies from the empirical correlations that decrease the level of forecasting accuracy.

Several regulations have provided guidelines for managing LDPs, including new data sources and data-enhancement tools such as pooling[2] of data with other banks (OeNB, 2004) or using alternative validation tools such as benchmarking and expert judgements (BCBS, 2005a; FSA, 2005). Recommendations for PD estimates are conservative, with default probabilities taken from the upper limit of a confidence interval, backtesting based on robust statistical approaches, good governance and control with banks documenting any aspect of internal models, and timely validation of models reviewed regularly to determine if they remain fully applicable.

Particular strategies have been proposed for PD estimates depending on the nature of the LDPs (FSA, 2005). In the absence of sufficient data (internally or externally) to derive the PD estimates, quantitative techniques have been proposed for use in identifying the appropriate upper limit of the PD, including Bayesian theory and credibility theory, among others (see Appendix A for a review). Such LDP models are largely focused on theoretical PD distributions based on prior knowledge (Löffler *et al*, 2004; Pluto and Tasche, 2005, 2006; Dwyer, 2007; Kiefer, 2009); however, the source

[1] Credit risk model should be validated in terms of discriminatory power (*ex ante* ability to distinguish defaulters and non-defaulters), stability of causal relationships between risk factors and creditworthiness, and calibration referred to the accuracy of PD quantification (OeNB, 2004; BCBS, 2005b).

[2] Collection of the required default data from multiple credit institutions that share their data sets.

of prior information could be little reliable, or even unavailable, for small portfolios with short default histories or with only very recent crisis. In addition, these LDP models are usually tested on *ad hoc* artificial samples (Forrest, 2005; Benjamin *et al*, 2006; Dwyer, 2007; Kiefer, 2009), and critical parameter decisions remain unsolved, such as the choice of the most suitable confidence intervals (FSA, 2005). The literature on real-world LDPs is quite limited (Sabato, 2006; Van der Burgt, 2008); it focuses on large data sets with few default experiences, a limited number of credit risk determinants are considered, and even if backtesting is recognised as the most important concern (Van der Burgt, 2008), no model validation or calibration is provided.

As in any portfolio, the PD estimates must be compared with the observed default rates to guarantee the reliability and usefulness of the LDP models (BCBS, 2005a). However, the credit-scoring techniques face two underlying problems when dealing with small default data sets: (i) the class imbalance between the defaulters and non-defaulters, which produces biased estimates that certain techniques are unable to handle (Brown and Mues, 2012; Marques *et al*, 2013); and (ii) the specification error that arises when estimating a large number of the model parameters from limited and insufficient data.

2.1. Handling imbalanced data sets in credit risk

The class imbalance problem emerges when the number of defaulters is much smaller than the number of non-defaulters (Brown and Mues, 2012; Marques *et al*, 2013). Different techniques have been proposed to manage the drop in performance for the minority class estimation in the presence of class imbalances and are largely focused on data re-sampling (oversampling and undersampling).

Oversampling attempts to balance the data set using replication of the minority class observations; the main drawbacks stem from over-fitting problems, which could be managed by the artificial generation of examples based on interpolation and *k*-nearest neighbour strategies (Chawla *et al*, 2002; Batista *et al*, 2004). In contrast, undersampling aims to balance the data set via the removal of majority class examples; the main drawbacks of this technique arise from a loss of information from the discarded examples, which could be reduced by the selective removal of redundant examples (Kubat and Matwin, 1997). Findings on the most efficient techniques are largely inconclusive, even if certain evidence suggests that oversampling methods may perform better than undersampling ones (Batista *et al*, 2004).

With a focus on credit scoring, Marques *et al* (2013) compare different techniques using five credit data sets with alternative imbalance ratios (20/80 to 7/93). The experiments include large- and medium-sized data sets and models are validated in terms of AUC, Type-I and Type-II errors. The results suggest that oversampling outperforms both the undersampling techniques and the pure imbalance models; however, in the presence of the smallest data set (317 observations), the oversampling and undersampling techniques obtain similar accuracies compared with that of models with no sampling modification.

These results suggest that the presence of scarce data sets reduces the efficacy of imbalanced data models. In this context, effects of over-fitting and information loss become critical because they increase the model specification error, leading to biased parameters and misleading predictive results that cannot be generalised (Sabato, 2006).

2.2. Handling specification error in credit risk

Specification error refers to the uncertainty in the suitability of a model to provide correct representation of the analysed phenomena due to missing variables (omitted variable bias), extraneous variables (irrelevant features, multicollinearity), over-parameterisation (in-sample bias), improper assumptions (linearity, additivity), or inappropriate functional distributions (exponential, dispersion assumptions) (Deegan, 1974; Horowitz, 1981).

The techniques that address specification error depend on the nature of such error. In the presence of small data sets with scarce defaults, three sources of specification error arise: (i) extraneous and correlated variables, (ii) in-sample bias (over-fitting), and (iii) high variance caused by estimators validated on very different partitions of data. Although the two first sources of error can be reduced by feature-selection processes, the third source is more complex and dramatically reduces the discriminatory power obtained over independent data sets (generalisation accuracy). In credit scoring, it leads to backtesting and calibration problems because the PD estimates are largely different than the real PD situations. Uncertainty in the out-of-sample discriminatory power can compromise the model selection, leading to false conclusions for the integrity of the classification model which is aggravated in the presence of small data sets (Beleites *et al*, 2005).

To manage the third source of specification error, robust techniques must be applied for inferring adequate estimates of the model's generalisation accuracy. The test-and-training and cross-validation approaches have been commonly used to estimate the generalisation accuracy of credit models (Thomas *et al*, 2001; Baesens *et al*, 2009; Brown and Mues, 2012; Marques *et al*, 2013). However, in the presence of scarce data sets, the test-and-training approach produces inefficient results because dividing the initial sample into two separated sets generates a significant loss of relevant information that leads to in-sample bias. In addition, small samples lead to a notably large variance of the cross-validated estimates, which suggests over-fitted results (Shao and Tu, 1995).

These results suggest that the traditional approaches to generalisation accuracy do not fit for scarce data sets. As an alternative, re-sampling techniques, such as the bootstrapped approaches proposed in Section 3.2, could be a more efficient alternative in the presence of scarce LDPs.

3. Methodology

3.1. A cooperative approach to credit scoring

Crook *et al* (2007) provide extensive reviews of the state-of-the-art analysis of credit-scoring models, including: (i) heuristic models such as rating questionnaires, expert systems, and fuzzy logic systems; (ii) causal models such as option-pricing models and cash-flow simulation models; (iii) statistical models such as LR, linear and quadratic discriminant analysis; and (iv) machine-learning models such as *k*-nearest neighbour, decision trees, ANNs, SVM, genetic programming, or neuro-fuzzy models. However, conflicting opinions exist as to the most suitable individual classifier, a choice that largely depends on the size and characteristics of the data sets (Baesens *et al*, 2003).

Focusing on scarce-default data sets, Brown and Mues (2012) provide a comparative analysis of 10 well-established credit-scoring techniques for use when facing alternative imbalance ratios. The results from medium-large samples confirm that the model suitability depends on the analysed data set; however, two techniques based on the combination of individual classifiers were observed to yield good performance results with reduced data size, providing initial support for the use of cooperative models for scarce LDPs.

Several theoretical arguments support the strength of cooperative models for handling class imbalance (Kotsiankis and Pintelas, 2003) as well as reducing specification error (Dietterich, 1997; Zheng and Padmanabhan, 2007): (i) the ensemble of models reduces the statistical risk of a wrong choice between similar but opposite hypotheses; (ii) the models cover a larger search space and reduce the risk of becoming stuck in local optima; and (iii) the models represent a larger number of hypotheses that obtain a better approximation to the true endogenous function. The literature has confirmed that an ensemble of models could consistently outperform the individual classifiers (Bauer and Kohavi, 1999), even in the presence of small data sets (Breiman, 1998). A necessary and sufficient condition is that the combined classifiers must be both accurate and diverse (Breiman, 2001). The first condition is a prerequisite of model-combining: the higher the individual accuracy, the higher the discriminatory power of the cooperative models. The second condition exploits hypotheses independence, leading to higher stability: the lower the correlation between individual classifiers, the higher the potential of removing errors by combining classifiers (Dietterich, 1997). Considering both, two broad alternatives have been proposed for building cooperative classifiers (Zheng and Padmanabhan, 2007): the ‘static parallel’ (SP) and the ‘perturb and combine’ (PC) approaches. The SP approach builds independent classifiers in parallel to address a common data set (Zheng and Padmanabhan, 2007). The PC approach uses a unique algorithm on different subsets of data to build individual models. The well-known PC techniques include bagging (based on bootstrapped training samples) (Breiman, 1998), boosting and arcing (based on sequential models of bootstrapped non-covered examples) (Schapire, 1990; Freund and Schapire, 1997).

The literature on credit scoring has recently highlighted the gain in accuracy produced by ensembles of classifiers for large data sets (Nanni and Lumini, 2009; Hsieh and Hung, 2010; Finlay, 2011; Wang *et al*, 2011). However, to the extent of our knowledge, the potential of cooperative models for managing scarce LDPs, although suggested in some works, has not yet been empirically tested.[3]

3.2. The adjusted-cooperative proposal for scarce LDPs

In this paper, a proposal based on cooperative models is introduced to handle scarce LDPs, being organised in three successive stages as analysed below. Because both the SP and PC approaches contain potential advantages in dealing with scarce LDPs, we use these approaches in different steps.[4] The SP approach will be considered for generating individual models, while a PC bootstrapped-based strategy will be used for generalisation of the accuracy estimates.

3.2.1. Generation of individual models. Five well-established scoring techniques are used as individual models.[5] A brief explanation of each of the techniques is presented below.

- *Linear discriminant analysis* (LDA) assigns an observation to the binary response y_i ($y \in \{0, 1\}$) with the largest posterior probability, depending on values of interval or dummy independent predictors (x_{ki}), where $p(y|x) = (p(x|y)p(y))/p(x)$. A discriminant function is obtained to separate the binary classes, $Z_{im} = \alpha_m + \Sigma_{k=1}^{K} \beta_{mk} X_{ik}$, with observations assigned to

[3] Cooperative models are also coherent with the Credibility Theory (FSA, 2005), which ‘is about taking several estimates of some quantity and then computing a weighted average composite estimate, with the size of the weights being determined by how credible the individual estimates are’ (p 10).

[4] The SP approach integrates classifiers with different assumptions, leading to a large search space that increases the generalisation accuracy while simultaneously preventing the loss of information. Individual models perform feature selection on different criteria such that extraneous variables are underweighted in the final cooperative approach. Over-fitting is reduced because models with uncorrelated errors are combined, but no individual strategies for handling imbalanced sets are proposed. Additionally, PC techniques reduce the deviation of individual classifiers using voting bootstrapped models, which could be a potential solution to the high-variance specification error problem and could increase stability. Building sequential models on particular divisions of data could also aid in handling imbalanced classes, but in the presence of scarce data sets, dividing the initial sample into multiple small subsets could lead to over-fitted and highly correlated classifiers, thus reducing the suitability of the ensemble model.

[5] These models are based on different assumptions and use different forms of parameter estimation (Kuncheva, 2004); therefore, they fulfil the independence condition.

each class with respect to a cut-off point ($Z(x) > Z_0$ or $Z(x) \leqslant Z_0$).

- *Logistic regression* (LR) is a generalised linear model that uses a logistic function for modelling the dependent variable, where $\text{logit}(p) = \log(p/(1-p)) = \alpha_m + \Sigma_{k=1}^{K}\beta_{mk}X_{ik}$, p is a score in [0, 1] that reflects the default probability.
- *K-nearest neighbour* (*k-nn*) is a local-search method that uses a distance function, eg the Euclidean distance $d(x_i, x_j) = [\Sigma_{k=1}^{K}(x_{ki} - x_{kj})^2]^{1/2}$, to gather the nearest k neighbours of an unclassified observation; the class most heavily represented in the neighbour area is assigned to the observation.
- *Support vector machine (SVM)* is a supervised learning model that performs a local search based on Vapnik's structural risk minimisation principle (Vapnik, 1995). The SVM uses quadratic programming to find a maximum-margin separating hyper-plane in a selected transformed feature space, so that the examples that are situated closest to the hyper-plane are referred to as support vectors. Both linear and non-linear kernel functions can be considered in defining the decision boundary between classes. In the linear SVM, the K function that measures the similarity of a stored training example $\vec{x}_i$ to the input $\vec{x}$ is linear, and the optimisation problem can be defined as:

$$\min_{\vec{\alpha}} \psi(\alpha) = \min_{\vec{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(\vec{x}_i, \vec{x})\alpha_i\alpha_j - \sum_{i=1}^{N}\alpha_i, 0 \leqslant \alpha_i \leqslant C, \quad \forall i, \sum_{i=1}^{N} y_i\alpha_i = 0, \qquad (1)$$

where u is the output of SVM, y_i is the desired binary output, and C is a tuning parameter to control the trade-off between maximising the margin and minimising the classification error.

- *Artificial neural network* (ANN) is a machine-learning approach that uses massively parallel systems of computing units (nodes) connected and organised into layers to simulate the computational efficiency of a biological nervous system (Bishop, 1995). Among the ANNs, the Multilayer Perceptron or MLP is the most widely used supervised model for classification and prediction. The MLP is organised into successive input layers (that receive exogenous variables), j hidden layers with H_j hidden units (that combine inputs through a weighting scheme) and output layers (that produce the model forecasting). During the training stage, the weights of the network are iteratively adjusted to minimise an error function (eg, the sum of squared errors).

3.2.2. Evaluation of individual models. Different measures have been proposed to evaluate model performance, such as the accuracy ratio, the Gini coefficient measures, Kolmogorov–Smirnov statistics, mean difference statistics, information value, Type-I and Type-II errors (Thomas *et al*, 2001; BCBS, 2005a; Hand, 2005). Although only a few measures are commonly implemented, characteristics of data sets as size, class distribution, or noise can deeply affect the accuracy estimates (Peng *et al*, 2011). In addition, each measure fails to include all information relevant in the context of the scoring problem (Hand, 2005), and, therefore, the conclusions reached could be misleading. In this paper, different measures are used to assess the discriminatory power based on their potential for addressing class imbalance and small samples:

- Accuracy rate (AR): the percent of correctly classified instances or AR = (TN + TP)/(TP + FP + FN + TN), where TN, TN, FP, and FN are the true positive, true negative, false positive, and false negative examples, respectively.
- Type-I error: the number of defaulters misclassified as non-defaulted clients.
- Type-II error: the number of non-defaulters misclassified as defaulted clients.
- ROC (receiver-operating characteristics) curve: the trade-off between the TP and FP rates in a two-dimensional measure of performance. The area under the curve (AUC) represents the accuracy of the classifier and is suggested as an appropriate performance evaluator independent of the class distribution (Sabato, 2006; Brown and Mues, 2012; Marques *et al*, 2013).

Statistical significance of differences between the AUCs derived from different classifiers of a sample will be tested using the DeLong *et al* (1988) non-parametric chi-squared test, which uses the theory of generalised U-statistics and the method of structural components to estimate the covariance matrix of the AUC (Baesens *et al*, 2004). In addition, Friedman's test (Friedman, 1940) will be adopted to compare the AUCs of different classifiers of multiple data sets (Demšar, 2006; Brown and Mues, 2012; Marques *et al*, 2013).

To obtain adequate estimates, the accuracy measures are evaluated using a PC-based strategy, the .632E bootstrapping approach, which has been proven efficient in controlling the specification error (Beleites *et al*, 2005). The .632E bootstrap method (Efron and Tibshirani, 1995) estimates the in-sample bias by randomly drawing B bootstrap sub-samples of size N with replacement, so that the true misclassification error is estimated as:

$$\hat{E}rr_{.632E} = 0.368 \times \overline{err} + 0.632 \times E_0, \qquad (2)$$

where $\overline{err}$ is the error on the training set (re-substitution error), and E_0 is the bootstrap average error on examples not included in each bootstrap sub-sample. Bootstrapping approaches have been found to produce tighter PD confidence intervals than those of the traditional Wald approach in the presence of LDPs (Schuermann and Hanson, 2004). In this paper, the non-parametric percentile-based bootstrapped intervals are obtained as follows:

$$\left[\hat{E}0_{\% low}, \hat{E}0_{\% up}\right] = \left[\hat{E}0_B^{(\alpha)}, \hat{E}0_B^{(1-\alpha)}\right], \qquad (3)$$

where $\left[\hat{E}0_{\%\,low}, \hat{E}0_{\%\,up}\right]$ is the bootstrapped interval, $\hat{E}0_B^{(\alpha)}$ is the 100·αth empirical percentile of the error distribution.

3.2.3. Combination of multiple models. The literature on cooperative models has tested different techniques to combine individual classifiers such as the voting principle, Bayesian formalism, Dempster–Shafer theory, neural networks, statistical regressions, or fuzzy integrals (Lee, 1995). In practice, variants of the voting principle are simple but accurate approaches, which include the unweighted vote, the sum rule, or the weighted vote (Hsieh and Hung, 2010; Finlay, 2011). Out of these, the weighted vote is one of the most accurate variants, which introduces the reliability of the individual classifiers (weights) in the final decision (Finlay, 2011).

Despite their simplicity, voting methods are based on the assumptions of independence between classifiers, which are not fulfilled in practice. This drawback reduces the gain of discriminatory power in presence of highly correlated individual models. To address this drawback, we propose to adjust the weighted vote scheme by considering model similarities, as follows. On the basis of the most accurate model (m*), the individual weights will be adjusted using two alternative factors[6]:

(a) A pseudo-correlation penalty measure in terms of region overlap (Ho, 1998) as:

$$\hat{s}_{m^*,j} = 1 - \frac{1}{N}\sum_{k=1}^{N} f(t_k), \tag{4}$$

where $f(t_k) = 1$ if both classifiers (m*,j) predict the same class for the k-th instance ($k = 1, \ldots, N$), and $f(t_k) = 0$ otherwise.

(b) A pseudo-independence measure in terms of credit scores as:

$$\hat{i}_{m^*,j} = \frac{1}{N}\sum_{k=1}^{N} |(s_{m^*} - s_j)|, \quad 1 \leqslant s_{m^*}, s_j \leqslant 0, \tag{5}$$

where s_{m^*}, s_j are the scores of the k-th instance in the m* and j classifiers, respectively.

4. Empirical research

4.1. Data set

The characteristics of the data sets used in evaluating classifiers are given in Table 1. The German data set contains 1000 instances of retail loans from a major German bank. These instances are classified as good (70%) or bad (30%) clients using a set of 20 numerical and categorical highly correlated attributes. The Japanese data set represents a greatly reduced portfolio that includes 125 instances of credit grant applicants with 85 instances classified as good (68%) and 40 classified as bad (32%). Both data sets are publicly available at the UCI repository (http://archive.ics.uci.edu/ml/) and have been previously used in the LDP and class imbalance literature (Brown and Mues, 2012; Marques *et al*, 2013).

Table 1 Empirical data sets

	Full data set			*Scarce training data set*			*Validation data set*		
	Good	*Bad*	*Total*	*Good*	*Bad*	*Total*	*Good*	*Bad*	*Total*
German data set	700	300	1000	70	30	100	630	270	900
Japanese data set	85	40	125	17	8	25	68	32	100

Because these data sets do not represent sparse portfolios on their own, they have been altered to produce small samples with a reduced number of defaults. This process was carried via random stratified sampling,[7] producing training samples of 10% (German data set) and 20% size (Japanese data set). As a result, sparse training samples were constructed that included a limited number of defaults (Table 1). The non-selected observations were used for out-of-time calibration purposes.

4.2. Experimental set-up

The categorical variables were pre-processed, with each category allocated to a separate variable. For cases in which a category contained very few observations, coarse classing was applied to merge the category with another category with a similar good/bad ratio (Thomas *et al*, 2001; Finlay, 2011). Preliminary feature selection was undertaken using step-wise (forward) regression based on Wilks's partial lambda with a significance level (p) of 5%.

Once features were selected, the individual classifiers were built next.[8] Preliminary experiments were performed to define the parameters of the k-nn, SVM, and ANN models. A set of 10 bins was obtained with replacement from the training sample and was used to build alternative models for different parameter values. The out-of-bin observations were used to test the parameter estimates such that the model with the highest average out-of-bin accuracy rate was selected. For the k-nn algorithm, different values of k were tested,[9] and the final values of $k = 10$ (German) and $k = 4$ (Japanese) were adopted.

[6] Weights are computed from the accuracy of each additive classifier, subsequently multiplied by the pseudo-correlation penalty measure (respectively, pseudo-independence measure), and finally normalised.

[7] For this empirical study, our focus is on the performance of cooperative techniques for small data sets with a reduced number of defaults, such as the emerged LDPs from the post-crisis scenario. Therefore, we have maintained the original imbalance ratio of the original sets (~70/30), which is the moderate, basis split used in the literature when analysing imbalanced data sets (Brown and Mues, 2012).

[8] A cut-off level point of 0.5 was used to calculate accuracy rate and errors.

[9] The proper choice of k largely depends on the data, with k commonly moving between 1 and $N-1$. The smaller the k, the higher the noise and the variance; the larger the k, the higher the bias. A rule of thumb is selection of $k' = N^{1/2}$ (10 and 5 for German and Japanese sets); we tested k ranging from 2 to $2k'$.

Following suggestions from the literature, the Euclidean distance was used as the neighbourhood function (Brown and Mues, 2012).

SVM were built with a linear kernel based on Platt's sequential minimal optimisation algorithm, which is particularly suited to the presence of a small data set. The SVM parameters were established following Platt's (1998) recommendations for classification task: the tolerance for accuracy (minimum relative improvement in the objective function at each step) was established at 0.001 to prevent over-fitting, and the *C*-value that controls margin failures was set to 1.

The ANN classifier architecture was adopted with a single hidden layer and one output node. The best performing number of hidden neurons (two units) was selected based on the 10-bin procedure previously presented.[10] Logistic activation functions were used in both data sets (Brown and Mues, 2012).

Four cooperative models were investigated: the original voting scheme (OVS), weighted-voting (WVS) based on the accuracy rate, adjusted-weighted voting based on the pseudo-correlation penalty measure [AWS(*s*)], and adjusted-weighted voting based on the pseudo-independence measure [AWS(*i*)]. The accuracy was computed using 50 bootstrapping samples,[11] models were run on each classifier using identical training and bootstrapping samples.

4.3. Results for discriminatory power

Table 2 reports the discriminatory power of the individual classifiers on the German and Japanese data sets in terms of out-of-time error (validation dataset). Detailed figures on re-substitution error (training sample), out-of-sample error (.632E bootstrapped subsamples), and confidence intervals are included in Appendix B.

Results show slight differences in terms of the generalisation error, ranging from 27.8% (SVM) to 29.1% (*k*-nn algorithm). The AUC measure points to a higher accuracy for the machine-learning models (SVM and ANN) over that of the pure statistical models (LDA, LR), which is in agreement with previous works on imbalanced sets (eg, Brown and Mues, 2012). The DeLong test confirms a significant difference between the AUC of the best and worst individual model using a significance level of 5%.

The cooperative models generate a smaller generalisation error (27.3–28.0%) than any individual model (except for SVM), and a greater AUC. A major reduction of Type-II error is observed, despite a slight increase in the Type-I error. No differences were found between the original voting strategy (OVS) and the weighting voting scheme (WVS). Although the adjusted-weighted voting strategy generates a small error reduction compared with previous voting methods, the increase in the AUC is rather large. The DeLong test confirmed the presence of significant differences between the best and worse cooperative strategies ($\alpha = 0.05$). Additionally, a significant difference was obtained in the comparison of the best individual and cooperative models.

An unsolved question in the LDP literature is the definition of the most suitable PD confidence interval. Most theoretical confidence levels have been proposed (50.0–99.9%), with certain authors arguing that a confidence level of less than 95% appears intuitively appropriate (Pluto and Tasche, 2005, 2006; Wilde and Jackson, 2006). In our empirical application, the comparison between generalisation and re-substitution overall error does not provide a clear upper confidence level for individual models, which varies between 90.8% (LDA, LR) and 99.8% (*k*-nn). Only the ANN re-substitution error is found to be more predictive (58.0%). The cooperative models obtain more adjusted confidence levels for the re-substitution error, namely, 98.3–98.9% (overall error), 99.8–99.9% (Type-II error), and 84.6–84.8% (Type-I error).

In addition, the bootstrapped error estimates from LDA, LR, and *k*-nn produce more accurate error estimates that are fairly close to the true generalisation error. Tight confidence levels[12] within 50.0–51.2% (overall error), 47.0–56.3% (Type-II error), 40.9–54.1% (Type-I error) were found for the statistical models, but these values deteriorate for the machine-learning approaches. Similarly, the bootstrapped confidence levels for the OVS and WVS are close to 50% for the overall error (55% for Type-II, 60% for Type-I), and those of the adjusted cooperative technique range 51.3–63.2%.

The Japanese data set represents a much sparse sample, and, therefore, the distortions on the error estimates are expected to be larger than those from the German data set. Results in Table 2 confirm the higher variance of the error confidence interval width in both the training and bootstrapping confidence intervals.

The generalisation error falls between 24.0% (SVM) and 39.0% (LDA, ANN) for the individual models. Once more, the classical statistical techniques are the most balanced models in terms of Type-I and Type-II errors. The techniques that achieve the highest AUC are the LR and SVM. The DeLong test confirms a significant difference between the best and worst AUC using a significance level of 5%. The cooperative models generate tighter generalisation errors in the range of 28.0% (AWS models) and 30.0% (OVS and WVS), slightly better than those of the LDA and ANN techniques. However, the AUC measure identifies cooperative approaches as the highest performing models, with significant increases over the individual techniques, particularly with respect to the OVS variant. Again, the DeLong test confirms a significant difference between the AUC of the best and worst cooperative models ($\alpha = 0.05$).

[10] One hidden layer is usually enough to characterise any arbitrary complex hidden function. However, no solutions have been provided to establish the best number of hidden neurons (*h*); a widely used rule of thumb considers $h = (K \cdot M)^{1/2}$, with *K* as the number of input neurons and *M* as the number of output neurons. We tested *h* in [2, 10].

[11] Efron and Tibshirani (1995) recommend at least 25 bootstrap sub-samples (*B*) for statistical purposes and no more than 200 for computational efficiency; 50 bootstrap sub-samples are usually enough to guarantee robust results.

[12] The closer the confidence level to 50%, the better the generalisation accuracy of the classifiers.

Table 2 Results of discriminatory power

	German data set				*Japanese data set*			
	Overall error	*Type-II error*	*Type-I error*	*AUC*	*Overall error*	*Type-II error*	*Type-I error*	*AUC*
LDA	0.286	0.167	0.563	0.720	0.390	0.324	0.531	0.745
	(0.908)	(0.918)	(0.755)		(0.991)	(0.949)	(0.957)	
	(0.512)	(0.520)	(0.530)		(0.563)	(0.567)	(0.394)	
LR	0.286	0.164	0.570	0.718	0.280	0.162	0.531	0.810
	(0.908)	(0.963)	(0.658)		(0.841)	(0.719)	(0.811)	
	(0.500)	(0.470)	(0.541)		(0.303)	(0.383)	(0.455)	
k-nn	0.291	0.167	0.582	0.656	0.260	0.103	0.594	0.665
	(0.998)	(0.999)	(0.950)		(0.412)	(0.093)	(0.884)	
	(0.503)	(0.563)	(0.409)		(0.172)	(0.261)	(0.419)	
SVM	0.278	0.137	0.607	0.723	0.240	0.015	0.719	0.746
	(0.952)	(0.984)	(0.792)		(0.500)	(0.214)	(0.696)	
	(0.461)	(0.605)	(0.496)		(0.155)	(0.635)	(0.216)	
ANN	0.286	0.156	0.589	0.720	0.390	0.324	0.531	0.742
	(0.580)	(0.463)	(0.613)		(0.991)	(0.949)	(0.957)	
	(0.250)	(0.446)	(0.378)		(0.481)	(0.689)	(0.289)	
OVS	0.273	0.137	0.593	0.725	0.300	0.191	0.531	0.910
	(0.983)	(0.998)	(0.846)		(0.972)	(0.825)	(0.955)	
	(0.503)	(0.547)	(0.594)		(0.461)	(0.428)	(0.460)	
WVS	0.273	0.137	0.593	0.725	0.300	0.191	0.531	0.856
	(0.983)	(0.998)	(0.846)		(0.894)	(0.565)	(0.955)	
	(0.503)	(0.547)	(0.594)		(0.461)	(0.428)	(0.460)	
AWS (s)	0.280	0.146	0.593	0.737	0.280	0.162	0.531	0.865
	(0.989)	(0.999)	(0.846)		(0.841)	(0.713)	(0.819)	
	(0.589)	(0.513)	(0.632)		(0.383)	(0.420)	(0.441)	
AWS (i)	0.277	0.143	0.594	0.726	0.280	0.162	0.531	0.777
	(0.987)	(0.999)	(0.848)		(0.841)	(0.713)	(0.819)	
	(0.580)	(0.529)	(0.632)		(0.383)	(0.420)	(0.441)	
DeLong test	SVM *versus* *k*-nn: $p < 0.001$ AWS(s) *versus* WVS: $p < 0.001$ SVM *versus* AWS (s): $p < 0.05$				DeLong test	LR *versus* *k*-nn: $p < 0.001$ OVS *versus* AWS(i): $p < 0.001$ LR *versus* OVS: $p < 0.001$		

Note: Generalisation error (confidence level for re-substitution error) (confidence level for bootstrapped error).

A significant difference was also observed between the AUC of the best individual and cooperative models.

In terms of defining the most suitable confidence intervals, the generalisation error shows a wider range of confidence levels than the German results (re-substitution error functions); such confidence level varies between 41.2% (*k*-nn) and 99.1% (LDA, ANN) for the overall error (9.3–94.9% and 69.6–95.7% for Type-II and Type-I errors, respectively). The cooperative models obtain much tighter and stable results, suggesting a confidence level around 84.1–97.2% for the overall error distribution. The bootstrapped estimates also produce a large range of confidence levels for the individual models, ranging 15.5–56.3% (overall error), 26.1–68.9% (Type-I error), and 21.6–45.5% (Type-II error); again, the LDA is the classifier with the closest bootstrapped estimates. The cooperative models are much nearer to the 50% bootstrapped confidence interval, which suggests their accuracy in predicting the out-of-time generalisation error.

Previous results suggest a potential AUC superiority of the cooperative models in both data sets, which was analysed by Friedman's test (Iman and Davenport's variant) (Demšar, 2006); although no consistent significant difference is observed between the AUC ranks for the individual ($p = 0.274$) or cooperative models ($p = 0.261$), the latter clearly outperforms the individual classifiers using a significance level of 5% ($p = 0.001$).[13]

4.4. Results of calibration

The previous results are complemented by the model calibration, which assesses the accuracy of the PD estimates for each rating grade. A rating system is considered well calibrated if the (*ex ante*) estimated risk measures deviate only marginally from the *ex post* observations (Castermans *et al*, 2010). On a first stage, a suitable number of ratings must be established to calibrate the models. Basel II does not require a minimum number of grades for retail exposures even if a sufficient number should be established for credit risk management. For the LDPs, Basel II recommends a reduction in the rating categories by combining

[13] Post-estimate Nemenyi tests were not performed given the reduced number of analysed data sets; under these conditions, the Nemenyi test exhibits little power (Demšar, 2006).

Table 3 Calibration of the German data set (multiple-grade models)

Calibration (PD estimates)	*Train PD (PR)*		*Bootstrap PD (PR)*		*Generalisation PD (validation data test)*				
	Mean	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$	*Real PD (DR)*	*Confid. level train* $(1-\alpha)$	*Confid. level .632E* $(1-\alpha)$	*DR versus PR distance (train) (%)*	*DR versus PR distance (.632E) (%)*
LDA *Rating 1*	0.122	(0.043, 0.197)	0.182	(0.108, 0.273)	0.147	0.704	0.208	20.49	−19.23
Rating 2	0.286	(0.119, 0.443)	0.289	(0.105, 0.642)	0.269	0.432	0.497	−5.94	−6.92
Rating 3	0.333	(0.102, 0.550)	0.369	(0.123, 0.755)	0.500	0.890	0.821	50.15	35.50
Rating 4	0.600	(0.337, 0.847)	0.456	(0.221, 0.853)	0.504	0.268	0.606	−16.00	10.53
Rating 5	1.000	(1.000, 1.000)	0.738	(0.451, 1.000)	0.597	—	0.157	−40.30	−19.11
LR *Rating 1*	0.128	(0.045, 0.206)	0.182	(0.108, 0.273)	0.142	0.613	0.203	10.94	−21.98
Rating 2	0.261	(0.119, 0.407)	0.289	(0.105, 0.642)	0.261	0.500	0.536	0.00	−9.69
Rating 3	0.333	(0.102, 0.527)	0.369	(0.123, 0.755)	0.500	0.915	0.873	50.15	35.50
Rating 4	0.800	(0.586, 1.000)	0.456	(0.221, 0.853)	0.536	0.018	0.549	−33.00	17.54
Rating 5	1.000	(1.000, 1.000)	0.738	(0.451, 1.000)	0.547	—	0.130	−45.30	−25.88
k-nn *Rating 1*	0.211	(0.122, 0.295)	0.252	(0.077, 0.487)	0.226	0.612	0.462	7.11	−10.32
Rating 2	0.278	(0.099, 0.446)	0.282	(0.102, 0.535)	0.276	0.492	0.471	−0.72	−2.13
Rating 3	0.714	(0.509, 0.907)	0.491	(0.263, 0.895)	0.504	0.041	0.545	−29.41	2.65
Rating 4	0.667	(0.206, 1.000)	0.491	(0.245, 0.877)	0.528	0.305	0.550	−20.84	7.54
Rating 5	1.000	(1.000, 1.000)	0.624	(0.368, 1.000)	0.542	—	0.398	−45.80	−13.14
SVM *Rating 1*	0.071	(0.015, 0.124)	0.157	(0.026, 0.279)	0.092	0.737	0.191	29.58	−41.40
Rating 2	0.250	(0.077, 0.413)	0.252	(0.145, 0.408)	0.241	0.465	0.454	−3.60	−4.37
Rating 3	0.391	(0.170, 0.599)	0.374	(0.172, 0.629)	0.478	0.748	0.867	22.25	27.81
Rating 4	1.000	(1.000, 1.000)	0.714	(0.389, 1.000)	0.588	—	0.234	−99.94	−17.65
Rating 5	n.a.	(n.a., n.a.)	n.a.	(n.a., n.a.)	n.a.	n.a.	n.a.	n.a.	n.a.
ANN *Rating 1*	0.175	(0.092, 0.253)	0.118	(0.000, 0.268)	0.059	0.009	0.254	−66.29	−50.00
Rating 2	0.235	(0.066, 0.394)	0.253	(0.143, 0.396)	0.258	0.591	0.423	9.79	1.98
Rating 3	0.556	(0.331, 0.768)	0.444	(0.226, 0.637)	0.514	0.376	0.711	−7.55	15.77
Rating 4	0.857	(0.514, 1.000)	0.663	(0.368, 1.000)	0.580	0.085	0.362	−32.32	−12.52
Rating 5	1.000	(1.000, 1.000)	n.a.	(n.a., n.a.)	n.a.	n.a.	n.a.	n.a.	n.a.
OVS *Rating 1*	0.095	(0.073, 0.116)	0.174	(0.096, 0.246)	0.122	0.980	0.138	28.42	−29.89
Rating 2	0.250	(0.144, 0.350)	0.260	(0.092, 0.463)	0.296	0.769	0.687	18.40	13.85
Rating 3	0.571	(0.229, 0.893)	0.406	(0.210, 0.613)	0.461	0.293	0.670	−19.26	13.55
Rating 4	0.750	(0.491, 0.994)	0.567	(0.276, 0.908)	*0.596*	0.151	0.751	−20.53	5.11
Rating 5	1.000	(1.000, 1.000)	0.775	(0.368, 1.000)	0.602	—	0.144	−43.20	−26.71
WVS *Rating 1*	0.095	(0.073, 0.116)	0.175	(0.096, 0.251)	0.126	0.980	0.127	32.63	−30.29
Rating 2	0.250	(0.144, 0.350)	0.260	(0.112, 0.433)	0.296	0.769	0.689	17.60	13.85
Rating 3	0.571	(0.229, 0.893)	0.402	(0.210, 0.589)	0.461	0.293	0.670	−19.26	14.68
Rating 4	0.750	(0.491, 0.994)	0.578	(0.276, 0.908)	*0.606*	0.173	0.712	−19.20	4.84
Rating 5	1.000	(1.000, 1.000)	0.782	(0.368, 1.000)	0.531	—	0.127	−46.90	−32.10

AWS (s) *Rating 1*	0.093	(0.071, 0.114)	0.171	(0.104, 0.237)	0.129	0.997	0.197	38.71	−24.56
Rating 2	0.267	(0.154, 0.373)	0.278	(0.098, 0.529)	0.298	0.679	0.604	11.61	7.19
Rating 3	0.533	(0.214, 0.384)	0.394	(0.196, 0.575)	0.461	0.351	0.712	−13.51	17.01
Rating 4	0.714	(0.425, 0.986)	0.532	(0.263, 0.840)	0.571	0.201	0.528	−20.03	7.33
Rating 5	1.000	(1.000, 1.000)	0.782	(0.368, 1.000)	0.591	—	0.136	−40.90	−24.42
AWS (i) *Rating 1*	0.098	(0.075, 0.119)	0.172	(0.106, 0.238)	0.122	0.969	0.150	25.51	−28.49
Rating 2	0.233	(0.129, 0.331)	0.266	(0.086, 0.517)	0.278	0.769	0.599	19.31	4.51
Rating 3	0.500	(0.191, 0.791)	0.381	(0.184, 0.563)	0.451	0.394	0.713	−9.80	18.37
Rating 4	0.750	(0.473, 1.000)	0.545	(0.276, 0.853)	0.566	0.130	0.519	−24.53	3.85
Rating 5	1.000	(1.000, 1.000)	0.782	(0.368, 1.000)	0.595	—	0.138	−40.50	−23.91

*Normal approximation to the binomial distribution (Dwyer, 2007).

Notes: Rating 1 represents the lowest risk (best rating); Rating 5 represents the highest risk; n.a.: 'not available'.

grades to make backtesting more meaningful (BCBS, 2005a). Therefore, the LDP literature has typically developed a two-class prediction model (default *versus* non-default). However, data-enhancement tools are expected to define a higher number of grades.

In this paper, the results are calibrated by: (i) a simple default model that forecasts good (score ⩾ 0.5) *versus* bad (score < 0.5) clients, and (b) an experimentally based multi-grade model that divides the full scoring scale on the higher potential number of grades (5 and 3 grades for the German and Japanese data sets, respectively). The experimental procedure searches for significant PD differences among the grades. Using the bootstrapped sub-samples, alternative *k* rating grades (3–7) were uniformly defined (OENB, 2004). Next, the observed PDs were obtained from examples taken out of the sub-samples and pooled on the pre-defined rating. The PD differences along all grades were tested using the Kruskal–Wallis non-parametric test (Sheskin, 2006). If the null hypothesis was rejected (significant differences exist between at least two groups), a Mann–Whitney U test was applied to analyse the PD differences between pairs of neighbouring grades. The process was initiated by the minimum grade definition (three groups) and repeated for an incremental number of grades until the null hypotheses of both tests were rejected.

Table 3 and Appendix C summarise the calibration results for the multi-grade and default models for the German data set, respectively. The confidence levels for the generalisation PD are computed together with the relative distance between the predicted default rate (PR) and the real default rate (DR), defined as $(DR-PR)/PR$. Appendix C reports small differences in the PD estimates between the default individual models. The non-default category obtains a reduced generalisation PD for the LDA and LR (DR < 0.23), and the SVM and ANN are the better predictors for the default class (DR > 0.55). In any case, the DR for the non-default grade lies in the 95–99.9% confidence level of the PD estimates (re-substitution model) with DR–PR distances of 11.3–37.7%. The bootstrapped estimates are much more accurate, with a maximum distance of 6% (ANN non-default) and a minimum of 1.67% (LDA default). Additionally, the bootstrapped confidence levels approach 50%. The cooperative models obtain similar results, with a slightly higher DR for the predicted non-defaults (23.00–24.00%) and a 54% DR for the predicted defaults. The adjusted weighted voting models are slightly more accurate than the original voting and weighted voting schemes. The confidence levels based on the re-substitution PD are over 99.9%, and the bootstrapped confidence levels are closer to 70% for the predicted non-defaults.

The result in Table 3 shows the differences among the individual classifiers for the multi-grade PD. The LDA and LR generate the widest PD range even if the intermediate ratings produce closer results; the real PD for the better grades reaches 14.2–14.7%. The machine-learning models (SVM and ANN) only recognise four grades as relevant, with DRs per grade that are much different (9.20–58.8% and 5.90–58.0% DR ranges, respectively). The bootstrapped PD estimates (ratings 2 and 4) are closer to DR than those from the extreme grades (ratings 1 and 5), which tend to be overestimated (LDA, LR). In contrast, rating 3 is underestimated in all models (but *k*-nn). These differences suggest skewness differences between the bootstrapped and the real PD distributions. The distances between the DR and PR are smaller for the LDA and LR and increase for machine-learning models.

The cooperative approaches produce higher PD differences among the grades, thus improving their utility for managing credit risk. Although the OVS and WVS do not adequately differentiate between ratings 4 and 5, the AWS models obtain a reduced DR for the predicted non-defaulters (12.2–12.9%) while still differentiating among the five grades. Ratings 4 and 5 produce the largest absolute DR *versus* PR distance, which suggests their potential integration.

No clear upper limits of confidence intervals are obtained for the individual models. For the best rating, the re-substitution PD is underestimated in most models (with the ANN exception). For the worst rating, the re-substitution PD is overestimated in all models. These results suggest a wider range for the estimated PD distribution than the true range, including a bias in the intermediate grades. The cooperative models obtain more stable results, with a confidence level ranging between 96.9% and 99.7% for the best rating, which is reduced to 67.9–76.9% for the second, and successively. Again, the distribution of the confidence levels suggests a narrower DR than that predicted.

The bootstrapped estimates are much closer to the real PD, and the confidence limits are below 50% for the two best grades and the two worse grades and over 50% for the intermediate grades. This result confirms the skewness bias between the DR and PR estimates. This difference is larger for the cooperative models, with confidence intervals 12.7–19.7% for the best grade (absolute DR *versus* PR distance is 24.6–30.3%).

Table 4 and Appendix C summarise the calibration results on the multi-grade and default models for the Japanese data set, respectively. For the default model, LR and ANN represent the best predictors for the non-defaulters (DR in 23.0–23.6%), and SVM is the best default model (DR in 90.0%). However, the *k*-nn is not able to distinguish between categories in the presence of such a small data set. The re-substitution confidence intervals vary in a wide range: 47.3–99.9% (non-default grade) and 8.7–68.8% (default grade). The bootstrapped estimates are much closer to the real PD, and the non-default DR *versus* PR distance ranges 8.0–16.9% (absolute values), and the confidence levels are 27.4–67.9%.

The cooperative models are the best calibrated, producing both a small real PD for the non-defaulters (23.0–23.6%) and a large DR for the defaulters (53.6–57.7%). The confidence levels are very similar between variants, approaching 99.9% for the non-defaulters (as in the German data set). The bootstrapped error estimates are very tight as well, with confidence levels ranging between 49.9% and 57.9% (both defaulters and non-defaulters). The distance between the non-default DR and PR is much reduced (3.5–6.9% in absolute values). Although slight differences emerge between the cooperative models, the adjusted approaches obtain the most accurate and close results.

For the multi-grade ratings, Table 4 reports different real PDs for the best rating, which range from 15.6% (LR) to 28.3% (LDA, ANN) for individual models. However, the machine-learning models are not able to distinguish between the second and third grade, and the LR obtains inadequate PD estimates (PD second grade higher than PD third grade). As expected, higher distortions are observed with respect to the German results.

In spite of this, a smaller distance is observed between DR and bootstrapped PR; eg the absolute distance between the bootstrapped estimates and the real PD falls between 19.41% (LDA) and 27.48% (ANN) in the first grade. However, the re-substitution PD estimates are much more distanced from DRs. Rating 1 is highly underestimated, particularly for the LR (102.60%), LDA (97.90%), and ANN (97.90%) models, such that the confidence level approaches 99.9% (with the exception of the SVM and *k*-nn). In contrast, rating 3 behaves more erratically, and no patterns of the confidence levels are obtained for any resubstitution or bootstrapped estimates.

The cooperative results overcome previous problems, providing robust PD estimates for any grade. The adjusted cooperative models are particularly efficient, producing a 16.1% real PD for the best grade, 56.8% for the second grade, and 100% for the third grade. The confidence level reaches 99.9% for the best grade (re-substitution estimates). The bootstrapped PD is more closely adjusted to the real PD, even if the skewness hypothesis is not confirmed. The absolute distances between the bootstrapped predictions and the real PDs range from 9.9–21.5% (first grade), 4.4–28.5% (second grade), and 72.1–75.8% (third grade). The bootstrapped confidence levels for the best grade approach 40%.

5. Conclusions

In a post-crisis scenario, notably large losses have arisen from small portfolios with a lack of historical default data and a reduced number of observations (ie, sovereigns, OECD banks, specialised retail lending). These portfolios have been largely forgotten in the literature on LDPs, which have provided a reduced number of theoretical and non-empirically tested proposals focused on large data sets.

In this comparative study, we examined a number of statistical and machine-learning techniques for modelling small portfolios with scarce defaults (scarce LDPs) in two real credit risk data sets. Four cooperative models were additionally proposed to address the class imbalance and error specification problems that characterise these samples. The performance of these techniques was assessed in terms of the out-of-sample (.632E bootstrapping estimates) and out-of-time (independent validation set) discriminatory power and calibration results. The discriminatory power was assessed through alternative measures (accuracy rate, Type-I error, Type-II error, AUC).

Table 4 Calibration of Japanese data set (multi-grade model)

Individual models (PD estimate)	*Train PD (PR)*		*Bootstrap PD (PR)*		*Generalisation PD (validation data test)*				
	Mean	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$	*Real PD (DR)*	*Confid. level train* $(1-\alpha)$	*Confid. level .632E* $(1-\alpha)$	*DR versus PR distance (train) (%)*	*DR versus PR distance (.632E) (%)*
LDA *Rating 1*	0.143	(0.066, 0.215)	0.237	(0.053, 0.455)	0.283	0.999	0.724	97.90	19.41
Rating 2	0.444	(0.304, 0.576)	0.405	(0.164, 0.796)	0.306	0.048	0.523	−31.08	−24.44
Rating 3	1.000	(1.000, 1.000)	0.603	(0.368, 1.000)	1.000	0.500	0.882	0.00	65.84
LR *Rating 1*	0.077	(0.041, 0.111)	0.210	(0.028, 0.431)	0.156	0.999	0.389	102.60	−25.71
Rating 2	0.500	(0.200, 0.782)	*0.530*	(0.184, 0.816)	*0.619*	0.750	0.854	23.80	16.79
Rating 3	0.750	(0.016, 1.000)	*0.507*	(0.276, 0.908)	*0.600*	0.365	0.603	−20.00	18.34
k-nn *Rating 1*	0.153	(0.104, 0.199)	0.328	(0.053, 0.455)	0.194	0.921	0.388	26.80	−16.74
Rating 2	*0.556*	(0.258, 0.836)	*0.461*	(0.204, 0.836)	0.613	0.627	0.721	10.25	32.97
Rating 3	*0.500*	[0.000, 1.000)	*0.407*	(0.184, 0.816)	n.a.	n.a.	n.a.	n.a.	n.a.
SVM *Rating 1*	0.238	(0.188, 0.285)	0.302	(0.119, 0.519)	0.230	*0.393*	0.287	−3.36	−23.84
Rating 2	n.a.	n.a.	n.a.	n.a.	1.000	n.a.	n.a.	n.a.	n.a.
Rating 3	0.750	(0.016, 1.000)	0.515	(0.276, 0.729)	0.889	0.626	0.994	18.53	110.17
ANN *Rating 1*	0.143	(0.095, 0.188)	0.222	(0.053, 0.508)	0.283	0.999	0.754	97.90	27.48
Rating 2	0.545	(0.278, 0.796)	0.494	(0.201, 0.833)	0.375	0.140	0.317	−31.19	−24.09
Rating 3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
OVS *Rating 1*	0.077	(0.041, 0.111)	0.202	(0.028, 0.479)	0.182	0.999	0.464	136.36	−9.90
Rating 2	0.500	(0.232, 0.752)	0.457	(0.184, 0.816)	0.477	0.439	0.426	−4.60	4.38
Rating 3	1.000	(1.000, 1.000)	0.569	(0.368, 1.000)	1.000	—	0.896	0.00	75.75
WVS *Rating 1*	0.077	(0.041, 0.111)	0.202	(0.028, 0.479)	0.182	0.999	0.464	136.36	−9.90
Rating 2	0.500	(0.232, 0.752)	0.457	(0.184, 0.816)	0.477	0.439	0.426	−4.60	4.38
Rating 3	1.000	(1.000, 1.000)	0.569	(0.368, 1.000)	1.000	—	0.896	0.00	75.75
AWS (s) *Rating 1*	0.077	(0.041, 0.111)	0.206	(0.028, 0.481)	0.161	0.999	0.370	109.09	−21.46
Rating 2	0.500	(0.232, 0.752)	0.461	(0.184, 0.816)	0.568	0.666	0.685	−13.60	23.21
Rating 3	1.000	(1.000, 1.000)	0.581	(0.368, 1.000)	1.000	—	0.955	0.00	72.12
AWS (i) *Rating 1*	0.077	(0.016, 0.134)	0.205	(0.028, 0.439)	0.161	0.999	0.371	136.36	−21.46
Rating 2	0.500	(0.372, 0.620)	0.442	(0.184, 0.816)	0.568	0.666	0.721	−4.60	28.51
Rating 3	1.000	(1.000, 1.000)	0.581	(0.368, 1.000)	1.000	—	0.953	0.00	72.12

*Normal approximation to the binomial distribution (Dwyer, 2007).

Notes: n.a.: 'not available'. In italics, inconsistent PD estimates (PD second grade higher than PD third grade).

The DeLong and Friedman tests were applied to verify the statistically significant differences between AUCs from the same and distinct samples, respectively. Calibration was based on two-grade and multiple-grade definitions, and confidence levels were obtained on the similarity between the predicted default rates and the real default rates for both the re-substitution and bootstrapped-based forecasts.

The results on the discriminatory power confirmed the class imbalance and error specification problems for individual classifiers in the presence of scarce LDPs: the smaller the data set, the lower the discriminatory power and the higher the variance of the predictions. LR and SVM were the most accurate individual classifiers applied to manage scarce LDPs. The cooperative approaches obtained higher discriminatory power than individual models in terms of AUC (Friedman test $p < 0.05$). In particular, the correlation-adjusted cooperative variants produced highly accurate results. The bootstrapping techniques were found to produce much accurate estimates of the generalisation error *versus* that of the pure re-substitution approaches while maintaining an acceptable balance between Type-I and Type-II errors, in line with Schuermann and Hanson (2004). Confidence levels approximately 40–60% of the bootstrapped error confidence intervals were observed in both data sets, which confirmed the model stability[14] in terms of shorter standard errors than individual classifiers.

The models were first calibrated using a default *versus* non-default rating scale. The PD estimates based on the training model showed a high distance to the real PD that made it difficult to define a confidence level for the individual classifiers. The suggested levels fall within the 95–99.9% interval (German data set) and 75–99.9% interval (Japanese data set). In contrast, the bootstrapped approaches produced more highly stretched estimates, with most real PDs in the 50–60% (German data set) and 35–65% (Japanese data set) confidence levels. The cooperative models produced more stable PD distributions, with confidence levels in the 99.9% range for any data set (training model) and tight bootstrapped estimates (particularly for the Japanese data set with bootstrapped confidence levels around 50%).

Finally, the models were calibrated on a multi-grade rating scale. Five and three different ratings were defined for the German and Japanese data sets, respectively, with certain individual classifiers unable to generate differentiated ratings for scarce data sets. However, the cooperative models produced both accurate and well-calibrated ratings with short distances between the real PD and .632E bootstrapped PD estimates. In particular, the correlation-adjusted cooperative variants were found to produce the most stable PD distributions in the presence of both data sets and obtained the tightest confidence levels (around 45–70% for intermediate grades).

The results suggest that cooperative models based on bootstrapped-based confidence levels are promising techniques for dealing with class imbalance and error specification in the presence of sparse LDPs, producing marginal but significant differences in terms of out-of-time discriminatory power (AUC). The calibration results also suggest the superiority of the cooperative models in obtaining shorter distances between the real PD and the predicted bootstrapped PD. As expected, the ensemble of pseudo-independent models have been able to represent a larger number of hypotheses, removing individual errors, and reducing over-fitting; such effects are particularly enhanced in the smallest sample (25 observations), where cooperative models visibly outperformed individual classifiers (AUC measure).

Several limitations and suggestions for further work arise from this study. First, the results are not representative of the full LDP problem, which includes very different sizes and PD scenarios. Instead, we have focused on small data sets with a moderate imbalance ratio, in line with some post-crisis LDPs. Consequently, further evidence is needed to obtain a broader picture of the accuracy and stability of cooperative approaches for alternative scenarios: models should be tested in highly imbalanced data sets and compared to undersampling and oversampling techniques (Brown and Mues, 2012; Marques *et al*, 2013); also, evidence should be added to test the potential increased accuracy of cooperative approaches on samples with a much reduced number of observations.

Second, it would be of interest to run a search procedure to find alternative individual models with a higher diversity and assess the effects on the cooperative models' performance. While the addition of models would produce a theoretical increase of accuracy and stability, it is expected to obtain an empirical 'stuck point' where adding extra classifiers will not enhance (or even reduce) performance.

Finally, considering the post-crisis nature of small LDPs, the performance of the cooperative approaches should be analysed in the presence of corporate scarce LDPs with consideration of financial information as a primary source for rating companies (or sovereigns). Further research on more advanced bootstrapping approaches should be tested (ie, the wild bootstrap for heteroscedastic residuals and small sample sizes), while alternative combinatorial strategies may be beneficial in the search for the most efficient credit-scoring models for scarce LDPs.

Acknowledgements—We express our appreciation to the referees and the special issue Editor, whose comments stimulated a range of significant improvements to the paper. This work has been partially supported by the Spanish Andalusian Government under the research projects SEJ-5061 and SEJ-111.

References

Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J and Vanthienen J (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society* **54**(6): 627–635.

[14] In this work, stability refers to the similarity between the confidence levels and the absolute distances along samples with different sizes (and/or imbalance ratios).

Baesens B, Verstraeten G, Van den Poel D, Egmont-Petersen M, Van Kenhove P and Vanthienen J (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research* **156**(2): 508–523.

Baesens B, Mues C, Martens J and Vanthienen J (2009). 50 years of data mining and OR: Upcoming trends and challenges. *Journal of the Operational Research Society* **60**(1): 816–823.

Basel Committee on Banking Supervision (BCBS) (2004). *International Convergence of Capital Measurements and Capital Standards: A Revised Framework*. Bank for International Settlements: Basel.

Basel Committee on Banking Supervision (BCBS) (2005a). Validation of low-default portfolios in the Basel II Framework, *Newsletter n. 6*, Bank for International Settlements: Basel.

Basel Committee on Banking Supervision (BCBS) (2005b). *Studies on the validation of internal rating systems*. Technical Report Working Paper No. 14. Bank for International Settlements: Basel.

Basel Committee on Banking Supervision (BCBS) (2010). *Basel III: A Global Regulatory Framework for More Resilient Banks and Banking Systems*. Bank for International Settlements: Basel.

Batista G, Prati RC and Monard MC (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter* **6**(1): 20–29.

Bauer E and Kohavi R (1999). An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning* **36**(1): 105–142.

Beirne J and Fratzscher M (2013). The pricing of sovereign risk and contagion during the European sovereign debt crisis. *Journal of International Money and Finance* **34**(4): 60–82.

Beleites C, Baumgartner R, Bowman C, Somorjai R, Steiner G, Salzer R and Sowa MG (2005). Variance reduction in estimating classification error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems* **79**(1–2): 91–100.

Benjamin N, Cathcart A and Ryan K. (2006). *Low default portfolios: A proposal for conservative estimation of default probabilities*. Working Paper. Financial Services Authority: London.

Bishop CM (1995). *Neural Networks for Pattern Recognition*. Oxford University Press: Oxford.

Brei M, Gambacorta L and von Peter G (2013). Rescue packages and bank lending. *Journal of Banking and Finance* **37**(2): 490–505.

Breiman L (1998). Arcing classifier. *The Annals of Statistics* **26**(3): 801–849.

Breiman L (2001). Random forests. *Machine Learning* **45**(1): 5–32.

Brown I and Mues C (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* **39**(3): 3446–3453.

Castermans G, Martens D, Van Gestel T, Hamers B and Baesens B (2010). An overview and framework for PD backtesting and benchmarking. *Journal of the Operational Research Society* **61**(3): 359–373.

Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16**(1): 321–357.

Crook JN, Edelman DB and Thomas LC (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research* **183**(3): 1447–1465.

Deegan J (1974). Specification error in causal models. *Social Science Research* **3**(3): 235–259.

DeLong E, DeLong DM and Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A non-parametric approach. *Biometrics* **44**(3): 837–845.

Demšar J (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* **7**(1): 1–30.

Dietterich T (1997). Machine Learning research: Four current directions. *AI Magazine* **18**(4): 97–136.

Dwyer DW (2007). The distribution of defaults and Bayesian model validation. *Journal of Risk Model Validation* **1**(1): 23–53.

Efron B and Tibshirani R (1995). *Cross validation and the bootstrap: Estimating the error rate of a prediction rule*. Technical Report 176. Stanford University, Department of Statistics.

Financial Services Authority (FSA) (2005). *Expert Group Paper on Low Default Portfolios*. Financial Services Authority: London.

Finlay S (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research* **210**(2): 368–378.

Forrest A (2005). *Likelihood approaches to Low Default Portfolios*. Joint Industry Working Group Discussion Paper. Credit Research Center (CRC), University of Edinburgh.

Freund Y and Schapire RE (1997). A decision theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1): 119–139.

Friedman M (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* **11**(1): 86–92.

Hand DJ (2005). Good practice in retail credit scorecard assessments. *Journal of the Operational Research Society* **56**(9): 1109–1117.

Ho TK (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8): 832–844.

Horowitz J (1981). Identification and diagnosis of specification errors in the multinomial logit model. *Transportation Research Part B: Methodological* **15**(5): 345–360.

Hsieh NC and Hung LP (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications* **37**(1): 534–545.

Huang H-H and Lee H-H (2013). Product market competition and credit risk. *Journal of Banking and Finance* **37**(2): 324–340.

Kiefer NM (2009). Default estimation for low-default portfolios. *Journal of Empirical Finance* **16**(1): 164–173.

Kotsiankis SB and Pintelas PE (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics* **1**(1): 46–55.

Kubat MS and Matwin S (1997). Addressing the curse of imbalanced training sets: One-sided selection. In: *Proceedings of the 14th International Conference on Machine Learning*; Nashville, TN, pp 179–186.

Kuncheva LI (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken: New Jersey.

Lee DS (1995). *A theory of classifier combination: The neural network approach*. PhD thesis, State University of New York at Buffalo.

Löffler G, Posch PN and Schoene C (2004). *Bayesian methods for improving credit scoring models*. Working Paper. DefaultRisk. http://www.defaultrisk.com/pp_score_46.htm, accessed 2 January 2012.

Marques AI, Garcia V and Sanchez JS (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society* **64**(7): 1060–1070.

Nanni L and Lumini A (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications* **36**(2): 3028–3033.

Oesterreischische Nationalbank (OeNB) (2004). *Rating Models and Validation: OeNB Guidelines on Credit Risk Management Series*. OeNB: Vienna.

Peng Y, Wang G, Kou G and Yong S (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing* **11**(2): 2906–2915.

Platt JC (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in kernel methods-Support Vector Learning*, Technical Report.

Pluto K and Tasche D (2005). *Thinking positively*. Working Paper. Deustche Bundesbank.

Pluto K and Tasche D (2006). Estimating probabilities of default for low default portfolios. In: Engelmann B and Rauhmeier R (eds). *The Basel II Risk Parameters*. Springer: Berlin: pp 79–103.

Sabato G (2006). *Managing credit risk for retail low-default portfolios*. Working Paper. Department of Banking, University of Rome 'La Sapienza'.

Schapire R (1990). Strength of weak learnability. *Journal of Machine Learning* **5**(2): 197–227.

Schuermann T and Hanson S (2004). *Estimating probabilities of default*. Working Paper. Federal Reserve Bank of New York: New York.

Schuermann T and Hanson S (2005). *Confidence Intervals for Probabilities of Default*. Federal Reserve Bank of New York: New York.

Shao J and Tu D (1995). *The Jackknife and Bootstrap*. Springer-Verlag: New York.

Sheskin DJ (2006). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall: Boca Raton, FL.

Thomas LC, Banasik J and Crook JN (2001). Recalibrating scorecards. *Journal of the Operational Research Society* **52**(9): 981–988.

Van der Burgt MJ (2008). Calibrating low-default portfolios using the cumulative accuracy profile. *Journal of Risk Model Validation* **1**(4): 1–17.

Vapnik V (1995). *The Nature of Statistical Learning Theory*. Springer: New York.

Wang GJ, Hao J, Ma H and Jiang H (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications* **38**(1): 223–230.

Wilde T and Jackson L (2006). Low-default portfolios without simulation. *Risk* **8**(1): 60–63.

Zheng Z and Padmanabhan B (2007). Constructing ensembles from data envelopment analysis. *INFORMS Journal on Computing* **19**(4): 486–496.

Appendix A

Table A1 A review of methodological proposals for managing LDPs

	Methodology	*Empirical sample*	*Credit risk determinants*	*Remarks and drawbacks*	*Out-of-sample validation*
Löffler *et al* (2004)	Credit-risk-scoring and Bayesian methodology (prior on external data sets)	4558 observations expanded to 29 500 by random sampling	Yes (six variables)	Accuracy similar to logistic regression. Training and validating samples on the same data set	100 random defaults (PD backtesting)
Schuermann and Hanson (2004, 2005)	Duration method (external rating transition matrices and Markov chains). Small data sets at grade level	Standard and Poor's credit rating history (50 585 observations)	Not used. External PD estimates	Bootstrapped intervals tighter than Wald confidence intervals. Insufficient to distinguish neighbouring grades with few default observations	No
Pluto and Tasche (2005, 2006)	Most prudent estimation based on binomial distributions. 0.50 – 99.9% PD confidence bounds	Not used	Not used. Theoretical PD distributions	PD scaling in presence of few defaults. Proposal for updating PD estimates over time based on accumulated PD frequencies. No solution for PD confidence interval selection	No
Forrest (2005)	Likelihood functions and ratios, and expert opinion (prior odds). PD 95% confidence interval	Artificial sample of 100 observations (for percentile estimates purposes)	Not used. Theoretical PD distributions	Multi-grade PD estimates. No unique PDs are obtained, selection based on maximising regulatory capital	No
Sabato (2006)	Credit-scoring logistic model. Default data are inferred based on shock variables	11 646 observations (Polish bank) and 3383 exposures (Czech bank)	Yes (20 variables)	Tested on real-world retail portfolios. Improved results over expert models and logistic over actual defaults	No
Benjamin *et al* (2006)	Mapping based on look-up tables of PD averages provided by regulators. 50–75% PD confidence intervals	Artificial sample of 500 obligors (7 rating grades)	Not used. *A priori* PD tables	Proposal for updating PD estimates over time based on accumulated look-up tables	No
Dwyer (2007)	Bayesian methodology (numerical integration, Monte Carlo)	Artificial (1000–50 000 observations)	Not used. Theoretical PD distributions	Proposal for PD calibration under shocks	No
Kiefer (2009)	Bayesian methodology and expert opinion. 95% PD confidence level	Artificial samples (100–300 observations)	Not used. Theoretical PD distributions	No out-of-sample validation	No
Van der Burgt (2008)	Calibration model based on ROC curve	Artificial samples (1700 and 4100 observations). Sovereigns (86 countries)	Not used. Mathematical estimates of ROC curve	PD estimated from AUC values. Values within the 95% confidence levels around real PD for artificial portfolios	Just artificial portfolios

Appendix B

Table B1 Discriminatory power results (re-substitution and bootstrapped estimates)

	German data set				*Japanese data set*			
	Train Error $\overline{err}$		*Bootstrap error* $\hat{E}_{0.632E}$		*Train Error* $\overline{err}$		*Bootstrap error* $\hat{E}_{0.632E}$	
	Mean	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$	*Mean*	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$
LDA overall error	0.230	(0.159, 0.297)	0.286	(0.209, 0.370)	0.200	(0.064, 0.328)	0.322	(0.183, 0.530)
Type-II error	0.114	(0.050, 0.175)	0.175	(0.091, 0.285)	0.176	(0.025, 0.320)	0.229	(0.065, 0.467)
Type-I error	0.500	(0.345, 0.646)	0.547	(0.407, 0.658)	0.250	[0.000, 0.511)	0.524	(0.092, 0.724)
LR overall error	0.230	(0.159, 0.297)	0.291	(0.214, 0.397)	0.200	(0.064, 0.328)	0.318	(0.175, 0.503)
Type-II error	0.100	(0.039, 0.157)	0.165	(0.085, 0.304)	0.118	[0.000, 0.239)	0.212	(0.043, 0.511)
Type-I error	0.533	(0.379, 0.678)	0.583	(0.433, 0.670)	0.375	(0.085, 0.648)	0.536	(0.138, 0.770)
k-nn overall error	0.180	(0.115, 0.241)	0.287	(0.214, 0.373)	0.280	(0.128, 0.423)	0.348	(0.200, 0.513)
Type-II error	0.071	(0.019, 0.120)	0.175	(0.063, 0.294)	0.235	(0.066, 0.394)	0.254	(0.087, 0.466)
Type-I error	0.433	(0.280, 0.577)	0.552	(0.382, 0.753)	0.375	(0.085, 0.648)	0.533	(0.138, 0.770)
SVM overall error	0.210	(0.141, 0.274)	0.276	(0.202, 0.342)	0.240	(0.095, 0.376)	0.319	(0.179, 0.476)
Type-II error	0.071	(0.019, 0.120)	0.133	(0.038, 0.238)	0.059	[0.000, 0.148)	0.095	(0.022, 0.393)
Type-I error	0.533	(0.379, 0.678)	0.607	(0.449, 0.778)	0.625	(0.315, 0.917)	0.790	(0.546, 0.862)
ANN overall error	0.277	(0.201, 0.348)	0.320	(0.242, 0.420)	0.200	(0.064, 0.328)	0.332	(0.200, 0.512)
Type-II error	0.160	(0.086, 0.230)	0.202	(0.095, 0.325)	0.176	(0.025, 0.320)	0.208	(0.065, 0.444)
Type-I error	0.563	(0.410, 0.707)	0.617	(0.428, 0.527)	0.250	[0.000, 0.511)	0.578	(0.163, 0.724)
OVS overall error	0.190	(0.124, 0.253)	0.271	(0.193, 0.340)	0.160	(0.036, 0.277)	0.297	(0.149, 0.488)
Type-II error	0.057	(0.010, 0.101)	0.143	(0.058, 0.264)	0.118	[0.000, 0.243)	0.208	(0.043, 0.503)
Type-I error	0.500	(0.345, 0.646)	0.568	(0.428, 0.680)	0.250	[0.000, 0.511)	0.500	(0.092, 0.724)
WVS overall error	0.190	(0.124, 0.253)	0.270	(0.186, 0.352)	0.200	(0.064, 0.328)	0.312	(0.164, 0.503)
Type -II error	0.057	(0.010, 0.101)	0.143	(0.062, 0.282)	0.176	(0.019, 0.323)	0.230	(0.065, 0.525)
Type-I error	0.500	(0.345, 0.646)	0.568	(0.428, 0.680)	0.250	[0.000, 0.511)	0.500	(0.092, 0.724)
AWS (s) overall error	0.190	(0.124, 0.253)	0.271	(0.186, 0.352)	0.200	(0.064, 0.328)	0.313	(0.164, 0.511)
Type-II error	0.057	(0.010, 0.101)	0.151	(0.062, 0.282)	0.118	[0.000, 0.243)	0.208	(0.043, 0.503)
Type-I error	0.500	(0.345, 0.646)	0.552	(0.411, 0.677)	0.375	(0.085, 0.648)	0.546	(0.138, 0.770)
AWS (i) overall error	0.190	(0.124, 0.253)	0.269	(0.149, 0.553)	0.200	(0.064, 0.328)	0.315	(0.138, 0.770)
Type-II error	0.057	(0.010, 0.101)	0.193	(0.062, 0.416)	0.118	[0.000, 0.243)	0.208	(0.043, 0.503)
Type-I error	0.500	(0.345, 0.646)	0.340	(0.273, 0.678)	0.375	(0.085, 0.648)	0.546	(0.164, 0.511)

Normal approximation to the binomial distribution (Dwyer, 2007).

Appendix C

Table C1 Calibration of data sets (default model)

	Train PD (PR)		*Bootstrapped PD (PR)*		*Generalisation PD (Validation set)*				
	Mean	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$	*Real PD (DR)*	*Confid. level train* $(1 - \alpha)$	*Confid. level .632E* $(1 - \alpha)$	*DR versus PR distance (train) (%)*	*DR versus PR distance (.632E) (%)*
German data set									
LDA Non-default	0.195	(0.171, 0.217)	0.220	(0.149, 0.282)	0.225	0.984	0.488	15.38	2.27
Default	0.652	(0.450, 0.842)	0.538	(0.404, 0.698)	0.529	0.151	0.498	−18.87	−1.67
LR Non-default	0.203	(0.179, 0.225)	0.222	(0.152, 0.288)	0.226	0.949	0.546	11.33	1.80
Default	0.667	(0.453, 0.868)	0.543	(0.403, 0.701)	0.530	0.138	0.398	−20.54	−2.39
k-nn Non-default	0.167	(0.145, 0.187)	0.226	(0.150, 0.293)	0.230	1.000	0.534	37.72	1.77
Default	0.773	(0.548, 0.984)	0.538	(0.363, 0.735)	0.518	0.027	0.440	−32.99	−3.72
SVM Non-default	0.198	(0.175, 0.220)	0.228	(0.152, 0.288)	0.232	0.993	0.513	17.17	1.75
Default	0.737	(0.501, 0.959)	0.572	(0.429, 0.745)	0.552	0.092	0.425	−25.10	−3.50
ANN Non-default	0.195	(0.172, 0.217)	0.224	(0.161, 0.279)	0.237	0.999	0.620	21.54	5.80
Default	0.778	(0.529, 1.000)	0.539	(0.429, 0.644)	0.566	0.075	0.595	−27.25	5.01
OVS Non-default	0.185	(0.163, 0.206)	0.219	(0.138, 0.276)	0.238	0.999	0.711	28.65	8.68
Default	0.789	(0.545, 1.000)	0.593	(0.427, 0.728)	0.538	0.041	0.192	−31.81	−9.27
WVS Non-default	0.185	(0.163, 0.206)	0.219	(0.138, 0.276)	0.238	0.999	0.711	28.65	8.68
Default	0.789	(0.545, 1.000)	0.593	(0.427, 0.728)	0.538	0.041	0.192	−31.81	−9.27
AWS (s) Non-def.	0.185	(0.163, 0.206)	0.216	(0.138, 0.276)	0.238	0.999	0.712	−28.65	10.19
Default	0.789	(0.545, 1.000)	0.593	(0.427,0.744)	0.538	0.041	0.454	−31.81	−9.27
AWS (i) Non-def.	0.177	(0.155, 0.198)	0.216	(0.135, 0.272)	0.234	0.999	0.691	32.22	8.33
Default	0.762	(0.539, 0.972)	0.583	(0.409, 0.744)	0.535	0.042	0.291	−29.79	−8.23

Table C1: *Continued*

	Train PD (PR)		*Bootstrapped PD (PR)*		*Generalisation PD (Validation set)*				
	Mean	*Confid. interval** $\alpha = 0.05$	*Mean*	*Confid. interval* $\alpha = 0.05$	*Real PD (DR)*	*Confid. level train* $(1-\alpha)$	*Confid. level .632E* $(1-\alpha)$	*DR versus PR distance (train) (%)*	*DR versus PR distance (.632E) (%)*
Japanese data set									
LDA Non-default	0.125	(0.083, 0.164)	0.232	(0.046, 0.465)	0.270	0.999	0.679	116.00	16.38
Default	0.667	(0.341, 0.974)	0.497	(0.245, 0.877)	0.405	0.087	0.439	−39.28	−18.51
LR Non-default	0.167	(0.122, 0.210)	0.250	(0.041, 0.480)	0.230	0.991	0.443	37.72	−8.00
Default	0.714	(0.331, 1.000)	0.510	(0.263, 0.895)	0.500	0.172	0.609	−29.97	−1.96
k-nn Non-default	0.304	(0.250, 0.355)	0.300	(0.112, 0.463)	0.327	0.765	0.656	7.57	9.00
Default	0.500	(0.000, 1.000)	0.407	(0.184, 0.816)	n.a.	n.a.	n.a.	n.a.	n.a.
SVM Non-default	0.238	(0.188, 0.285)	0.296	(0.178, 0.438)	0.256	*0.729*	0.340	7.56	−13.51
Default	0.750	(0.231, 1.000)	0.479	(0.276, 0.908)	0.900	0.688	0.902	20.00	87.89
ANN Non-default	0.238	(0.188, 0.285)	0.284	(0.088, 0.444)	0.236	*0.473*	0.274	−0.84	−16.90
Default	0.750	(0.231, 1.000)	0.457	(0.276, 0.716)	0.536	0.242	0.669	−28.53	17.29
OVS Non-default	0.118	(0.079, 0.155)	0.227	(0.043, 0.410)	0.236	0.999	0.579	100.00	3.96
Default	0.750	(0.383, 1.000)	0.520	(0.276, 0.908)	0.536	0.274	0.541	−28.53	3.08
WVS Non-default	0.125	(0.083, 0.164)	0.228	(0.046, 0.413)	0.236	0.999	0.577	88.80	3.51
Default	0.667	(0.341, 0.974)	0.488	(0.245, 0.877)	0.536	0.248	0.553	−19.64	9.84
AWS (s) Non-def.	0.167	(0.122, 0.210)	0.245	(0.061, 0.428)	0.230	0.999	0.499	37.72	−6.12
Default	0.714	(0.331, 1.000)	0.507	(0.263, 0.895)	0.577	0.272	0.537	−19.19	13.81
AWS (i) Non-def.	0.167	(0.122, 0.210)	0.230	(0.061, 0.480)	0.230	0.999	0.499	37.72	−6.88
Default	0.714	(0.331, 1.000)	0.577	(0.263, 0.895)	0.577	0.272	0.537	−19.19	12.04

Normal approximation to the binomial distribution (Dwyer, 2007).
Note: n.a.: ‘not available’.

Received 17 January 2012;
accepted 21 August 2013 after two revisions