# Effects of missing data in credit risk scoring. A comparative analysis of methods to gain robustness in presence of scarce data

Raquel Florez-Lopez
Department of Economics and Business Administration
University of Leon
SPAIN
e-mail: raquel.florez@unileon.es

The 2004 Basel II Accord has pointed the usefulness of credit risk management through internal ratings-based models (IRB approach), which use each bank's internal data for categorizing borrowers into risk grades and for estimating risk components: probability of default (PD), loss given default, exposure at default and effective maturity.

Internal data are considered as the primary source of information for the estimation of PD. Banks are allowed to use statistical default prediction models to estimate the PD of borrowers, conditioned to the fulfilment of some requirements related to the accuracy, completeness and appropriateness of the data, which must be representative of the population of actual borrowers.

But in practice internal records are usually incomplete or not contain enough history for the estimation of PD. The presence of missing data is more problematic in presence of low default portfolios, which are characterised by scarce default records that make difficult to design a statistically significant prediction model; nevertheless this problem has been largely ignored in the analysis of credit risk.

Several methods might be used for dealing with missing data, such as listwise deletion, application-specific listwise deletion, substitution techniques or imputation models (simple and multiple variants). Listwise deletion is an easy-to-use method that has been widely applied by social scientists; but in presence of scarce datasets it generates a substantial loss of cases and reduces information's diversity, resulting in a bias in the model's parameters, results and inferences.

The selection of the best method for solving the missing data problem largely depends on the nature of the missing values; nevertheless, there is a lack of empirical analyses about such nature on credit risk, which limits the validity of consequent models.

In this paper, we analyse the nature and effects of missing data in credit risk modelling (MCAR, MAR and NMAR processes). we consider real scarce datasets on consumer and corporate borrowers that include different percents and distributions of missing data.

The findings are used to analyse the performance of several methods for dealing with missing data, such as likewise deletion, simple imputation methods, MLE models and advanced multiple imputation alternatives, based on MarkovChain-MonteCarlo and resampling methods.

Results are evaluated and discussed between models, in terms of robustness, accuracy and complexity; particularly, some multiple imputation models are found to provide very valuable solutions for credit risk missing data.