Using model performance to assess representativeness of data for model development and calibration

C Kruger

orcid.org 0009-0009-0651-6749

Dissertation accepted in partial fulfilment of the requirements for the degree *Master of Science in Risk Analytics* at the North-West University

Supervisor:	Prof WD Schutte
Co-supervisor:	Prof T Verster

Graduation December 2023

ACKNOWLEDGEMENTS

Firstly, I would like to thank my Heavenly Father for His grace and love and for granting me the opportunity and capability to complete my MSc study.

I wish to express my deepest gratitude toward the following individuals:

- My supervisor, Prof WD Schutte for all of his support, guidance and insights during my research. Thank you for continually encouraging me throughout the process.
- My co-supervisor, Prof T Verster for all of her invaluable advice and assistance. Thank you for sharing your knowledge and mentoring me.
- Thank you to my husband, Johann, for your unwavering love and support. Thank you for your endless patience and for always believing in me.
- Finally, thank you to my family for your constant encouragement. It kept me motivated during the ups and downs of the project.

PREFACE

This thesis is written in article format. The research reported in this thesis was done in conjunction with my promoter, Professor WD Schutte and co-promoter, Professor Tanja Verster and consists of one published article.

The thesis is structured as follows: Chapter 1 consists of a general overview of the thesis. The manuscript that was drafted as the primary output of our research and accepted (and published) by the academic journal, can be found in Chapter 2. In Chapter 3, the main findings of the thesis are reported along with conclusions and opportunities for future research. Appendix A contains the author guidelines of the Risks journal, which is a requirement of North-West University.

The article is:

> *Kruger, C. Schutte, WD. & Verster. T. 2021. Using model performance to assess the representativeness of data for model development and calibration in financial institutions. Risks (Special Issue Quantitative Risk Modeling and Management—New Regulatory Challenges) 9: 204. https://doi.org/10.3390/risks9110204.*

The editor confirmed that there is no copyright involved in the publishing of the article as part of the thesis. The promoters agreed on co-authorship and gave consent for the use of this article as part of the final thesis. The student, together with the supervisors, planned the thesis and the student was then responsible for all literature investigations, coding of algorithms, interpretation of results, and the initial drafting of the manuscript.

STATEMENT BY CO-AUTHORS

Herewith is a statement of the co-authors giving permission that the article may form part of this thesis.

I hereby approve the article and give my consent that this article may be published as part of the thesis for the degree *Master of Science in Risk Analytics of Mrs C Kruger.*

ABSTRACT

The main objective of this thesis is to propose a novel methodology that can be employed to assess the representativeness of external or pooled data when it is used in the development and calibration of regulatory models by banks. Currently, there is no formal methodology to assess representativeness, which highlights the significance of this research.

In this thesis, we provide a review of existing regulatory literature to identify the requirements that need to be considered when assessing representativeness. We emphasise that both qualitative and quantitative aspects need to be considered to ensure a comprehensive analysis.

Our proposed methodology is designed to assess the representativeness of external data by utilising model performance as a metric. The methodology is applied to two case studies to demonstrate its effectiveness. In the first case study, we investigate whether a pooled data source from Global Credit Data (GCD) is representative when considering the enrichment of internal data with pooled data in the development of a regulatory loss-given default (LGD) model. The second case study differs from the first by illustrating which other countries in the pooled data set could be representative when enriching internal data during the development of an LGD model.

To validate the effectiveness of our methodology, we compared it with the Multivariate Prediction Accuracy Index (MPAI). Using these case studies as examples, our proposed methodology provides users with a generalised framework to identify subsets of the external data that are representative of their country's or bank's data. This makes our methodology universally applicable for banks to assess the representativeness of external data before utilising it in their regulatory model development and calibration process.

The methodology is not without shortcomings. We have applied our methodology using a linear model and mean squared error (MSE) as performance measure, but it could also be investigated whether the methodology delivers similar performance when a different type of model (e.g. logistic regression) or a different performance measure (e.g.Gini coefficient) are used. This study did not address the validation of external data's representativeness in the absence of internal data. Therefore, it presents an intriguing opportunity for future research to explore how a financial institution can accomplish this task.

Keywords: *representativeness; regulation; LGD; model performance; Global Credit Data (GCD); pooled data*

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

Introduction and motivation

The banking industry has a wide array of business lines (Bessis, 2002) of which providing credit is a major focus. Providing credit is a risky business, as borrowers differ in their ability and willingness to pay (Anderson, 2007). A bank becomes the financial intermediary that brings lenders and borrowers together. Basel is a set of international banking regulations established by the Basel Committee on Banking Supervision (BCBS, 2006). It prescribes minimum capital requirements for financial institutions, intending to minimize credit risk.

In a credit risk context, Basel proposes that regulatory capital be determined by either a standardised approach or an Internal Ratings Based (IRB) approach (BCBS, 2006). Under the IRB approach, the formula suggested for calculating Regulatory Capital (RC) is based on the Asymptotic Risk Factor (ASRF) which is based on the Merton (1974) model that was initially developed for the pricing of corporate debt.

The primary inputs needed in this RC formula are estimates of probability of default (PD), loss given default (LGD) and exposure at default (EAD). Banks who have obtained approval to use the advanced IRB approach usually obtain these estimates from complex models developed in-house.

The minimum requirements for internal rating systems with regards to the data needed are quite extensive and strict, e.g., “Plausible, intuitive and current input data” (BCBS, 2006) and the fact that all relevant information must be taken into account. Without data, quantitative credit models cannot be developed (Anderson, 2007). Data and information are fundamental to the success of credit risk models and provide a competitive edge (McNab & Wynn, 2003).

Generally, internal and external data is considered for developing credit risk models (PD, LGD, EAD). The bank may decide to develop credit models based only on internal data or may choose to supplement this data from external sources such as credit bureaus, central claims repositories, geodemographic data providers, and so forth. Both internal and external data intended for modelling need to be evaluated for reliability and applicability (Siddiqi, 2006).

For any modelling task, a sufficient number of observations should be used to guarantee meaningful models. Some portfolios (e.g. retail credit) have large volumes of data, including the number of defaults experienced. However, other portfolios have limited default experience (so-called 'low-default portfolios' or LDPs) (BCBS, 2006). Examples of LDPs (Benjamin et al, 2006) include low-risk portfolios (e.g. banks, sovereigns, insurers, highly rated corporations), portfolios with a low number of counterparties (e.g. niche markets, public-private partnerships), and portfolios with a lack of historical data (e.g. a new entrant into a market).

Many banks regularly use external data to enrich their internal data to build credit models, especially for data-scarce portfolios. For banks to use external data, they should adhere to regulatory guidelines such as Basel Regulation (BCBS, 2006):

"*(where) external data is used, the bank must demonstrate that its estimates are representative of long run experience*".

Many institutions provide external data, for example, Global Credit Data (GCD). GCD was established in 2004 and is a non-profit association owned by more than 50 banks (GCD, 2022a). The mission of GCD is to provide data for banks by banks. Some of the services provided are pooling credit loss data from banks (focusing on low default portfolios), providing benchmarks, sharing knowledge and other research services (GCD, 2022a). Only member banks have access to the pooled data. The GCD LGD database totals over 195 000 non-retail defaulted loan facilities from around the world (GCD, 2020). The PD database consists of internal ratings from 27 banks where GCD monitors the rating migrations on a basis of 92 000 Large Corporate borrowers for the last 12 years (GCD, 2022b). GCD focuses on high-quality data and works on a "give to get" basis. The robustness of GCD's data collection process ensures a global standard for credit risk data pooling (GCD, 2020).

This thesis will only focus on the LGD and not the PD and EAD components of the RC model. The possibility to use GCD data to model LGD in the South African context will be investigated. For regulatory approval, the external data needs to be tested for representativeness, before it could be used in the development of regulatory models. The thesis will entail testing the representativeness of GCD data within the context of a regulatory LGD model. Note that this research problem originated from the banking industry as there is currently no formal methodology to assess representativeness.

The objective of the thesis

The **objective of this thesis** is to evaluate the criteria of representativeness specifically in an LGD context by proposing a methodology to assess the representativeness of the GCD LGD data in the South African context.

The layout of the thesis

Since this is an article-based thesis, the published paper contains the proposed methodology and is reproduced in Chapter 2. The paper is entitled "Using model performance to assess the representativeness of data for model development and calibration in financial institutions". As stated in the paper the article presents a novel methodology and applies it to two case studies.

Chapter 3 concludes the thesis with some summary remarks and recommendations for future research. Finally, we hope that the thesis could result in an industry-wide accepted methodology to assess the representativeness of data.

REFERENCES (Chapter 1)

Anderson, R. (2007). The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation. New York: Oxford University Press.

BCBS. (2006). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Basel: Bank for International Settlements. Available online: https://www.bis.org/publ/bcbs128.htm (accessed on 19 January 2018).

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.

Bessis, J. (2002) Risk Management in Banking. 2nd Edition, John Wiley & Sons, Ltd., Chichester

GCD. (2020). LGD Report 2020 - Large Corporate Borrowers. Retrieved April 11, 2023, from Global Credit Data: https://globalcreditdata.org/gcd_library/lgd-report/

GCD. (2022a). About GCD. Retrieved April 12, 2023, from Global Credit Data: https://globalcreditdata.org/about-gcd/

GCD. (2022b). PD and Rating Dashboard 2022. Retrieved April 6, 2023, from Global Credit Data: https://tinyurl.com/ew6wpn67

McNab, H & Wynn, A. 2003. Principles and Practice of Consumer Credit Risk Management (2nd edition). Canterbury: Financial World Publishing

Merton, R. C. (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. Journal of Finance, 2, 49-71.

CHAPTER 2: ARTICLE

Article title: Using model performance to assess the representativeness of data for model development and calibration in financial institutions

Authors:

Kruger, C. Schutte, WD. & Verster. T.

The article was published in the Risks Journal (Special Issue Quantitative Risk Modeling and Management—New Regulatory Challenges) in 2021, 9 (204).

risks

MDPI

Article

Using Model Performance to Assess the Representativeness of Data for Model Development and Calibration in Financial Institutions

Chamay Kruger [1], Willem Daniel Schutte [1,2,*] and Tanja Verster [1]

[1] Centre for Business Mathematics and Informatics, North-West University, Potchefstroom 2531, South Africa; Chamay.Oelofse@nwu.ac.za (C.K.); Tanja.verster@nwu.ac.za (T.V.)
[2] National Institute for Theoretical and Computational Sciences (NITheCS), Pretoria 0001, South Africa
* Correspondence: Wd.schutte@nwu.ac.za

Abstract: This paper proposes a methodology that utilises model performance as a metric to assess the representativeness of external or pooled data when it is used by banks in regulatory model development and calibration. There is currently no formal methodology to assess representativeness. The paper provides a review of existing regulatory literature on the requirements of assessing representativeness and emphasises that both qualitative and quantitative aspects need to be considered. We present a novel methodology and apply it to two case studies. We compared our methodology with the Multivariate Prediction Accuracy Index. The first case study investigates whether a pooled data source from Global Credit Data (GCD) is representative when considering the enrichment of internal data with pooled data in the development of a regulatory loss given default (LGD) model. The second case study differs from the first by illustrating which other countries in the pooled data set could be representative when enriching internal data during the development of a LGD model. Using these case studies as examples, our proposed methodology provides users with a generalised framework to identify subsets of the external data that are representative of their Country's or bank's data, making the results general and universally applicable.

Keywords: representativeness; regulation; LGD; model performance; Global Credit Data (GCD); pooled data

Citation: Kruger, Chamay, Willem Daniel Schutte, and Tanja Verster. 2021. Using Model Performance to Assess the Representativeness of Data for Model Development and Calibration in Financial Institutions. *Risks* 9: 204. https://doi.org/10.3390/risks9110204

Academic Editor: Jiří Witzany

Received: 20 September 2021
Accepted: 28 October 2021
Published: 10 November 2021

1. Introduction

The Basel Committee on Banking Supervision (BCBS) establishes guidelines for how banks should be regulated. These regulations relate to all aspects of the models used to estimate risk parameters, amongst others. More specifically, the Basel regulation states the following: "Internal estimates of probability of default (PD), loss given default (LGD) and exposure at default (EAD) models must incorporate all relevant, material and available data, information and methods. A bank may utilise internal data and data from external sources (including pooled data). Where internal or external data is used, the bank must demonstrate that its estimates are *representative* of long-run experience" (BCBS 2006). These regulatory requirements provide the milieu of this research. The aim of this paper is to develop a methodology to measure representativeness when using external data in regulatory models. The most common regulatory models include the PD, LGD and EAD models (Baesens et al. 2016). This research problem originated from the banking industry, as there is currently no formal methodology to assess representativeness.

The concept of using a smaller sample to make an inference about a larger population is an everyday practice, which originated in the statistical literature. Furthermore, we note that the existing literature on assessing whether a smaller sample is representative of an original, more sizable sample (i.e., population) is quite vast (e.g., Mountrakis and Xi 2013; Thompson 2012). However, we are interested in whether a larger data set in terms of the

number of observations (i.e., containing internal and external data) is representative of the smaller sample (i.e., only the internal data) when regulatory model development and calibration takes place. This specific topic has not been widely researched and is, in part, an aim of this study to enable the proposal of a methodology to assess representativeness. This paper will specifically focus on validating whether this larger sample is representative of the smaller sample, where (in most cases) the one sample is not a subset of the other, i.e., disjoint sets. Our proposed methodology provides users with a generalised framework to assess the representativeness of subsets of data.

The layout of the paper is as follows: The paper commences by giving an overview of the literature in Section 2. First, the regulatory requirements on representativeness are provided. From a regulatory perspective, the assessment of representativeness falls into two categories: qualitative aspects and quantitative aspects, which are also discussed in the literature review. Section 3 contains the main contribution of our paper, where we propose a methodology of how model performance could be used to assess representativeness quantitatively. To illustrate the potential uses of our proposed methodology, we apply it to two case studies in Section 4. The first case study exemplifies our methodology when investigating whether a pooled data source is representative, considering the enrichment of internal data with pooled data in developing a regulatory LGD model for a hypothetical South African (SA) bank. The second case study employs our methodology in the context of identifying potential subsets (e.g., countries) within the pooled data that could be considered representative when developing a LGD model (with internal and external data) in the SA context. Section 5 concludes the paper and suggests future research topics.

2. Literature Review

2.1. Regulatory Perspective

Basel requires banks to demonstrate that the data used to develop regulatory models are representative of the population of the bank's actual borrowers or facilities (BCBS 2006). Basel also requires that, where internal or external data is used, the bank must demonstrate that its estimates are representative of long-run experience. Furthermore, the European Capital Requirement Regulations (2013) states that the data used to build a model should be representative of the population of the institution's actual obligors or exposures. Where external data is used by an institution to build models, the Prudential Regulation Authority (PRA) expects the institution to assess the representativeness of the data by considering whether the data are appropriate to their own experience and whether adjustments are necessary (Prudential Regulation Authority 2019).

When developing a regulatory model, an institution must ensure that the population of exposures represented in the data used for an estimation are comparable with those of the institution's exposures and standards. Furthermore, where pooled data is used, the institution should confirm that the pool is representative of the portfolio for which the data is used. Additionally, it is also required that institutions validate their internal estimates using quantitative validation tools and comparisons with relevant external data sources (European Capital Requirement Regulations 2013).

Although the above-mentioned references regularly refer to the concept of representativeness, a methodology to assess representativeness is absent. However, the European Banking Authority (EBA) provides some guidelines. The EBA (2017) splits data representativeness into two sub-sections, namely requirements for the data used in model development and for the data used in the calibration of risk parameters (i.e., for the data used to calculate the long-run average default rate and the long-run average LGD).

Specifically, for assessing the representativeness of data for model development, the focus is on four aspects:

(a) the scope of application;
(b) the definition of default;
(c) the distribution of the relevant risk characteristics;
(d) the lending standards and recovery policies.

For assessing the representativeness of data for calibration, one further aspect is mentioned:

(e) the current and foreseeable economic or market conditions.

The EBA (2017) additionally specifies that, for LGD models, the analysis of (c) should be done separately for non-defaulted and defaulted exposures. In essence, the above list could be broken down into the qualitative and quantitative aspects of representativeness. Although no clear split exists, aspects (a), (b), (d) and (e) are regarded as qualitative aspects and will be discussed in the section that follows. Aspects (c) and (e) relate more to the quantitative assessment of representativeness and will be discussed after the qualitative aspects. Our paper will contribute by proposing a methodology of how a model's performance could be used to assess representativeness, focusing on the quantitative aspects. This is additionally motivated by the US Federal Reserve (Office of the Comptroller of the Currency OCC (2011)), which states that there should be a rigorous assessment of data quality and relevance and that developers should be able to demonstrate that such data are suitable for the model and are also consistent with the theory behind the approach and with the chosen methodology.

2.2. Qualitative Aspects of Representativeness

As mentioned above, regulations deal with both the qualitative and quantitative aspects concerning representativeness. Our focus is predominantly on the quantitative features of representativeness, but it is important to assess the qualitative aspects as well. The EBA (2017) guidelines of testing data representativeness for model development specify some qualitative aspects, namely: the scope of application, the definition of default, lending standards and recovery policies.

The qualitative aspects are crucial in assessing whether data (internal or external) are representative for model development. Expert judgement should be used to determine whether the scope of application will make sense when using this data. The definition of default in the data to be used should be in line with the definition of default of the model that is developed. It is essential to understand the difference in the lending standard and recovery policies in the data used in the modelling and the environment in which the developed model will be implemented. A significant component of the qualitative aspect of assessing whether internal or external data are representative for the model development is to use business knowledge (i.e., common sense joined with experience). Additionally, the EBA (2017) also adds that the current and foreseeable economic or market conditions should be considered in the qualitative aspect of testing for representativeness. The most important concept to validate when assessing the qualitative aspects of representativeness is whether each aspect in your external data is more or less aligned to the conditions applicable where the model will be applied. Since our focus is predominantly on the quantitative aspects—which will be discussed next—some further remarks on the qualitative aspects can be found in Engelman and Rauhmeier (2011).

2.3. Quantitative Aspects of Representativeness

While the qualitative aspects are notable, the quantitative aspects concerning representativeness will be the focus of this research. Assessing the representativeness could refer to both internal and external data and should be considered under multiple dimensions. In this section, we will consider existing quantitative approaches that could be used to assess representativeness, followed by our proposed methodology to measure representativeness by using model performance metrics (Section 3). One of the regulatory requirements is to test whether the distribution of the risk drivers is similar when comparing one data set to another (i.e., internal and external data). Note that if the bank does not have enough data to build a model, comparing the distribution of the risk drivers of the external data with that of the internal data is nearly impossible. If the bank does have enough data, the distribution of the feature (e.g., the PD) that will be modelled could be compared with the distribution of the same feature using the external data. Representativeness can thus be

analysed by cross-sectional comparison of the distribution of the risk factors and some other key factors (such as countries, regions, industry sectors, company type, obligor size, etc.) for each sample. In this context, frequency plots and tables ordered by the frequency of each realisation (for discrete factors) can be particularly useful. For risk factors on a continuous measurement scale, statistical tests such as the Kolmogorov–Smirnov and Anderson–Darling tests can be used (D'Agostino and Stephens 1986). These tools can be supplemented with basic descriptive statistics (e.g., difference of the medians of both samples relative to their standard deviation or the ratio of the standard deviations on both samples) (Engelman and Rauhmeier 2011). In summary, the list of risk drivers needs to be determined, and the distribution of each risk driver could then be compared.

Following from the above, we need to determine the degree of similarity between the distributions and not necessarily the equality. It is important to remember that the reason for using external data is to enrich the internal data. If the distributions are identical, the internal data will not be enriched but simply be expanded with more observations with the same characteristics. Formal statistical tests on assessing the similarity of distributions across samples were not found to be helpful, since the question is not whether distributions are identical (typically, they are not) but whether they are sufficiently similar for the extrapolation of results and estimates derived from one sample to the other sample (Engelman and Rauhmeier 2011). An example of this is where a bank's current population age is between 30 and 50 while the external data is spread across a wider range, for example, a population age between 20 and 60. This will enrich the data available for modelling and ensure that the developed model will cater to a broader population.

Engelman and Rauhmeier (2011) gave some guidelines on a potential methodology when the data are found to be unrepresentative. The most important aspect is to ascertain whether the problem occurs only for a few risk factors or for the majority. In the first case, the reasons for the differences have to be analysed, and the development samples should be adjusted accordingly. One reason might be that the distributions of obligors across regions or industry sectors are different. The development sample can then be adjusted by reducing the number of obligors in those regions or industry sectors that are overrepresented in the development sample. In the second case, a variety of approaches can be considered, depending on the specific situation. Examples include the reduction of the range of the risk factors so that it only includes areas that are observable in both the development and the target samples. Furthermore, the weight of a risk factor found to be insufficiently representative can be reduced manually, or it can be excluded from the analysis.

Other methods to compare the distribution of two data sets include the Kolmogorov–Smirnov test, the chi-square test, population stability indices, etc. Although both the chi-square test and population stability indices are mentioned, the chi-square test and the PSI are essentially the same measure when the PSI is appropriately normed. Ramzai (2020) considered the population stability index (PSI) and the characteristic stability index (CSI) as the two most widely used metrics in credit risk to assess whether the model is still relevant and reliable when, for example, applying the model to a data set following the development of that model. The PSI and CSI establish whether there are any major differences when comparing these two data sets, especially shifts in distributions. The PSI can evaluate the overall population distribution (of the two sets), while the CSI can narrow it down to the specific features that are causing fluctuations in the distributions. For a discussion on the PSI and some other tests relating to the comparisons of distributions, see Prorokowski (2018), Siddiqi (2006) and Taplin and Hunt (2019). Furthermore, the topic of comparing distributions is still relevant when considering the recent work of Yurdakul and Naranjo (2020). Although the PSI is currently widely used in the industry as a "traffic light indicator approach" by employing "rule of thumb" threshold values to assess changes from the original data, limited studies on the statistical properties (and the thresholds) of the PSI exist. In this regard, Yurdakul and Naranjo (2020) examined the statistical properties and proposed a data-dependent approach to obtain the thresholds for the PSI. An alternative to

the PSI was also recently proposed in the form of the Prediction Accuracy Index (PAI) that overcomes several disadvantages of the PSI (Taplin and Hunt 2019). Since we proposed a methodology to test representativeness, we would prefer to compare it to some standard measure. In the absence of a formal methodology or measure prescribed by regulations, we utilised the PAI as a potential measure to relate our results. For this reason, the PAI will be briefly discussed as part of the case study results.

In summary, when considering the recent literature together with the banking regulations, and since regulations are not prescriptive of a methodology, we concluded that the following guidelines can potentially be applied to assess the representativeness of the data during model development:

- assessing the qualitative aspects of representativeness using expert judgement;
- assessing the quantitative aspects of representativeness using distributional comparisons, for example, the use of the PSI;
- using the Prediction Accuracy Index (PAI) as an alternative to the PSI when comparing distributions (Taplin and Hunt 2019).

In this regard, our research focused on proposing a methodology to assess representativeness.

3. Generic Methodology to Assess Data Representativeness Quantitatively

Following our investigation into the regulatory requirements and literature pertaining to the assessment of the representativeness of external data, we propose the following methodology to assess the representativeness of data for model development and calibration. We also introduce a notation that will be used throughout.

The proposed methodology is to assess whether the data set in question (Data set Q) is representative of a base data set (Data set B). Without a loss of generality, the dependent variable is referred to as LGD (defined in more detail below), since our case studies apply the methodology in an LGD context. Any other dependent variable could also be used. The methodology consists of five steps:

Step 1: Split the base data set (Data set B) into disjoint subsets: one part for building a model, namely Data set BB (Base Build), and another part to evaluate the model that was developed, say, Data set BT (Base Test).

In predictive modelling, the typical strategy for an honest assessment of the model performance is data splitting (Breed and Verster 2017). Data splitting is the method of dividing a sample into two parts and then developing a hypothesis or estimation method using one part and testing it on the other part (Barnard 1974). Picard and Berk (1990) reviewed data splitting in the context of regression and provided specific guidelines for validation in regression models, i.e., use 20–50% of the data for testing.

Step 2: Develop a model using Data set BB (Base Build). We refrain from specifying any model building technique, as the methodology is generic, and any technique could potentially be applied.

Step 3: Join the data set in question (Q) with the subsample of the base data set (BB) and develop another model (it should be the same class of models as the one developed in Step 2) on this augmented data (Data set Q + BB).

Step 4: Evaluate the model performance (e.g., mean squared error (MSE) or any other measure) of these two models on the test subsample of the base data set (Data set BT) and determine whether the model performance improved or remained similar using the following construct:

1. Define $MSE_{Q+BB,BT}$ as the MSE (or any other model performance measure) of Data set BT using the model developed on Data set Q + BB and $MSE_{BB,BT}$ as the MSE of Data set BT using the model developed on Data set BB.
2. If $MSE_{Q+BB,BT} < MSE_{BB,BT}$, the model developed on the augmented data has improved the model performance compared to the model developed only on the base data. Suppose that more substantiation is required regarding the significance of the difference between the MSEs, then the formal tests proposed in Step 5 could

be optionally performed. However, if $MSE_{Q+BB,BT} \geq MSE_{BB,BT}$, the formal tests proposed in Step 5 should be performed.

If the MSE is used as the performance measure, the equations of $MSE_{BB,BT}$ and $MSE_{Q+BB,BT}$ are given as:

$$MSE_{BB,BT} = \sum_{i=1}^{N_{BT}} \frac{\left(LGD_i - \widehat{LGD}_{i,BB,BT}\right)^2}{N_{BT}}, \quad (1)$$

and

$$MSE_{Q+BB,BT} = \sum_{i=1}^{N_{BT}} \frac{\left(LGD_i - \widehat{LGD}_{i,Q+BB,BT}\right)^2}{N_{BT}}, \quad (2)$$

where

- LGD_i indicates the observed outcome for observation i,
- $\widehat{LGD}_{i,Q+BB,BT}$ indicates the predicted outcome for observation i calculated on Data set BT using the model build on Data set Q + BB and
- $\widehat{LGD}_{i,BB,BT}$ indicates the predicted outcome for observation i calculated on Data set BT using the model build on Data set BB for $i = 1, \ldots, N_{BT}$, where N_{BT} is the number of observations in Data set BT.

Step 5: During Step 4, if it was found that $MSE_{Q+BB,BT} \geq MSE_{BB,BT}$, a dependent two-sample test (e.g., a parametric or a nonparametric test) is performed to determine whether the model developed on the augmented data has a similar model performance to the model developed only on the base data. The tests most suitable for this scenario are either a parametric test (e.g., a paired t-test) or a nonparametric test, e.g., the Sign test and/or the Wilcoxon rank-sum test (Sprent and Smeeton 2001).

The assumptions associated with the preferred test should also be checked during this step. For illustrative purposes, the paired t-test based on two different residual statistics will be used to describe the methodology in the case studies that follow. If the test concludes that the $MSE_{Q+BB,BT}$ is not statistically different from the $MSE_{BB,BT}$, we deduce that the model developed on the augmented data (Data set Q + BT) has a similar model performance to the model developed only on the base data (Data set BB). Data set Q is therefore not atypical (i.e., unrepresentative) for model development and calibration. In our view, this translates into evidence that the representativeness of the data (Data set Q) has been assessed and is appropriate to our own experience (Data set B) and aligned with the requirement of the PRA.

Two residual statistics were proposed when performing the test, namely, the absolute error and the squared error. Either set of residual statistics (or both) can be used. The absolute error calculated on Data set BT using the model developed on Data set BB $(Absolute\ Error_{BB,BT})$ is calculated as follows, for $i = 1, \ldots, N_{BT}$:

$$Absolute\ Error_{i,BB,BT} = \left|LGD_i - \widehat{LGD}_{i,BB,BT}\right|, \quad (3)$$

and the absolute error calculated on Data set BT using the model developed on Data set Q + BB $\left(Absolute\ Error_{Q+BB,BT}\right)$ is calculated as follows:

$$Absolute\ Error_{i,Q+BB,BT} = \left|LGD_i - \widehat{LGD}_{i,Q+BB,BT}\right|, \quad (4)$$

where LGD_i indicates the observed outcome value. Furthermore, $\widehat{LGD}_{i,BB,BT}$ and $\widehat{LGD}_{i,Q+BB,BT}$ were defined above.

Similarly, the squared error calculated on Data set BT using the model developed on Data set BB $(Squared\ Error_{BB,BT})$ is calculated as follows:

$$Squared\ Error_{i,BB,BT} = \left(LGD_i - \widehat{LGD}_{i,BB,BT}\right)^2, \quad (5)$$

and the squared error calculated on Data set BT using the model developed on Data set Q + BB $\left(Squared\ Error_{Q+BB,BT}\right)$ is calculated as follows:

$$Squared\ Error_{i,Q+BB,BT} = \left(LGD_i - \widehat{LGD}_{i,Q+BB,BT}\right)^2, \quad (6)$$

with all the symbols defined earlier.

3.1. Remarks

This methodology is formulated in a generic way. References to the dependent variable could be anything, such as the PD, EAD or LGD. These models could be developed for any type of portfolio, e.g., international ship finance loans, Small and Medium Enterprises (SMEs) in Italy or large corporations in Japan. The modelling technique is also not limited, and any technique, such as logistic regression, linear regression, decision trees, survival analysis, etc., can be used. All the underlying assumptions of the chosen modelling technique should, however, be assessed. In the case of nonlinear models, e.g., neural networks, an augmentation of the data set may lead to model overfit. There are various techniques to manage overfitting, e.g., data splitting and limiting the degrees of freedom (by preselecting useful inputs and reducing the number of hidden nodes). These techniques should be considered when nonlinear models are used in the application of our proposed methodology.

Furthermore, many other measures could be used to assess the model performance (such as goodness-of-fit measures) and will depend on the type of model developed. Although we proposed the MSE as a performance measure, many alternatives exist, e.g., the Gini coefficient (frequently used for PD models), R-squared statistic (for regression models), Akaike information criterion (AIC), Bayesian information criterion (BIC), likelihood ratio statistic, Wald statistic (Neter et al. 1996) and the Diebold Mariano test (Diebold 2015). Specifically, for LGD models such as the one used in the case study, Li et al. (2009) provided a set of quantitative metrics that can be used in the LGD model validation process. When generalising from the literature, it is evident that the main areas when measuring the model performance are accuracy, ranking and stability (Prorokowski 2018). The following definitions of these areas are provided by Baesens et al. (2016):

- Stability measures to what extent the population that was used to construct the rating system is similar to the current population.
- Discrimination measures how well the rating system provides an ordinal ranking of the risk.
- Calibration measures if there is a deviation of the estimated risk measure from what has been observed ex-post.

These three areas of Baesens et al. (2016) can easily be translated into the terminology of Prorokowski (2018), i.e., accuracy is comparable to calibration, ranking is similar to discrimination and stability is self-explanatory. Our aim is to assess the data representativeness of both model development and model calibration. Therefore, an accuracy measure rather than a ranking measure should be used. As MSE is one of the most common measures of accuracy, we propose this measure in our methodology when assessing the representativeness of the external data. Additionally, from a statistical perspective, the MSE measures both bias and variance.

Our methodology is, however, not without limitations, and we have identified the following aspects that could be improved upon. We acknowledge that the methodology

requires the user to choose between several alternatives when it comes to the following aspects:

- the model methodology;
- the performance measures;
- the type of dependent two-sample test;
- the significance level.

Furthermore, for each of these choices, the underlying assumptions should be carefully evaluated, and if the assumptions are not met, the specific choice needs to be re-evaluated. Apart from the limitations, we should also consider how our proposed methodology relates to familiar techniques such as data splitting and cross-validation. The purpose of data splitting or cross-validation is "model assessment" or "model methodology selection" (James et al. 2013; Hastie et al. 2009; Sheather 2009; Zhang and Yang 2015). Model methodology selection refers to estimating the features of different models in order to choose the best one, while model assessment implies that a final model has been chosen and we need to estimate its prediction error on new data (Hastie et al. 2009). It should be clear from the above that the focus of cross-validation is on the different aspects of the model (i.e., what model to use and which model is more accurate), while our focus is on the characteristics of the data that are used during model development and validation (i.e., Would a model that was developed/validated on augmented data be representative of own experience?). Arlot and Celisse (2010) offered a comprehensive review of cross-validation procedures and their uses in model selection, while Zhang and Yang (2015) clear up some misconceptions relating to cross-validation that exist in the literature.

3.2. Roadmap/Summary to Assessing Representativeness

To summarise the aspects of representativeness discussed so far, we graphically illustrate this in Figure 1. This is also our proposed framework to assess representativeness. First, we assess the data on a qualitative basis. This includes the scope of application, definition of default, lending standards, recovery policies and business opinion (discussed in Section 2). Second (and within the focus of our research), the quantitative aspects should receive attention, namely, the investigation of the PSI/CSI/PAI, the comparison of the distribution of risk drivers (also discussed in Section 2) and the application of our proposed methodology to assess the representativeness. Our definition of representativeness consists of an assessment across multiple dimensions, i.e., we consider one data set (external data) to be representative of another data set (internal data) if the former data set displays sufficiently similar characteristics to those of the latter data set. Under these conditions, the data sets can be joined when developing/validating a model that should be representative of one's own experience. We will use this definition of representativeness in the application of our proposed methodology in the sections that follow.

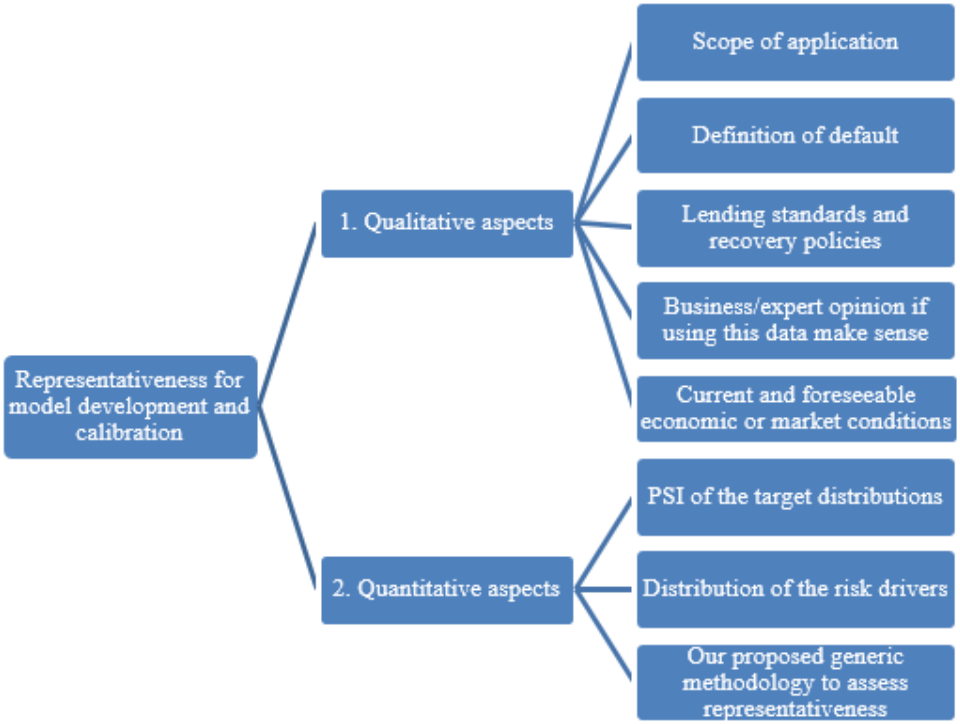

Figure 1. Tree-based diagram to define representativeness (across multiple dimensions).

4. Case Studies

Two case studies will be discussed to illustrate the methodology proposed in Section 3. In the first case study, we illustrate how our proposed methodology can be applied when a bank is in the process of investigating whether a subset of a pooled data source could be used as a representative source of external data for the enrichment of internal data in developing a regulatory LGD model. For such a model, LGD would be defined as one minus the recovery rate, where the recovery rate is equal to the discounted recovered amount divided by the EAD (de Jongh et al. 2017).

In the second case study, we test our methodology when investigating potential subsets (e.g., countries) within the pooled data source that could be considered representative when developing a LGD model. In both cases, pooled data of Global Credit Data (GCD 2019) are used. GCD is a non-profit association owned by its member banks from around the world. The mission of GCD is to assist banks in improving their credit risk models through data pooling and benchmarking activities. GCD can be summarised by the phrase: "By banks, for banks" (GCD 2019). We will start this section by first describing the methodology and the data used for both case studies, followed by the presentation of the results.

4.1. Methodology and Data

4.1.1. Modelling Technique Used

Many techniques are available to model LGD. Joubert et al. (2018a), for example, made use of the default weighted survival analysis to directly model LGD. This survival analysis method is then compared with other techniques to model LGD, namely beta regression, ordinary least squares, fractional response regression, inverse beta, run-off triangle and Box–Cox model. Indirect modelling methodologies can also be used to predict LGD using two components, namely the loss severity component and the probability component. Examples of models used to predict the loss severity and the probability component are haircut survival analysis models (Joubert et al. 2018b). In other literature, quantile regression was used to predict the LGD (Krüger and Rösch 2017). In this last-mentioned reference, quantile regression was compared with the ordinary least squares, fractional

response model, beta regression, regression tree and finite mixture models. Log semi-nonparametric distributions have also proved helpful in modelling skewed and fat-tailed distributions (Cortés et al. 2017). In our paper, however, we use ordinary least squares in the case study, but the proposed methodology is generic and can be applied independently of the modelling technique used. Our choice of modelling technique (i.e., linear regression) is not an uncommon method, as linear regression is frequently used in LGD model settings (Loterman et al. 2012). The focus, however, is not on the development of a superior LGD model but on the demonstration of the proposed methodology.

When using linear regression, we will report both the coefficient of determination (R-squared) and the adjusted R-squared value to assess the goodness-of-fit of the regression model. The R-squared statistic is the proportion of variance in the dependent variable (LGD) that can be predicted from the independent variables. Note that this is an overall measure of the strength of association and does not reflect the extent to which any independent variable is associated with the dependent variable. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model, which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in the R-squared statistic would simply be due to chance variation in that sample. The adjusted R-squared statistic attempts to yield a more honest value by estimating the R-squared for the population and by adjusting for the number of predictors in the model (Neter et al. 1996). Furthermore, the underlying assumptions (Neter et al. 1996) of linear regression are:

- The model is linear in the parameters and variables.
- The error terms are normally distributed.
- The regressors are independent of one another (no collinearity).
- The error terms are independently distributed.
- The error terms have constant variance (no heteroscedasticity).

All these assumptions were checked before proceeding with the rest of the analysis. When using linear regression, we will determine whether a variable is statistically significant by considering a significance value of 5%. Note that the *p*-value to use for the selection of variables should be adjusted with respect to the sample size. Typically, larger samples should use smaller *p*-values (Wasserstein and Lazar 2016).

4.1.2. Data

We used the unique loss data base of GCD to construct the subsamples of data for both case studies. The data base includes detailed loss information on a transaction basis of all of the member banks from around the world. For both case studies, we used the obligor level data, and throughout, the base data (Data set B) was randomly split into an 80% build data set (Data set BB) and a 20% test data set (Data set BT). Furthermore, all analyses were generated using SAS/STAT software, Version 9.4 (TS1M3). Copyright © 2021 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA. The specific detail concerning the data for each case study can be found in the introduction of the results, although the data characteristics common to both case studies are given next, including the data preparation that was done.

4.1.3. Dependent Variable Used

The GCD data set (GCD 2018) provides users with different LGD values, calculated depending on how advances after default are treated. For the one LGD value, advances after default are treated as cash flows and are included in the loss calculations. For the other LGD value, these advances are treated in the EAD and are included in the default amount. We used the latter in our research. We only used data from 2000 to 2015 due to the lack of data prior to 2000. To ensure a sufficient workout period when calculating the LGD, we only used data up to and including 2015. This is to address the resolution bias caused

by cured cases (GCD 2018). The rationale behind this is to wait for the data to mature before using it in model development (Cutaia 2017).

The reference data set used in this study comprises the large corporates and SMEs aggregated on the obligator level. The definitions of large corporates and SMEs are given in Table 1. The LGD values are provided in Table 2. We used the data of H2/2018 (GCD 2019).

Table 1. GCD definition of the asset classes (GCD 2019).

Asset Class	GCD Definition
SME	Borrowers in the Corporate Asset Class as defined in the Basel II Accord §218 and §273, where the reported sales for the consolidated group of which the firm is a part is less than €50 million and where the exposure is not treated as retail, i.e., group exposure > €1 million.
Large corporate	Borrowers in the Corporate Asset Class as defined in the Basel II Accord §218 and §273, where the reported sales for the consolidated group of which the firm is a part is above or equal than €50 million but which is not reported in a more specialised Asset Class.

Table 2. LGD values with the associated number of observations for SMEs and large corporates.

SMEs	n	Large Corporates	n
$LGD < -0.01$	175	$LGD < -0.01$	14
$-0.01 \leq LGD \leq 1.5$	3600	$-0.01 \leq LGD \leq 1.5$	231
$LGD > 1.5$	0	$LGD > 1.5$	0

Within the South African context, we considered a business with a turnover of more than R400,000 as a large corporate (South African Reserve Bank SARB (2015)). Another complicating factor is to investigate the effect of inflation on these figures. In a developing country such as South Africa, inflation plays a significant role in any rand values assessed over time. This could, however, be a topic of further research. We used the country of jurisdiction of the loan to identify the country.

4.1.4. Independent Variables Used

The following independent variables were considered in the modelling process:

- EAD: Exposure at default.
- Facility type: Represents the different loan types. Member banks are responsible for mapping their own internal facilities denominations to the GCD Facility types.
- Seniority code: Debt grouped and assigned a code according to seniority level (e.g., Super Senior, Pari-Passu, Subordinated and Junior).
- Guarantee indicator: Indicates whether a loan has underlying protection in the form of a guarantee, a credit default swap or support from a key party.
- Collateral indicator: Indicates whether a loan has underlying protection in the form of collateral or a security.
- Industry code: The industry that accounts for the largest percentage of the entity's revenues.

Note that these six variables were used in another study that modelled the LGD using GCD data (Krüger and Rösch 2017). Li et al. (2009) also mentioned that these variables are typical of LGD models.

4.1.5. Data Preparation on Independent Variables

Previously, it was mentioned that the choice of the modelling technique, the performance measure and the type of dependent two-sample test will be done by the institution and will depend on the motivation with respect to specific modelling requirements. Simi-

larly, the data preparation will differ from model to model, and in these case studies, the specific choices exercised in terms of the data preparation are only for illustrative purposes.

The typical first step in predictive modelling is data preparation, including the binning of variables. We used a clustering algorithm (SAS Institute 2019) on each variable considered to ensure that each variable was binned using a similar methodology. Among the practical advantages of binning are the removal of the effects of outliers and a way to handle missing values (Verster 2018). Each bin was then quantified to ensure that all types of variables (categorical and numerical) were measured on the same scale. A further motivation to quantify each bin is an alternative to using dummy variables. When we bin the variables, we need to modify the bins, as regression cannot use categorical variables as-is. The default method that is used in regression is using a dummy variable for each class. Expanding the categorical inputs into dummy variables can significantly increase the dimension of the input space (SAS Institute 2010). One alternative is to quantify each bin using the target value (in our case, the LGD value). An example of this in credit scoring is to use the natural logarithm of the good/bad odds (i.e., the weights of evidence). For example, see Lund and Raimi (2012) for a detailed discussion. In our case, we will use the average LGD value in each bin. The main disadvantage of binning and quantification of bins is the loss of information (Lund and Raimi 2012). However, the quantification of the bins has the following advantages:

- Missing values will also be coded as the average LGD value and will therefore be used in model fit (else these rows will not be used in modelling).
- Outliers will have little effect on the fit of the model (as all high values (or all the low values) will have the same LGD value if they are in the same bin).
- Binning can capture some of the generalisation (required in predictive modelling) (Verster 2018).
- Binning can capture possible nonlinear trends (Siddiqi 2006).
- Using the average LGD value for each bin ensures that all variables are of the same scale (i.e., average LGD value). Note that many measures could have been used to quantify each bin, and the average was arbitrarily chosen.
- Using the average LGD value ensures that all types of variables (categorical, numerical, nominal and ordinal) will be transformed into the same measurement type.

For both case studies, two data sets were used for the above binning process:

- Using Data set BB to bin and then applying the binning results to both Data set BB and Data set BT.
- Using Data set Q + BB to bin and then applying the binning results to Data set Q + BB and then to Data set BT.

The results of the binning will not be shown, but the general trend observed when considering the binning will be given. Note that these general trends were observed for both the case studies:

- Seniority code: Senior debt is associated with less risk (lower LGD) than junior debt.
- Guarantee indicator: Debt with a guarantee indicator is associated with less risk (lower LGD values).
- Collateral indicator: Debt with a collateral indicator is associated with less risk (lower LGD values).
- Industry code: Some industries (e.g., mining) are associated with less risk than other industries, e.g., education.
- Type of loan: Some types of loans (e.g., revolver loans) are associated with less risk (lower LGD) than other types of loans, e.g., overdrafts.
- Exposure at default: We used ten equal-sized bins for the EAD. The risk increases as the EAD decreases. This seems counterintuitive, and the reason might be due to the fact that both large corporates and SMEs were included. Typically, large corporates are associated with higher loan amounts but are typically lower risk companies. The loan size of SMEs will usually be smaller but could be associated with a higher risk.

4.1.6. The Multivariate Prediction Accuracy Index (MPAI) as a Potential Measure to Relate Our Result

Taplin and Hunt (2019) recently proposed the Prediction Accuracy Index (PAI) for a setting where risk models are developed on one data set but applied to other/new data. Their focus is on "assessing whether a model remains fit-for-purpose by considering when review data is inappropriate for the model, rather than just different to the development data". The MPAI is defined as the average variance of the estimated mean outcome for the review data divided by the average variance of the estimated mean outcome at development. In our view, such a measure could potentially be applied in our setting to establish whether the external data (translate to review data in the MPAI setting) are representative of the internal data (i.e., development data in the MPAI setting). Both a univariate (PAI) and a multivariate (MPAI) measure were proposed by Taplin and Hunt (2019), and we will be using the MPAI in line with our application of a linear regression model containing several independent variables. From the definition of the MPAI, a value above one occurs when, for the review data, the independent variables have values that result in a variance of the predicted outcome that is higher than the corresponding variance for the development data. For example, a MPAI of 1.5 implies that the variance of the predicted mean response at review is 50% higher than the variance of the mean response at development (on average). In this regard, Taplin and Hunt (2019) proposed the following interpretation for the values of the PAI when applied in their setting: values less than 1.1 indicate no significant deterioration in the predictive accuracy of the model, values from 1.1 to 1.5 indicate a deterioration requiring further investigation and values exceeding 1.5 indicate the predictive accuracy of the model has deteriorated significantly. As such, MPAI values below 1.1 are preferred, as this indicates an almost similar or lower (improved) variance of the average predicted response at review compared to development. Compared to our setting, where we want to assess the representativeness of external data when compared to internal data, large differences in the estimated mean responses (i.e., both a significant improvement and reduction in the average variance of the estimated mean outcome using the external data compared to the internal data) are indicative that the external data are not exhibiting similar characteristics compared to the internal data. In that regard, we propose a "two-sided" threshold (but still related to the one-sided thresholds) for the MPAI when applying the measure in our setting: MPAI values between 0.9 and 1.1 indicate data that are typical to the internal or base data. MPAI values between 0.5 and 0.9 or between 1.1 and 1.5 indicate that the data should be cautiously used, and further investigation should be done. MPAI values between 0 and 0.5 or greater than 1.5 reflect data that are atypical when compared to the internal or base data set. As evident from the proposed thresholds, an ideal MPAI value would be in the order of 1, as it would signify that the external data does not result in a vastly different variance of the estimated mean response.

4.2. Results of Case Study 1

In this first case study, we assess our methodology when investigating whether a pooled data source is representative when considering the enrichment of internal data with pooled data in developing a LGD model for regulatory purposes. The proposed methodology aims at assessing the premise of whether the data set in question (Data set Q) is representative with respect to a base data set (Data set B). First, we define Data set Q and Data set B.

We commence by using the SA GCD data set and applied several exclusions. We exclude all observations before the year 2000 and after 2015, as well as the observations from asset classes other than large corporates and SMEs. We only use observations where the LGD is between -0.01 and 1.5. The summary statistics of this generated SA LGD data set that we use in our modelling are displayed in Table 2.

Going forward, we will consider this SA GCD LGD data set, which contains 3831 observations. Due to the confidential nature of some information in the data set, not all the details can be provided. The difference between the median and average LGD for South

Africa was 26.04%. The large difference between the mean and the median can be explained by the bimodal distribution of the LGD data, which is also positively skewed (see Figure 2). Furthermore, the standard deviation of the LGD variable in the SA GCD data set was 0.4. The LGD values of the SA GCD data set are depicted in Figure 2. The typical distribution of LGD is expected to be a bimodal distribution (GCD 2018; Riskworx 2011).

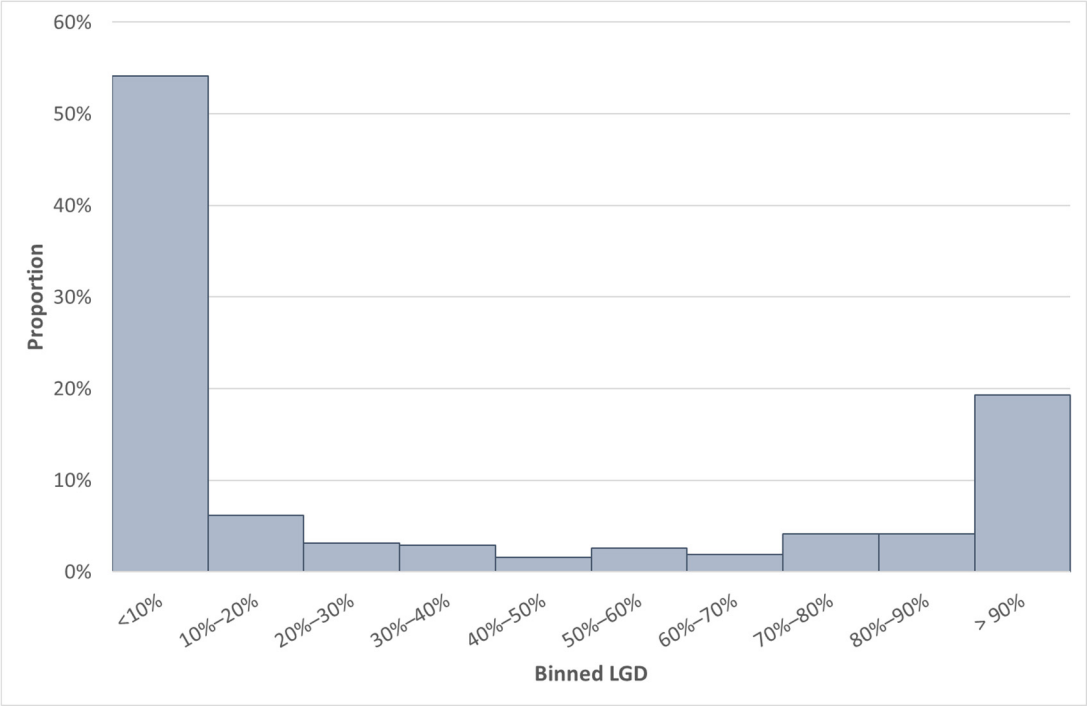

Figure 2. SA GCD LGD distribution.

In the GCD data set we used, it is possible to identify the country of origination of a loan, but the identity of the specific bank is protected[1]. To "synthetically create" a bank's internal data, we used a 25% random sample from the SA GCD data set, representing a hypothetical SA bank, say ABC Bank. This resulted in a sample of 958 observations from the original 3831 observations in the SA GCD LGD data. The remaining 2873 observations will then be regarded as the data set in question (Data set Q).

The ideal situation would have been to use an actual bank's internal data and not just a random sample, as described above. This case study, however, aims to illustrate the methodology. We will assume that ABC Bank is considering building a LGD model with only internal data or augmenting their internal data with the pooled data set created as described above. If ABC Bank wants to augment their internal data with this pooled data when developing a regulatory model, they need to assess whether the pooled data set is representative. ABC Bank's internal data will be the base data (i.e., Data set B), and the SA GCD data will be the data set to be assessed for representativeness (Data set Q). We will illustrate the proposed methodology from the perspective of ABC Bank.

Step 1: Split the base data set (Data set B) into one part for developing a model, namely Data set BB (Base Build), and another part to evaluate the model that was build, say, Data set BT (Base Test).

We split ABC Bank's data set randomly into an 80% build data set (Data set BB with 767 observations) and a 20% test data set (Data set BT with 191 observations). We will use the Data set BB to build a model (Step 2), and we will evaluate the model performance on the test data set (Data set BT) in Step 4. We will build a second model (Step 3) on the augmented data set (but excluding the test data of ABC Bank), i.e., build a model on Data set Q + BB. This results in 3640 observations, either by calculating $3831 - 191 = 3640$ (SA

GCD data set less Data set BT) or 2873 + 767 = 3640 (Data set Q + BB). We will evaluate both models on the test data set of ABC Bank, Data set BT (Step 4), as shown in Figure 3.

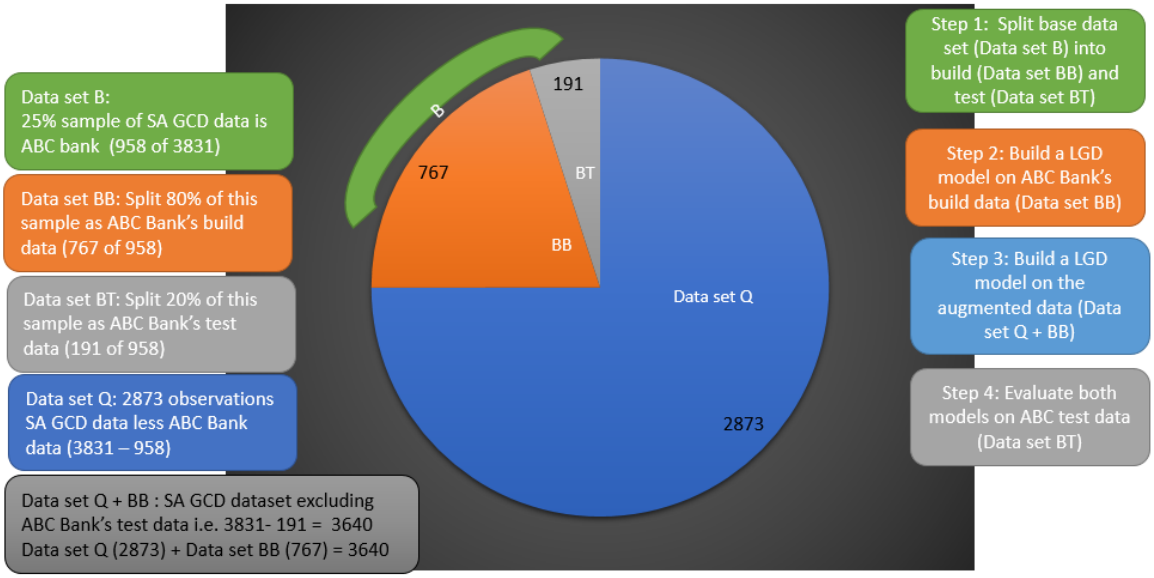

Figure 3. Visual illustration of the data used in case study 1.

Step 2: Build a model on Data set BB (Base Build).

A linear regression model was fitted to the build data set of ABC Bank's data (Data set BB), using LGD as the dependent variable. The underlying assumptions, as stated in Section 4.1, were checked and found satisfactory. Only five of the six predictor variables (discussed above) were statistically significant at a level of 5%. The results are shown in Table 3. One exception was made with the variable Facility type when fitting the model on ABC bank's data. For this variable, the *p*-value was 8%, and the variable was not excluded from the regression, since the literature confirmed that the variable is an important LGD driver. Furthermore, in all other analyses (see Table 4 and Section 4.3), this variable (Facility type) had a *p*-value of less than 1%.

Table 3. Regression results of the ABC build data set (Data set BB).

Parameter Estimates		
Variable (Binned, Average LGD)	**Parameter Estimate**	***p*-Value**
Intercept	1.31	0.02
Facility type	0.41	0.08
Industry code	0.25	0.02
Collateral indicator	0.59	<0.01
Seniority code	−4.89	0.01
Exposure at default	0.81	<0.01
Goodness-of-fit statistics		
R-squared on ABC build data set (Data set BB)		32.67%
Adjusted R-squared on ABC build data set (Data set BB)		32.23%

Table 4. Regression results of the SA GCD data (Data set Q + BB).

Parameter Estimates		
Variable (Binned, Average LGD)	**Parameter Estimate**	***p*-Value**
Intercept	0.59	0.01
Facility type	0.32	<0.01
Guarantee indicator	−0.40	0.05
Industry code	0.40	<0.01
Collateral indicator	0.68	<0.01
Seniority code	−2.52	<0.01
Exposure at default	0.72	<0.01
Goodness-of-fit statistics		
R-squared on SA GCD data (Data set Q + BB)		32.16%
Adjusted R-squared on SA GCD data (Data set Q + BB)		32.05%

The R-squared statistic was 32.68%. This value indicates that 32.68% of the variance in LGD can be explained by the five variables. The adjusted R-squared statistic was 32.15%. R-squared values in these ranges are not uncommon for LGD models, as evident from Loterman et al. (2012), who reported R-squared values for LGD models in the range of 4–43%.

Step 3: Add the data set in question (Q) together with the base data set (BB) and build a model on this augmented data (Data set Q + BB).

Next, a linear regression was fitted on the Data set Q + BB (this is the SA GCD data set excluding the test data set of ABC Bank and contains 3640 observations). The results of this regression are shown in Table 4. All six of the variables are statistically significant at the 5% level. The R-squared statistic was 32.16%, and the adjusted R-squared statistic was 32.05%.

Step 4: Evaluate the model performance (e.g., MSE) of these two models on the base test data set (Data set BT) and determine whether the model performance has improved or is similar using the following construct.

Step 4.1: Define $MSE_{Q+BB,BT}$ as the MSE of Data set BT using the model build on Data set Q + BB and $MSE_{BB,BT}$ as the MSE of Data set BT using the model build on Data set BB.

The models fitted in Steps 2 and 3 were applied on the test data set of ABC Bank (191 observations, Data set BT), and the MSE results on both the build and test data are shown in Table 5. The first subscript of the MSE indicates the development data set, and the second subscript indicates on what data set the MSE was calculated. The MPAI was also calculated for the instances where the development and test samples were not identical (Taplin and Hunt 2019). When the MPAI is calculated for examples with identical data sets used during model development and testing, the MPAI will be equal to one, as evident in Table 5.

Table 5. MSE and MPAI results of case study 1.

	MSE	**MPAI**
$MSE_{BB,BB}$	10.84%	1
$MSE_{BB,BT}$	11.11%	0.86
$MSE_{Q+BB,Q+BB}$	10.70%	1
$MSE_{Q+BB,BT}$	10.78%	1.09

Step 4.2: If $MSE_{Q+BB,BT} < MSE_{BB,BT}$, the model developed on the augmented data has improved the model performance over the model developed only on the base data.

We observe that the $MSE_{Q+BB,BT}$ is indeed smaller than $MSE_{BB,BT}$ and conclude that the model developed on the augmented data has improved the model performance. Optionally, we could have performed Step 5 to confirm the significance of the difference

between $MSE_{Q+BB,BT}$ and $MSE_{BB,BT}$. The above results imply that, when using the augmented data in the development of the model, improved predictive accuracy of the internal observations (Data set BT) resulted. Based on this argumentation, Step 5 is not applicable in case study 1 due to the outcome of Step 4. We can then conclude that it is safe to continue using the SA GCD data for LGD model development and calibration for ABC Bank. This result is to be expected given the manner in which Data set B was constructed for ABC Bank, i.e., a random sample. The purpose of this case study, however, was to provide a step-by-step implementation guide on how a member bank of an external data provider could use our proposed methodology to assess representativeness.

The results from the MPAI indicate that the average variance of the estimated mean outcome at testing when developing the model using only Data set BB is lower than the average variance of the estimated mean outcome at development. Furthermore, the value of 0.86 is just outside our proposed threshold of 0.9 to 1.1, indicating that further investigation is required. This marginal difference between the conclusion drawn from our proposed technique and the MPAI could be expected as the MPAI was developed for a different objective than our methodology. This is potentially an indication that our proposed thresholds for the MPAI might be too strict, and further refinement might be required. On the positive side, the value indicates that the estimated mean variance of the response at testing is lower than at development. For the model developed on the augmented data (Data set Q + BB), the MPAI of 1.09 indicates that the average variance of the estimated mean outcome at testing is higher than at development but within our proposed range of 0.9 to 1.1, validating our conclusion.

4.3. Results of Case Study 2

In this case study, our proposed methodology is demonstrated with a slightly different aim: we consider which countries in the global GCD data set could be used to augment the SA GCD data in the case of LGD modelling.

Step 1: Split the base data set (Data set B) into subsets, with one subset for building a model, namely Data set BB (Base Build), and another subset to evaluate the model that was build, say, Data set BT (Base Test).

Using the methodology as described in Section 3, we will use the SA GCD data set as the base data set (Data set B) split into an 80% build data set (Data set BB with 3065 observations) and a 20% test data set (Data set BT with 766 observations) to build a LGD model on Data set BB. We will then investigate another country on the GCD data set and append this country's data (Data set Q) to the SA data set (Data set BB) and build a LGD model. We will evaluate these two models on Data set BT. We repeated this for all the countries available in the GCD data set. We considered 42 countries after the filters were applied based on a threshold of minimum facilities.

Some summary statistics of Data set BB are shown in Table 6 and the distribution of the GCD LGD value (of Data set BB) in Figure 4.

Table 6. SA Build data statistics (Data set BB).

Variable: LGD	
Number of observations	3065
Difference between the mean and median	25.94%
Standard deviation	0.40

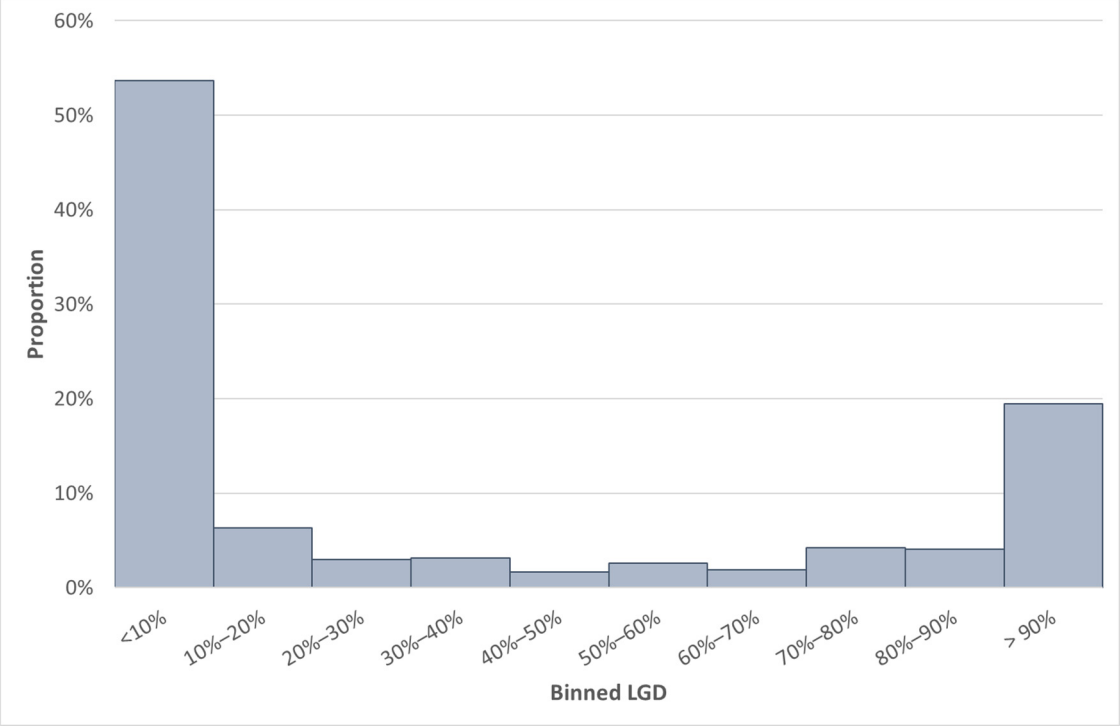

Figure 4. Distribution of LGD of the build data set (Data set BB) used in case study 2.

Step 2: Build a model on Data set BB (Base Build).

A linear regression model was fitted on the SA GCD build data set (Data set BB) after the independent variables were binned, quantified and the resulting binning info was applied to the SA test data set (Data set BT). Once more, all underlying model assumptions were acceptably adhered to. Five of the six variables were statistically significant at the level of 5%. The results are shown in Table 7.

Table 7. Regression results of the build data set of the SA GCD LGD.

Parameter Estimates		
Variable (Binned, Average LGD)	**Parameter Estimate**	***p*-Value**
Intercept	2.24	<0.01
Facility type	0.32	<0.01
Industry code	0.44	<0.01
Collateral indicator	0.70	<0.01
Seniority code	−7.91	<0.01
Exposure at default	0.69	<0.01
Goodness-of-fit statistics		
R-squared on SA build data set		32.51%
Adjusted R-squared on SA build data set		32.40%

The R-squared value was 32.51%, and the adjusted R-squared value was 32.40%. Note that the results of Step 1 of case study 2 are comparable to the results of Step 2 of case study 1, as both used the SA GCD LGD data set. The difference, however, is that, for case study 1, the test data set of ABC Bank was excluded, and for case study 2, the test data set for SA was excluded.

Step 3: Add the data set in question (Q) together with the base data set (BB) and build a model on this augmented data (Data set Q + BB).

Next, a linear regression was fitted on the augmented data set. The augmented data set is the SA GCD build data set (Data set BB) and one country from the GCD data set

(Data set Q). Note that, before fitting a linear regression, the independent variables were binned and quantified on Data set Q + BB, and the resulting binning info was applied to the SA test data set (Data set BT). We repeated this exercise and fitted linear regressions to each of the 42 countries.

We first display the expanded results of two countries: Country L in Table 8 and Country AH in Table 9. The reason for choosing these two countries is because Country L performed the best on the test results and Country AH performed the worst. We then present the abbreviated results of all 42 countries in Table 10.

Table 8. Regression of South Africa GCD build data plus Country L (Data set Q + BB).

Parameter Estimates		
Variable (Binned, Average LGD)	**Parameter Estimate**	***p*-Value**
Intercept	0.28	0.08
Facility type	0.40	<0.01
Industry code	0.46	<0.01
Collateral indicator	0.68	<0.01
Seniority code	−2.12	<0.01
Exposure at default	0.72	<0.01
Goodness-of-fit statistics		
R-squared on SA plus Country L		28.25%
Adjusted R-squared on SA plus Country L		28.14%

Table 9. Regression of South Africa GCD build data plus Country AH (Data set Q + BB).

Parameter Estimates		
Variable (Binned, Average LGD)	**Parameter Estimate**	***p*-Value**
Intercept	−0.94	<0.01
Facility type	0.80	<0.01
Guarantee indicator	0.46	<0.01
Industry code	0.91	<0.01
Collateral indicator	0.79	<0.01
Seniority code	0.66	<0.01
Exposure at default	0.60	<0.01
Goodness-of-fit statistics		
R-squared on SA plus Country AH		7.42%
Adjusted R-squared on SA plus Country AH		7.39%

Considering Country L, the resulting augmented data set contained 3337 observations (3065 from SA build plus 272 from Country L). Five of the six variables were significant at a 5% level. The R-squared value was 28.26% and the adjusted R-squared value 28.14%, as observed in Table 8.

Considering Country AH, the resulting augmented data set contained 21,645 observations (3065 from SA build plus 18,580 from Country AH). All six variables were significant at the 5% level. The R-squared value was 7.42% and the adjusted R-squared value 7.39% (much lower than previous models), as observed from Table 9.

Step 4: Evaluate the model performance (e.g., MSE) of these two models on the base test data set (Data set BT) and determine whether the model performance has improved or is similar.

Step 4.1: Define $MSE_{Q+BB,BT}$ as the MSE of Data set BT using the model build on Data set Q + BB and $MSE_{BB,BT}$ as the MSE of Data set BT using the model build on Data set BB.

The models fitted in Step 2 (one model, Data set BB) and Step 3 (42 models, Data set Q + BB) were applied to the test data set of the SA GCD LGD data (766 observations, Data set BT) in Step 4.

Step 4.2: If $MSE_{Q+BB,BT} < MSE_{BB,BT}$ 0, the model developed on the augmented data has improved the model performance compared to the model developed only on the base data.

The model developed in Step 2 resulted in $MSE_{BB,BT} = 10.64\%$, and not one of the 42 models developed in Step 2 obtained MSE values ($MSE_{Q+BB,BT}$) lower than this. However, many of the $MSE_{Q+BB,BT}$ values were closely related to the $MSE_{BB,BT}$ of the model in Step 2.

Step 5: If $MSE_{Q+BB,BT} \geq MSE_{BB,BT}$, a dependent two-sample test (by comparing residual statistics) is performed to determine whether the model developed on the augmented data has a similar model performance to the model developed only on the base data. If the test concludes that the $MSE_{Q+BB,BT}$ is not statistically different from the $MSE_{BB,BT}$, we deduce that the model developed on the augmented data has a similar model performance than the model developed only on the base data.

In this step, we assess whether the model performances were statistically significantly different from one another. We created paired observations for the absolute error and paired observations for the squared error. The normality assumption for the *t*-test was checked, and both the absolute error and the squared error followed a normal distribution. Using the paired *t*-test, the observations were compared to see if the average errors were statistically different (i.e., *p*-value < 0.05). This was repeated for all 42 countries using both error statistics and is shown in Table 10. We also calculated the test statistics and associated *p*-values for the nonparametric tests (sign test and Wilcoxon signed-rank test). In all cases, similar conclusions followed based on the results of the nonparametric tests, and therefore, the results were omitted from Table 10. Given the results, 12 countries were identified that had a *p*-value of 0.05 and were greater on both the squared error and the absolute error, together with a MPAI value between 0.9 and 1.1. When considering either the squared error or the absolute error, some more countries were identified that could potentially be used to enrich the base data for model development and calibration. The highlighted cells in Table 10 indicate either the absolute error or the squared error or where the MPAI signifies that the models developed on the augmented data (of these countries) have similar model performances compared to the models developed only on the base data. When focusing on MPAI values less than 0.9 (and greater than 1.1), there is an exact correspondence to our methodology where either the squared error or the absolute error has a *p*-value less than 0.05. When changing the direction of comparison by inspecting the *p*-values obtained from our proposed methodology, there are only three countries out of 42 where both the squared error and the absolute error had *p*-values less than 0.05, with a corresponding MPAI value between 0.9 and 1.1 (i.e., conflicting results between our proposed methodology and the MPAI). This marginal difference between our proposed methodology and the MPAI could be expected, as we have already indicated that the MPAI was developed for a different objective than our methodology.

In summary, when considering the last three columns of Table 10, 12 countries were found to have $MSE_{Q+BB,BT}$ that was not statistically different from the $MSE_{BB,BT}$, and we deduced that, for these 12 counties, the model developed on the augmented data (Data set Q + BT) had a similar model performance to the model developed only on the base data (Data set BB).

Table 10. Paired *t*-test results of the 42 countries.

Country	R-Squared (Q + BB)	MSE (Q + BB, Q + BB)	MSE (Q + BB, BT)	*p*-Value of *t*-Test (Squared Error) *	*p*-Value of *t*-Test (Absolute Error) *	MPAI *
Country A	29.45%	10.77%	11.31%	0.04	0.23	0.952
Country B	30.35%	10.98%	11.11%	0.05	0.04	0.974
Country C	25.75%	11.78%	11.32%	<0.01	<0.01	0.69
Country D	28.54%	11.11%	11.05%	0.16	0.01	0.983
Country E	19.03%	10.85%	13.08%	<0.01	<0.01	0.019
Country F	32.02%	10.58%	10.95%	0.71	0.32	0.912
Country G	29.67%	11.07%	11.09%	0.05	0.15	0.983
Country H	29.80%	11.09%	11.14%	0.05	<0.01	0.853
Country I	28.25%	11.22%	11.04%	0.2	0.06	0.9
Country J	29.26%	11.10%	11.19%	0.01	0.01	1.012
Country K	10.56%	14.35%	13.41%	<0.01	<0.01	0.002
Country L	28.25%	11.11%	10.94%	0.75	0.12	0.905
Country M	30.19%	10.84%	11.04%	0.16	0.16	0.925
Country N	28.49%	11.29%	11.13%	0.02	<0.01	0.9
Country O	14.99%	9.15%	13.85%	<0.01	<0.01	0.01
Country P	19.58%	13.45%	13.13%	<0.01	<0.01	0.033
Country Q	21.86%	10.58%	13.18%	<0.01	<0.01	0.105
Country R	27.23%	10.96%	11.21%	0.1	<0.01	0.704
Country S	13.96%	11.45%	13.13%	<0.01	<0.01	0.534
Country T	30.80%	10.62%	11.00%	0.34	0.54	0.914
Country U	28.54%	11.21%	11.10%	0.07	<0.01	0.981
Country V	14.70%	13.67%	12.85%	<0.01	<0.01	0.052
Country W	29.67%	11.04%	10.97%	0.52	0.28	0.956
Country X	26.44%	11.46%	11.20%	0.02	<0.01	0.874
Country Y	31.18%	10.78%	11.01%	0.23	0.63	0.933
Country Z	22.93%	10.37%	13.32%	<0.01	<0.01	0.047
Country AA	28.11%	10.89%	11.05%	0.18	0.08	0.912
Country AB	14.95%	12.53%	13.73%	<0.01	<0.01	0.772
Country AC	16.74%	13.37%	12.83%	<0.01	<0.01	0.099
Country AD	29.88%	11.09%	11.09%	0.1	0.04	1.013
Country AE	15.14%	12.60%	12.92%	<0.01	<0.01	0.515
Country AF	25.02%	11.87%	11.29%	<0.01	<0.01	0.752
Country AG	28.51%	11.36%	11.04%	0.31	<0.01	0.817
Country AH	7.42%	11.98%	14.27%	<0.01	<0.01	0.289
Country AI	26.86%	11.23%	11.09%	0.22	<0.01	0.922
Country AJ	25.59%	10.84%	11.12%	0.12	<0.01	0.712
Country AK	27.74%	12.40%	11.81%	<0.01	<0.01	0.486
Country AL	31.24%	10.74%	10.97%	0.31	0.97	0.985
Country AM	30.61%	10.90%	11.03%	0.08	0.15	0.96
Country AN	28.31%	11.36%	11.15%	0.04	<0.01	1.023
Country AO	29.54%	11.18%	11.03%	0.12	<0.01	0.941
Country AP	31.54%	10.68%	10.97%	0.43	0.34	0.979

* Highlighted cells indicate either the absolute error or the squared error or where the MPAI signifies that the models developed on the augmented data (of these countries) have similar model performances compared to the models developed only on the base data.

To use our methodology for calibration purposes, it is essential to observe the level of the LGD values. We first focus on Country L (the best-performing Country on the test data set). When comparing Table 11 with Table 6 and Figure 4 with Figure 5, we observed that Country L has a mean LGD value that is more than 10% lower compared to the mean SA LGD. However, the median LGDs of these countries are closely related.

Table 11. LGD summary statistics of Country L.

Variable: LGD	
Number of observations	272
Mean	27%
Standard deviation	0.35
Median	9.71%

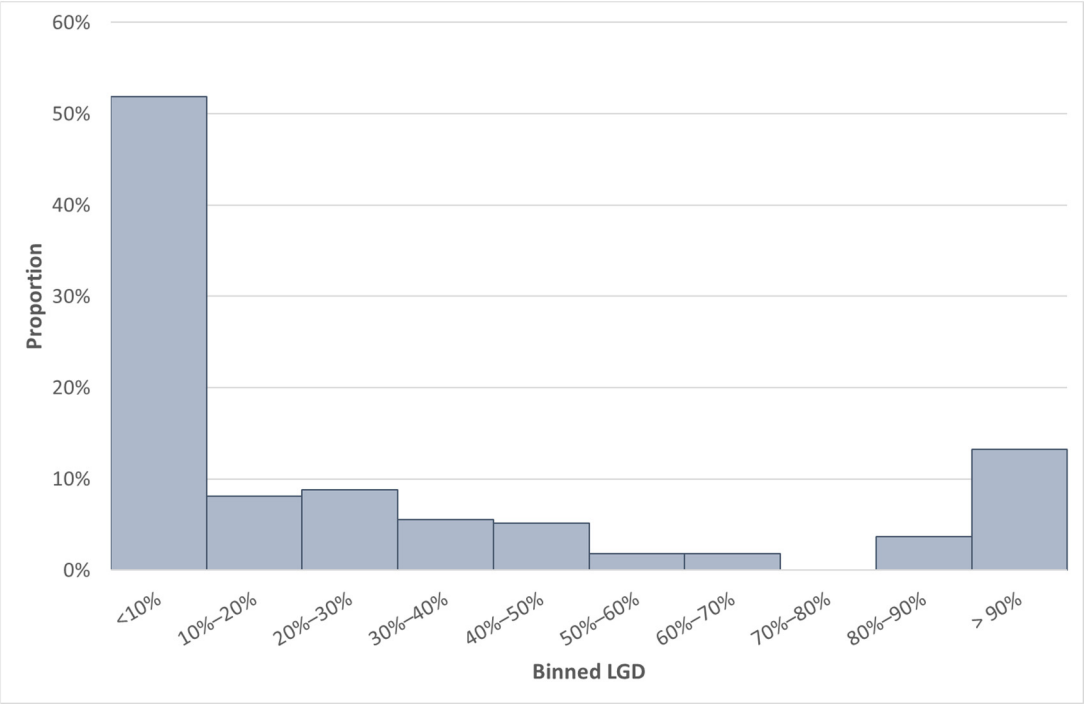

Figure 5. LGD distribution of Country L.

Next, we compare the South African data with the worst-performing MSE. When comparing Table 12 with Table 6 and Figure 4 with Figure 6, we note that Country AH had 18,580 observations. We observed that Country AH has a mean (median) LGD value of 35.3% (7.4%). The mean LGD is lower than the SA mean LGD, but the median LGD are once more closely related.

The summary LGD statistics of all 12 countries that have an absolute and squared error that are not statistically different when compared to the SA data are provided in Table 13.

Table 12. LGD summary statistics of Country AH.

Variable: LGD	
Number of observations	18,580
Mean	28.45%
Standard deviation	0.35
Median	7.47%

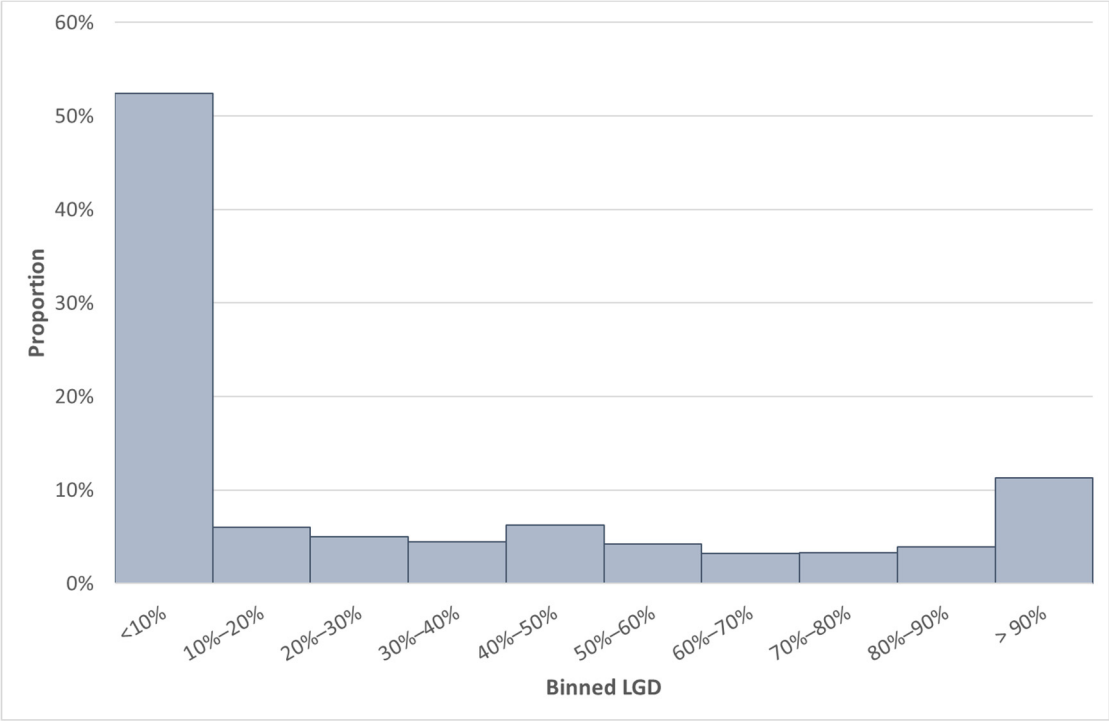

Figure 6. LGD distribution of Country AH.

Table 13. Summary statistics of the LGD of 12 countries.

Country	Mean LGD	Standard Deviation of LGD	Median LGD
Country F	10.56%	0.25	0.54%
Country G	30.31%	0.39	6.71%
Country I	29.19%	0.38	4.28%
Country L	27.00%	0.35	9.71%
Country M	27.48%	0.33	10.94%
Country T	12.61%	0.26	1.08%
Country W	27.37%	0.38	1.96%
Country Y	29.26%	0.36	8.87%
Country AA	21.41%	0.31	3.31%
Country AL	19.87%	0.31	4.82%
Country AM	23.32%	0.37	4.03%
Country AP	17.02%	0.30	3.39%

5. Conclusions and Recommendations

This paper draws together the existing literature on the representativeness of data and classifies these into qualitative and quantitative aspects. Remaining with the quantitative aspects, the paper's main contribution is the development of a novel methodology that utilises model performance to assess the representativeness of data for model development and calibration. We also evaluated our methodology with the MPAI proposed by Taplin and Hunt (2019), although the original purpose of the MPAI was different from our purpose of testing representativeness.

The proposed methodology uses the following premise: if the model developed on the augmented data (Data set Q + BB) has improved or has a similar model performance (on an out-of-sample subset of internal data) when compared with the model developed only on the base data (Data set BB), then Data set Q is not atypical (i.e., unrepresentative) for model development and calibration. This translates into the belief that it is safe to continue using Data set Q for model development and calibration based on the evidence obtained after executing the proposed methodology.

This proposed methodology was illustrated in two case studies. In case study 1, we investigated whether a pooled data source was representative when considering the enrichment of the internal data with pooled data when developing a LGD model for South African large corporates and SMEs. The results showed that the MSE improved when we augmented the "internal" data with the SA GCD data. We conclude that it would be valid to continue using the SA GCD data for model development and calibration.

In case study 2, we investigated which subsets in the pooled data set could be representative when enriching the data for LGD model development. In this case study, following the application of our proposed methodology, we identified the data of 12 countries that are typical to the base (South African) data when considering the absolute error, squared error and MPAI. More countries could be added if either the absolute error or the squared error is used. Based on the results, we suggest that using the data from these countries is valid when enriching the internal data set to model the LGD. Although these case studies are specific to South Africa, our proposed methodology is generic and applicable to settings unrelated to South Africa, rendering it universally applicable. We also expanded on the proposed thresholds of Taplin and Hunt (2019) when using the MPAI in our research setting. In that regard, we propose two-sided thresholds for the MPAI, taking both improvements and reductions in the average variance of the estimated mean outcome into account. The results showed an exceptional overlap in the conclusions drawn from our proposed methodology compared to the MPAI. The benefit of our proposed methodology is that it is founded on the well-known concept of the hypothesis testing of error statistics.

In terms of future research ideas, we propose investigations into the following:

- The modelling technique used to illustrate our proposed methodology was linear regression. Many other modelling techniques could be investigated, and it would be ideal to evaluate the performance of our proposed methodology and the MPAI in a simulation design by fitting different models to simulated data and comparing the outcomes under controlled conditions;
- A similar simulation design could be employed to assess the p-value cut-offs for our proposed methodology and to evaluate the MPAI thresholds proposed by Taplin and Hunt (2019) and those proposed for our setting of assessing representativeness;
- We used a clustering algorithm to bin industries together, although many other methods exist. A future research study could be to bin the industries using other techniques, such as the classification used by Krüger and Rösch (2017).

Author Contributions: Conceptualisation, W.D.S., C.K. and T.V.; formal analysis, C.K., T.V. and W.D.S.; investigation, C.K., T.V. and W.D.S.; methodology, T.V., W.D.S. and C.K.; software, C.K., T.V. and W.D.S.; validation, C.K., T.V. and W.D.S.; visualisation, C.K., T.V. and W.D.S.; writing—original draft, T.V., W.D.S. and C.K. and writing—review and editing, C.K., T.V. and W.D.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors acknowledge that this research idea benefited from the input of Dries de Wet, and the authors would like to extend their appreciation for these valuable contributions to this manuscript. This work was based on research supported in part by the Department of Science and Innovation (DSI) of South Africa. The grant holder acknowledges that the opinions, findings and conclusions or recommendations expressed in any publication generated by DSI-supported research are those of the authors and that the DSI accepts no liability whatsoever in this regard.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of the data; in the writing of the manuscript or in the decision to publish the results.

Note

[1] A loan from a specific country or region can originate from any global bank that submits data to the GCD.

References

Arlot, Sylvain, and Alain Celisse. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys* 4: 40–79. [CrossRef]

Baesens, Bart, Daniel Rosch, and Harald Scheule. 2016. *Credit Risk Analytics: Measurement Techniques, Applications and Examples in SAS*. Hoboken: Wiley & Sons.

Barnard, George Alfred. 1974. Discussion of Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society* 36: 133–35.

BCBS. 2006. *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Basel: Bank for International Settlements. Available online: https://www.bis.org/publ/bcbs128.htm (accessed on 19 January 2018).

Breed, Douw Gerbrand, and Tanja Verster. 2017. The benefits of segmentation: Evidence from a South African bank and other studies. *South African Journal of Science* 113: 1–7. [CrossRef]

Cortés, Lina Marcela, Andres Mora-Valencia, and Javier Perote. 2017. Measuring firm size distribution with semi-nonparametric densities. *Physica A: Statistical Mechanics and its Applications* 485: 35–47. [CrossRef]

Cutaia, Massimo. 2017. *Isn't There Really Enough Data to Produce Good LGD and EAD Models?* Edinburgh: Credit Research Centre, Business School, University of Edinburgh. Available online: https://www.crc.business-school.ed.ac.uk/sites/crc/files/2020-11/17-Massimo_Cutaia.pdf (accessed on 15 March 2019).

D'Agostino, Ralph, and Michael Stephens. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker Inc.

de Jongh, Pieter. Juriaan, Tanja Verster, Elsabe Reynolds, Morne Joubert, and Helgard Raubenheimer. 2017. A Critical Review of the Basel Margin of Conservatism Requirement in a Retail Credit Context. *International Business & Economics Research Journal* 16: 257–74. Available online: https://clutejournals.com/index.php/IBER/article/view/10041/10147 (accessed on 3 December 2020).

Diebold, Francis X. 2015. Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics* 33: 1. [CrossRef]

EBA. 2017. Guidelines on PD Estimation, EAD Estimation and the Treatment of Defaulted Exposures. Available online: https://www.eba.europa.eu/regulation-and-policy/model-validation/guidelines-on-pd-lgd-estimation-and-treatment-of-defaulted-assets (accessed on 18 June 2019).

Engelman, Bernd, and Robert Rauhmeier. 2011. *The Basel II Risk Parameters: Estimation, Validation, and Stress Testing*, 2nd ed. Berlin: Springer. [CrossRef]

European Capital Requirement Regulations. 2013. *Regulation (EU) No 575/2013 of the European Parliament and of the Council*. Luxembourg: Official Journal of the European Union. Available online: https://eur-lex.europa.eu/eli/reg/2013/575/oj (accessed on 5 December 2019).

GCD. 2018. *LGD Report 2018—Large Corporate Borrowers*; Reeuwijk: Global Credit Data. Available online: https://www.globalcreditdata.org/library/lgd-report-large-corporates-2018 (accessed on 21 February 2020).

GCD. 2019. Global Credit Data. Available online: https://www.globalcreditdata.org/ (accessed on 5 December 2019).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Joubert, Morne, Tanja Verster, and Helgard Raubenheimer. 2018a. Default weighted survival analysis to directly model loss given default. *South African Statistical Journal* 52: 173–202. Available online: https://hdl.handle.net/10520/EJC-10cdc036ea (accessed on 5 December 2019).

Joubert, Morne, Tanja Verster, and Helgard Raubenheimer. 2018b. Making use of survival analysis to indirectly model loss given default. *Orion* 34: 107–32. [CrossRef]

Krüger, Steffen, and Daniel Rösch. 2017. Downturn LGD modeling using quantile regression. *Journal of Banking and Finance* 79: 42–56. [CrossRef]

Li, David, Ruchi Bhariok, Sean Keenan, and Stefano Santilli. 2009. Validation techniques and performance metrics for loss given default models. *The Journal of Risk Model Validation* 3: 3–26. [CrossRef]

Loterman, Gert, Iain Brown, David Martens, Christophe Mues, and Bart Baesens. 2012. Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* 28: 161–70. [CrossRef]

Lund, Bruce, and Steven Raimi. 2012. Collapsing Levels of Predictor Variables for Logistic Regression and Weight of Evidence Coding. In *MWSUG 2012: Proceedings*. Paper SA-03. Minneapolis: Midwest SAS Users Group, Inc. Available online: http://www.mwsug.org/proceedings/2012/SA/MWSUG-2012-SA03.pdf (accessed on 19 January 2018).

Mountrakis, Giorgos, and Bo Xi. 2013. Assessing the reference dataset representativeness through confidence metrics based on information density. *ISPRS Journal of Photogrammetry and Remote Sensing* 78: 129–47. [CrossRef]

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. 1996. *Applied Linear Statistical Models*, 4th ed. New York: WCB McGraw-Hill.

OCC. 2011. *Supervisory Guidance on Model Risk Management*; Attachment to Supervisory Letter 11-7. Washington, DC: Board of Governors of the Federal Reserve System. Available online: https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107a1.pdf (accessed on 19 January 2018).

Picard, Richard, and Kenneth Berk. 1990. Data splitting. *The American Statistician* 44: 140–47. [CrossRef]
Prorokowski, Lukasz. 2018. Validation of the backtesting process under the targeted review of internal models: Practical recommendations for probability of default models. *Journal of Risk Model Validation* 13: 109–47. [CrossRef]
Prudential Regulation Authority. 2019. *Internal Ratings Based (IRB) Approaches (Supervisory Statement SS11/13)*; London: Bank of England. Available online: https://www.bankofengland.co.uk/prudential-regulation/publication/2013/internal-ratings-based-approaches-ss (accessed on 21 February 2020).
Ramzai, Juhi. 2020. PSI and CSI: Top 2 Model Monitoring Metrics. Available online: https://towardsdatascience.com/psi-and-csi-top-2-model-monitoring-metrics-924a2540bed8 (accessed on 1 March 2021).
Riskworx. 2011. LGD Distributions. Available online: http://www.riskworx.co.za/resources/LGD%20Distributions.pdf (accessed on 20 February 2020).
SARB. 2015. Bank's Act Reporting. Available online: https://www.resbank.co.za/Lists/News%20and%20Publications/Attachments/6864/07%20Chapter%202%20credit%20risk.pdf (accessed on 19 January 2018).
SAS Institute. 2010. *Predictive Modelling Using Logistic Regression*. Cary: SAS Institute.
SAS Institute. 2019. *The Modeclus Procedure (SAS/STAT 14.3 User's Guide)*. Cary: SAS Institute. Available online: http://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=statug&docsetTarget=statug_modeclus_toc.htm&locale=en (accessed on 2 February 2018).
Sheather, Simon. 2009. *A Modern Approach to Regression with R*. New York: Springer Science & Business Media.
Siddiqi, Naeem. 2006. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken: John Wiley & Sons.
Sprent, Peter, and Nigel C. Smeeton. 2001. *Applied Nonparametric Statistical Methods*. London: Chapman & Hall/CRC.
Taplin, Ross, and Clive Hunt. 2019. The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring. *Risks* 7: 53. [CrossRef]
Thompson, Steven K. 2012. *Sampling*, 3rd ed. Hoboken: Wiley.
Verster, Tanja. 2018. Autobin: A Predictive Approach towards Automatic Binning Using Data Splitting. *South African Statistical Journal* 52: 139–55. Available online: https://hdl.handle.net/10520/EJC-10ca0d9e8d (accessed on 5 June 2020).
Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA's statement on p-values: Context, process and purpose. *The American Statistician* 70: 129–33. [CrossRef]
Yurdakul, Bilal, and Joshua Naranjo. 2020. Statistical properties of the population stability index. *Journal of Risk Model Validation* 14: 89–100. [CrossRef]
Zhang, Yongli, and Yuhong Yang. 2015. Cross-validation for selecting a model selection procedure. *Journal of Econometrics* 187: 95–112. [CrossRef]

CHAPTER 3: CONCLUSIONS

This thesis's **main contribution** was the development of a novel methodology that utilises model performance to assess the representativeness of data for model development and calibration. We also evaluated our methodology with the MPAI proposed by Taplin and Hunt (2019), although the original purpose of the MPAI was different from our purpose of testing representativeness.

The proposed methodology uses the following premise: if the model developed on augmented data has improved or similar model performance (on an out-of-sample subset of internal data) when compared with the model developed only on the internal development data, then the additional dataset is not atypical (i.e. unrepresentative) for model development and calibration.

This proposed methodology was illustrated in two case studies. In the first case study, we investigated whether a pooled data source is representative when considering the enrichment of internal data with pooled data when developing an LGD model for South African large corporates and SMEs. In the second case study, we investigated which subsets in the pooled data set could be representative when enriching data for LGD model development. We have shown that the proposed methodology is able to evaluate external or pooled data for representativeness in both case studies.

Some limitations of the study could be seen as possible future research ideas.

1. The modelling technique used to illustrate our proposed methodology was linear regression. Many other modelling techniques could be investigated as future research ideas e.g. random forests, logistic regression, support vector machines (Engelman & Rauhmeier (2011) and Lessmann, Baesens, Seow, & Thomas (2015)). The underlying assumptions of the chosen modelling technique should be carefully evaluated, and if the assumptions are not met, the specific technique needs to be re-evaluated.
2. Future research could include a simulation design to be employed to assess the p-value cut-offs for our proposed methodology and to evaluate the MPAI thresholds proposed by Taplin and Hunt (2019) and those proposed for our setting of assessing representativeness.
3. The data preparation phase will be different for each model, and in these case studies the specific choices of data preparation are only for illustrative purposes, e.g. we used a clustering algorithm to bin industries together, although many other methods exist (Baesens, Rosch, & Scheule, 2016) that could potentially be investigated.

4. The MPAI was used to compare our proposed methodology against some other measure. An interesting area for future research could be to investigate other measures similar to the MPAI. In the study by du Pisanie, Allison, Budde, & Visagie (2023), they discuss the different population stability testing procedures and note that the MPAI is not well-defined for logistic regression and might give incorrect results in specific scenarios.
5. Definitions for representativeness exist (see e.g. D'Exelle (2014)), however, a formal definition for representativeness in the context of validating whether a larger sample is representative of a smaller sample does not exist. A further research idea could be to formally define and contrast "appropriateness" and "representativeness" specifically for this context of representativeness which can be used as an industry-wide accepted definition.
6. In this proposed study, we illustrated a methodology where the bank's internal data was augmented with external data, but the question arises for cases where no internal data exist. This creates an exciting avenue for future research to determine how a financial institution can validate whether external data is representative if no internal data exist.
7. In this study we focus on improving prediction accuracy and not necessarily on the interpretability of models. It could happen that models built on supplemented data differ significantly (in terms of coefficient estimates and selected explanatory variables) from models built on internal data. In connection with this, future research could include to calibrate the same model (in terms of type and explanatory variables) on both supplemented data and internal data and then comparing the coefficient estimates.
8. Although we propose the MSE as a performance measure, many alternatives exist and could be investigated in future (Neter, Kutner, Nachtsheim, & Wasserman, 1996).
9. In our proposed methodology, the paired t-test based on both the absolute error and the squared error was used. Alternative tests exist (Sprent & Smeeton, 2001) but again the underlying assumptions should be evaluated. Different significance levels (e.g. 5% or 10%) could also be used and will depend on the specific scenario.
10. In this paper there were multiple p-values that resulted from multiple testing. Another area that can be researched in future is the possibility of creating an automated strategy where only representative data sets are selected from a selection of external data sets. If such a strategy can be devised, multiple testing can possibly be avoided.
11. Our target variable used was LGD, but without loss of generality, any other dependent variable could be used, e.g. PD or EAD. These models could be developed for any type of portfolio, e.g. Large Corporates in Germany.

12. A further research idea would be to apply our methodology to an actual data set. Due to confidentiality, we did not have access to a real bank's data. Our methodology could be seen as a step-by-step guide for an institution and a valuable addition would be if this could be applied to real data.

REFERENCES (Chapter 3)

Baesens, B., Rosch, D., & Scheule, H. (2016). Credit Risk Analytics. New Jersey: SAS Institute, Wiley.

D'Exelle, B. (2014). Encyclopaedia of Quality of Life and Well-Being Research. (A. Michalos, Ed.) doi:10.1007/978-94-007-0753-5_2476

du Pisanie, J., Allison, J. S., Budde, C. J., & Visagie, I. J. H. (2023). A Critical Review of Existing and New Population Stability Testing Procedures in Credit Risk Scoring (Preprint). Retrieved April 19, 2023, from arXiv.org e-Print archive: https://arxiv.org/pdf/2303.01227.pdf

Engelman, B., & Rauhmeier, R. (2011). The Basel II Risk Parameters: Estimation, Validation, and Stress Testing (second edition) (2nd ed.). Berlin: Springer.

Lessmann, S., Baesens, B., Seow, H., & Thomas, L. (2015). Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research. European Journal of Operational Research, 247(1), 124-136. https://doi.org/10.1016/j.ejor.2015.05.030.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied Linear Statistical Models (4th ed.). New York: WCB McGraw-Hill.

Sprent, Peter, and Nigel C. Smeeton. 2001. Applied Nonparametric Statistical Methods. London: Chapman & Hall/CRC.

Taplin, Ross, and Clive Hunt. 2019. The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring. Risks 7: 53.

APPENDIX A: AUTHOR GUIDELINES FOR RISKS JOURNAL

Risks

Instructions for Authors

Shortcuts

- Manuscript Submission Overview
- Manuscript Preparation
- Preparing Figures, Schemes and Tables
- Supplementary Materials, Data Deposit and Software Source Code
- Research and Publication Ethics
- Reviewer Suggestions
- English Corrections
- Preprints and Conference Papers
- Authorship
- Editorial Independence
- Conflicts of Interest
- Editorial Procedures and Peer-Review
- Promoting Equity, Diversity and Inclusiveness Within MDPI Journals

Submission Checklist

Please:

1. Read the **Aims & Scope** to gain an overview and assess if your manuscript is suitable for this journal;
2. Use the **Microsoft Word template** or **LaTeX template** to prepare your manuscript;
3. Make sure that issues about **publication ethics**, **copyright**, **authorship**, **figure formats**, **data** and **references format** have been appropriately considered;
4. Ensure that all authors have approved the content of the submitted manuscript.
5. Authors are encouraged to add a **biography** (optional) to the submission and post it to **SciProfiles**.

Manuscript Submission Overview

Types of Publications

Full experimental details must be provided so that the results can be reproduced. *Risks* requires that authors publish all experimental controls and make full datasets available where possible (see the guidelines on **Supplementary Materials** and references to unpublished data).

Manuscripts submitted to *Risks* should neither be published previously nor be under consideration for publication in another journal. The main article types are listed below and a comprehensive list of article types can be found **here.**

- *Article:* These are original research manuscripts. The work should report scientifically sound experiments and provide a substantial amount of new information. The article should include the most recent and relevant references in the field. The structure should include an Abstract, Keywords, Introduction, Materials and Methods, Results, Discussion, and Conclusions (optional) sections, with a suggested minimum word count of 4000 words. Please refer to the journal webpages for specific instructions and templates.
- *Review:* Reviews offer a comprehensive analysis of the existing literature within a field of study, identifying current gaps or problems. They should be critical and constructive and provide recommendations for future research. No new, unpublished data should be presented. The structure can include an Abstract, Keywords, Introduction, Relevant Sections, Discussion, Conclusions, and Future Directions, with a suggested minimum word count of 4000 words.

Submission Process

Manuscripts for *Risks* should be submitted online at **susy.mdpi.com**. The submitting author, who is generally the corresponding author, is responsible for the manuscript during the submission and peer-review process. The submitting author must ensure that all eligible co-authors have been included in the author list (read the **criteria to qualify for authorship**) and that they have all read and approved the submitted version of the manuscript. To submit your manuscript, register and log in to the **submission website**. Once you have registered, **click here to go to the submission form for *Risks***. All co-authors can see the manuscript details in the submission system, if they register and log in using the e-mail address provided during manuscript submission.

Accepted File Formats

Authors are encouraged to use the **Microsoft Word template** or **LaTeX template** to prepare their manuscript. Using the template file will substantially shorten the time to complete copy-editing and publication of accepted manuscripts. The total amount of data for all files must not exceed 120 MB. If this is a problem, please contact the Editorial Office **risks@mdpi.com**. Accepted file formats are:

- *Microsoft Word:* Manuscripts prepared in Microsoft Word must be converted into a single file before submission. When preparing manuscripts in Microsoft Word, we encourage you to use the ***Risks* Microsoft Word template file**. Please insert your graphics (schemes, figures, *etc.*) in the main text after the paragraph of its first citation.
- *LaTeX:* Manuscripts prepared in LaTeX must be collated into one ZIP folder (including all source files and images, so that the Editorial Office can recompile the submitted PDF). When preparing manuscripts in LaTeX, we encourage you to use the ***Risks* LaTeX template files**. You can now also use the online application **writeLaTeX** to submit articles directly to *Risks*. The MDPI LaTeX template file should be selected from the **writeLaTeX template gallery**.
- *Supplementary files:* May be any format, but it is recommended that you use common, non-proprietary formats where possible (see **below** for further details).

Disclaimer: Usage of these templates is exclusively intended for submission to the journal for peer-review, and strictly limited to this purpose and it cannot be used for posting online on preprint servers or other websites.

Free Format Submission

Risks now accepts free format submission:

- We do not have strict formatting requirements, but all manuscripts must contain the required sections: Author Information, Abstract, Keywords, Introduction, Materials & Methods, Results, Conclusions, Figures and Tables with Captions, Funding Information, Author Contributions, Conflict of Interest and other Ethics Statements. Check the Journal Instructions for Authors for more details.
- Your references may be in any style, provided that you use the consistent formatting throughout. It is essential to include author(s) name(s), journal or book title, article or chapter title (where required), year of publication, volume and issue (where appropriate) and pagination. DOI numbers (Digital Object Identifier) are not mandatory but highly encouraged. The bibliography software package *EndNote*, ***Zotero***, *Mendeley*, *Reference Manager* are recommended.
- When your manuscript reaches the revision stage, you will be requested to format the manuscript according to the journal guidelines.

Cover Letter

A cover letter must be included with each manuscript submission. It should be concise and explain why the content of the paper is significant, placing the findings in the context of existing work. It should explain why the manuscript fits the scope of the journal.

Any prior submissions of the manuscript to MDPI journals must be acknowledged. If this is the case, it is strongly recommended that the previous manuscript ID is provided in the submission system, which will ease your current submission process. The names of proposed and excluded reviewers should be provided in the submission system, not in the cover letter.

All cover letters are required to include the statements:

- We confirm that neither the manuscript nor any parts of its content are currently under consideration or published in another journal.
- All authors have approved the manuscript and agree with its submission to (journal name).

Author Biography

Authors are encouraged to add a biography (maximum 150 words) to the submission and post it to **SciProfiles**. This should be a single paragraph and should contain the following points:

1. Authors' full names followed by current positions;
2. Education background including institution information and year of graduation (type and level of degree received);
3. Work experience;
4. Current and previous research interests;
5. Memberships of professional societies and awards received.

Manuscript Preparation

General Considerations

- **Research manuscripts** should comprise:
    - **Front matter**: Title, Author list, Affiliations, Abstract, Keywords.

- **Research manuscript sections:** Introduction, Results, Discussion, Materials and Methods, Conclusions (optional).
- **Back matter:** Supplementary Materials, Acknowledgments, Author Contributions, Conflicts of Interest, References.

- **Review manuscripts** should comprise the **front matter**, literature review sections and the **back matter**. The template file can also be used to prepare the front and back matter of your review manuscript. It is not necessary to follow the remaining structure. Structured reviews and meta-analyses should use the same structure as research articles and ensure they conform to the **PRISMA** guidelines.

- **Graphical Abstract:**
  A graphical abstract (GA) is an image that appears alongside the text abstract in the Table of Contents. In addition to summarizing the content, it should represent the topic of the article in an attention-grabbing way. Moreover, it should not be exactly the same as the Figure in the paper or just a simple superposition of several subfigures. Note that the GA must be original and unpublished artwork. Any postage stamps, currency from any country, or trademarked items should not be included in it.
  The GA should be a high-quality illustration or diagram in any of the following formats: PNG, JPEG, TIFF, or SVG. Written text in a GA should be clear and easy to read, using one of the following fonts: Times, Arial, Courier, Helvetica, Ubuntu or Calibri.
  The minimum required size for the GA is 560 × 1100 pixels (height × width). The size should be of high quality in order to reproduce well.

- **Acronyms/Abbreviations/Initialisms** should be defined the first time they appear in each of three sections: the abstract; the main text; the first figure or table. When defined for the first time, the acronym/abbreviation/initialism should be added in parentheses after the written-out form.

- **SI Units** (International System of Units) should be used. Imperial, US customary and other units should be converted to SI units whenever possible.

- **Equations:** If you are using Word, please use either the Microsoft Equation Editor or the MathType add-on. Equations should be editable by the editorial office and not appear in a picture format.

- **Research Data and supplementary materials:** Note that publication of your manuscript implies that you must make all materials, data, and protocols associated with the publication available to readers. Disclose at the submission stage any restrictions on the availability of materials or information. Read the information about **Supplementary Materials** and Data Deposit for additional guidelines.

- **Preregistration:** Where authors have preregistered studies or analysis plans, links to the preregistration must be provided in the manuscript.

- **Guidelines and standards:** MDPI follows standards and guidelines for certain types of research. See **https://www.mdpi.com/editorial_process** for further information.

Front Matter

These sections should appear in all manuscript types

- **Title:** The title of your manuscript should be concise, specific and relevant. It should identify if the study reports (human or animal) trial data, or is a systematic review, meta-analysis or replication study. Please do not include abbreviated or short forms of the title, such as a running title or head. These will be removed by our Editorial Office.

- **Author List and Affiliations:** Authors' full first and last names must be provided. The initials of any middle names can be added. The PubMed/MEDLINE standard format is used for affiliations: complete address information including city, zip code, state/province, and country. At least one author should be designated as the corresponding author. The email addresses of all authors will be displayed on published papers, and hidden by Captcha on the website as standard. It is the responsibility of the corresponding author to ensure that consent for the display of email addresses is obtained from all authors. If an author (other than the corresponding author) does not wish to have their email addresses displayed in this way, the corresponding author must indicate as such during proofreading. After acceptance, updates to author names or affiliations may not be permitted. Equal Contributions: authors who have contributed equally should be marked with a superscript symbol (†). The symbol must be included below the affiliations, and the following statement added: “These authors contributed equally to this work”. The equal roles of authors should also be adequately disclosed in the author contributions statement. Please read the criteria to qualify for authorship.

- **Abstract:** The abstract should be a total of about 200 words maximum. The abstract should be a single paragraph and should follow the style of structured abstracts, but without headings: 1) Background: Place the question addressed in a broad context and highlight the purpose of the study; 2) Methods: Describe briefly the main methods or treatments applied. Include any relevant preregistration numbers, and species and strains of any animals used. 3) Results: Summarize the article's main findings; and 4) Conclusion: Indicate the main conclusions or interpretations. The abstract should be an objective representation of the article: it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.
- **Keywords:** Three to ten pertinent keywords need to be added after the abstract. We recommend that the keywords are specific to the article, yet reasonably common within the subject discipline.

Research Manuscript Sections

- **Introduction:** The introduction should briefly place the study in a broad context and highlight why it is important. It should define the purpose of the work and its significance, including specific hypotheses being tested. The current state of the research field should be reviewed carefully and key publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work and highlight the main conclusions. Keep the introduction comprehensible to scientists working outside the topic of the paper.
- **Results:** Provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.
- **Discussion:** Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible and limitations of the work highlighted. Future research directions may also be mentioned. This section may be combined with Results.
- **Materials and Methods:** They should be described with sufficient detail to allow others to replicate and build on published results. New methods and protocols should be described in detail while well-established methods can be briefly described and appropriately cited. Give the name and version of any software used and make clear whether computer code used is available. Include any pre-registration codes.
- **Conclusions:** This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.
- **Patents:** This section is not mandatory but may be added if there are patents resulting from the work reported in this manuscript.

Back Matter

- **Supplementary Materials:** Describe any supplementary material published online alongside the manuscript (figure, tables, video, spreadsheets, etc.). Please indicate the name and title of each element as follows Figure S1: title, Table S1: title, etc.
- **Funding:** All sources of funding of the study should be disclosed. Clearly indicate grants that you have received in support of your research work and if you received funds to cover publication costs. Note that some funders will not refund article processing charges (APC) if the funder and grant number are not clearly and correctly identified in the paper. Funding information can be entered separately into the submission system by the authors during submission of their manuscript. Such funding information, if available, will be deposited to FundRef if the manuscript is finally published. Please add: “This research received no external funding” or “This research was funded by [name of funder] grant number [xxx]” and
“The APC was funded by [XXX]” in this section. Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.
- **Acknowledgments:** In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

- **Author Contributions:** Each author is expected to have made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; or have drafted the work or substantively revised it; AND has approved the submitted version (and version substantially edited by journal staff that involves the author's contribution to the study); AND agrees to be personally accountable for the author's own contributions and for ensuring that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and documented in the literature. For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; Methodology, X.X.; Software, X.X.; Validation, X.X., Y.Y. and Z.Z.; Formal Analysis, X.X.; Investigation, X.X.; Resources, X.X.; Data Curation, X.X.; Writing – Original Draft Preparation, X.X.; Writing – Review & Editing, X.X.; Visualization, X.X.; Supervision, X.X.; Project Administration, X.X.; Funding Acquisition, Y.Y.", please turn to the **CRediT taxonomy** for the term explanation. For more background on CRediT, see **here**.
  "**Authorship must include and be limited to those who have contributed substantially to the work. Please read the section concerning the criteria to qualify for authorship carefully**".
- **Data Availability Statement:** In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "**MDPI Research Data Policies**". You might choose to exclude this statement if the study did not report any data.
- **Conflicts of Interest:** Authors must identify and declare any personal circumstances or interest that may be perceived as influencing the representation or interpretation of reported research results. If there is no conflict of interest, please state "The authors declare no conflict of interest." Any role of the funding sponsors in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state "The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results". For more details please see **Conflict of Interest**.
- **References:** A reference list is always arranged alphabetically. All sources are listed by the last names of the authors and listed individually at the end of the manuscript. We recommend preparing the references with a bibliography software package, such as **EndNote**, **ReferenceManager** or **Zotero** to avoid typing mistakes and duplicated references. We encourage citations to data, computer code and other citable research material. If available online, you may use reference style 8. below.
- Citations and References in Supplementary files are permitted provided that they also appear in the main text and in the reference list.

All the references mentioned in the text should be cited in the "Author-Date" format, for example (Woodward 1987), (Schuman and Scott 1987); An author-date citation in running text or at the end of a block quotation consists of the last (family) name of the author, followed by the year of publication of the work in question. In this context, author may refer not only to one or more authors or an institution but also to one or more editors, translators, or compilers. No punctuation appears between author and date. Abbreviations such as ed. or trans. are omitted.

The Reference list should include the full title as recommended by the Chicago style guide. The style file for endnote, MDPI.ens, can be found at **https://mdpi-res.com/data/chicago-mdpi-v2.ens** References should be described as follows depending on the type of work:

- Journal Articles: 1. Author 1, and Author 2. Year. Title of the Article. *Journal Title* 6: 100–10.
- Books and Book Chapters: 2. Author 1, and Author 2. 2008. Book Title, rev. ed. Publication place: Publisher, pp. 154–96. 3. Author 1, and Author 2. 2008. Title of the chapter. In Book Title, 2nd ed. Edited by Editor 1 and Editor 2. Publication place: Publisher, vol. 3, pp. 54–96.
- Unpublished work, submitted work, personal communication: 4. Author 1, and Author 2. Title of Unpublished Work. Journal Title, phrase indicating stage of publication. 5. Author 1 (University, City, State, Country), and Author 2 (Institute, City, State, Country). Year. Personal communication.
- Conference Proceedings: 6. Author 1, Author 2, and Author 3. Year. Title of Presentation. In Title of the Collected Work (if available). Paper presented at Name of the Conference, Location of Conference, Date of Conference. Edited by Editor 1, Editor 2 and Editor 3 (if available). Publication place: Publisher (if available); Abstract Number (optional), Pagination (optional).

- Thesis: 7. Author 1. Year. Title of Thesis. Level of Thesis, Degree-Granting University, City, Country. Date (if available). Identification information (if available).
- Websites: 8. Author 1, and Author 2. Year. Title of the article. *Magazine/Journal Name*. Available online: URL (accessed on Day Month Year). Unlike published works, websites may change over time or disappear, so we encourage you create an archive of the cited website using a service such as **WebCite**. Archived websites should be cited using the link provided as follows: 9. Title of Site. Available online: URL (archived on Day Month Year).

See the **Reference List and Citations Guide** for more detailed information.

Preparing Figures, Schemes and Tables

- File for Figures and Schemes must be provided during submission in a single zip archive and at a sufficiently high resolution (minimum 1000 pixels width/height, or a resolution of 300 dpi or higher). Common formats are accepted, however, TIFF, JPEG, EPS and PDF are preferred.
- *Risks* can publish multimedia files in articles or as supplementary materials. Please contact the editorial office for further information.
- All Figures, Schemes and Tables should be inserted into the main text close to their first citation and must be numbered following their number of appearance (Figure 1, Scheme I, Figure 2, Scheme II, Table 1, *etc.*).
- All Figures, Schemes and Tables should have a short explanatory title and caption.
- All table columns should have an explanatory heading. To facilitate the copy-editing of larger tables, smaller fonts may be used, but no less than 8 pt. in size. Authors should use the Table option of Microsoft Word to create tables.
- Authors are encouraged to prepare figures and schemes in color (RGB at 8-bit per channel). There is no additional cost for publishing full color graphics.

Supplementary Materials, Data Deposit and Software Source Code

MDPI Research Data Policies

MDPI is committed to supporting open scientific exchange and enabling our authors to achieve best practices in sharing and archiving research data. We encourage all authors of articles published in MDPI journals to share their research data. Individual journal guidelines can be found at the journal 'Instructions for Authors' page. Data sharing policies concern the minimal dataset that supports the central findings of a published study. Generated data should be publicly available and cited in accordance with journal guidelines.

MDPI data policies are informed by **TOP Guidelines** and **FAIR Principles**.

Where ethical, legal or privacy issues are present, data should not be shared. The authors should make any limitations clear in the Data Availability Statement upon submission. Authors should ensure that data shared are in accordance with consent provided by participants on the use of confidential data.

Data Availability Statements provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study.

Below are suggested Data Availability Statements:

- Data available in a publicly accessible repository The data presented in this study are openly available in [repository name e.g., FigShare] at [**doi**], reference number [reference number].
- Data available in a publicly accessible repository that does not issue DOIs Publicly available datasets were analyzed in this study. This data can be found here: [link/accession number]

- Data available on request due to restrictions eg privacy or ethical The data presented in this study are available on request from the corresponding author. The data are not publicly available due to [insert reason here]
- 3rd Party Data Restrictions apply to the availability of these data. Data was obtained from [third party] and are available [from the authors / at URL] with the permission of [third party].
- Data sharing not applicable No new data were created or analyzed in this study. Data sharing is not applicable to this article.
- Data is contained within the article or supplementary material The data presented in this study are available in [insert article or supplementary material here]

Data citation:

- [dataset] Authors. Year. Dataset title; Data repository or archive; Version (if any); Persistent identifier (e.g., DOI).

Computer Code and Software

For work where novel computer code was developed, authors should release the code either by depositing in a recognized, public repository or uploading as supplementary information to the publication. The name and version of all software used should be clearly indicated.

Supplementary Material

Additional data and files can be uploaded as "Supplementary Files" during the manuscript submission process. The supplementary files will also be available to the referees as part of the peer-review process. Any file format is acceptable, however we recommend that common, non-proprietary formats are used where possible. For more information on supplementary materials, please refer to **https://www.mdpi.com/authors/layout#_bookmark83**.

Unpublished Data

Restrictions on data availability should be noted during submission and in the manuscript. "Data not shown" should be avoided: authors are encouraged to publish all observations related to the submitted manuscript as Supplementary Material. "Unpublished data" intended for publication in a manuscript that is either planned, "in preparation" or "submitted" but not yet accepted, should be cited in the text and a reference should be added in the References section. "Personal Communication" should also be cited in the text and reference added in the References section. (see also the MDPI reference list and citations style guide).

Remote Hosting and Large Data Sets

Data may be deposited with specialized service providers or institutional/subject repositories, preferably those that use the DataCite mechanism. Large data sets and files greater than 60 MB must be deposited in this way. For a list of other repositories specialized in scientific and experimental data, please consult databib.org or re3data.org. The data repository name, link to the data set (URL) and accession number, doi or handle number of the data set must be provided in the paper. The journal Data also accepts submissions of data set papers. .

References in Supplementary Files

Citations and References in Supplementary files are permitted provided that they also appear in the reference list of the main text.

Research and Publication Ethics

Research Ethics

Research Involving Human Subjects

When reporting on research that involves human subjects, human material, human tissues, or human data, authors must declare that the investigations were carried out following the rules of the Declaration of Helsinki of 1975 (**https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/**), revised in 2013. According to point 23 of this declaration, an approval from the local institutional review board (IRB) or other appropriate ethics committee must be obtained before undertaking the research to confirm the study meets national and international guidelines. As a minimum, a statement including the project identification code, date of approval, and name of the ethics committee or institutional review board must be stated in Section 'Institutional Review Board Statement' of the article.

Example of an ethical statement: "All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of XXX (Project identification code)."

For non-interventional studies (e.g. surveys, questionnaires, social media research), all participants must be fully informed if the anonymity is assured, why the research is being conducted, how their data will be used and if there are any risks associated. As with all research involving humans, ethical approval from an appropriate ethics committee must be obtained prior to conducting the study. If ethical approval is not required, authors must either provide an exemption from the ethics committee or are encouraged to cite the local or national legislation that indicates ethics approval is not required for this type of study. Where a study has been granted exemption, the name of the ethics committee which provided this should be stated in Section 'Institutional Review Board Statement' with a full explanation regarding why ethical approval was not required.

A written informed consent for publication must be obtained from participating patients. Data relating to individual participants must be described in detail, but private information identifying participants need not be included unless the identifiable materials are of relevance to the research (for example, photographs of participants' faces that show a particular symptom). Patients' initials or other personal identifiers must not appear in any images. For manuscripts that include any case details, personal information, and/or images of patients, authors must obtain signed informed consent for publication from patients (or their relatives/guardians) before submitting to an MDPI journal. Patient details must be anonymized as far as possible, e.g., do not mention specific age, ethnicity, or occupation where they are not relevant to the conclusions. A template permission form is available to download. A blank version of the form used to obtain permission (without the patient names or signature) must be uploaded with your submission. Editors reserve the right to reject any submission that does not meet these requirements.

You may refer to our sample form and provide an appropriate form after consulting with your affiliated institution. For the purposes of publishing in MDPI journals, a consent, permission, or release form should include unlimited permission for publication in all formats (including print, electronic, and online), in sublicensed and reprinted versions (including translations and derived works), and in other works and products under open access license. To respect patients' and any other individual's privacy, please do not send signed forms. The journal reserves the right to ask authors to provide signed forms if necessary.

If the study reports research involving vulnerable groups, an additional check may be performed. The submitted manuscript will be scrutinized by the editorial office and upon request, documentary evidence (blank consent forms and any related discussion documents from the ethics board) must be supplied. Additionally, when studies describe groups by race, ethnicity, gender, disability, disease, etc., explanation regarding why such categorization was needed must be clearly stated in the article.

Ethical Guidelines for the Use of Animals in Research

The editors will require that the benefits potentially derived from any research causing harm to animals are significant in relation to any cost endured by animals, and that procedures followed are unlikely to cause offense to the majority of readers. Authors should particularly ensure that their research complies with the commonly-accepted '3Rs [1]':

- Replacement of animals by alternatives wherever possible,
- Reduction in number of animals used, and
- Refinement of experimental conditions and procedures to minimize the harm to animals.

Authors must include details on housing, husbandry and pain management in their manuscript.

For further guidance authors should refer to the Code of Practice for the Housing and Care of Animals Used in Scientific Procedures [2], American Association for Laboratory Animal Science [3] or European Animal Research Association [4].

If national legislation requires it, studies involving vertebrates or higher invertebrates must only be carried out after obtaining approval from the appropriate ethics committee. As a minimum, the project identification code, date of approval and name of the ethics committee or institutional review board should be stated in Section ‘Institutional Review Board Statement’. Research procedures must be carried out in accordance with national and institutional regulations. Statements on animal welfare should confirm that the study complied with all relevant legislation. Clinical studies involving animals and interventions outside of routine care require ethics committee oversight as per the American Veterinary Medical Association. If the study involved client-owned animals, informed client consent must be obtained and certified in the manuscript report of the research. Owners must be fully informed if there are any risks associated with the procedures and that the research will be published. If available, a high standard of veterinary care must be provided. Authors are responsible for correctness of the statements provided in the manuscript.

If ethical approval is not required by national laws, authors must provide an exemption from the ethics committee, if one is available. Where a study has been granted exemption, the name of the ethics committee that provided this should be stated in Section ‘Institutional Review Board Statement’ with a full explanation on why the ethical approval was not required.

If no animal ethics committee is available to review applications, authors should be aware that the ethics of their research will be evaluated by reviewers and editors. Authors should provide a statement justifying the work from an ethical perspective, using the same utilitarian framework that is used by ethics committees. Authors may be asked to provide this even if they have received ethical approval.

MDPI endorses the ARRIVE guidelines (**arriveguidelines.org/**) for reporting experiments using live animals. Authors and reviewers must use the ARRIVE guidelines as a checklist, which can be found at **https://arriveguidelines.org/sites/arrive/files/documents/ARRIVE%20Compliance%20Questionnaire.pdf**. Editors reserve the right to ask for the checklist and to reject submissions that do not adhere to these guidelines, to reject submissions based on ethical or animal welfare concerns or if the procedure described does not appear to be justified by the value of the work presented.

1. NSW Department of Primary Industries and Animal Research Review Panel. Three Rs. Available online: **https://www.animalethics.org.au/three-rs**
2. Home Office. Animals (Scientific Procedures) Act 1986. Code of Practice for the Housing and Care of Animals Bred, Supplied or Used for Scientific Purposes. Available online: **https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_dat a/file/388535/CoPanimalsWeb.pdf**
3. American Association for Laboratory Animal Science. The Scientific Basis for Regulation of Animal Care and Use. Available online: **https://www.aalas.org/about-aalas/position-papers/scientific-basis-for-regulation-of-animal-care-and-use**
4. European Animal Research Association. EU regulations on animal research. Available online: **https://www.eara.eu/animal-research-law**

Research Involving Cell Lines

Methods sections for submissions reporting on research with cell lines should state the origin of any cell lines. For established cell lines the provenance should be stated and references must also be given to either a published paper or to a commercial source. If previously unpublished *de novo* cell lines were used, including those gifted from another laboratory, details of institutional review board or ethics committee approval must be given, and confirmation of written informed consent must be provided if the line is of human origin.

An example of Ethical Statements:

The HCT116 cell line was obtained from XXXX. The $MLH1^{+}$ cell line was provided by XXXXX, Ltd. The DLD-1 cell line was obtained from Dr. XXXX. The DR-GFP and SA-GFP reporter plasmids were obtained from Dr. XXX and the Rad51K133A expression vector was obtained from Dr. XXXX.

Research Involving Plants

Experimental research on plants (either cultivated or wild) including collection of plant material, must comply with institutional, national, or international guidelines. We recommend that authors comply with the **Convention on Biological Diversity** and the **Convention on the Trade in Endangered Species of Wild Fauna and Flora**.

For each submitted manuscript supporting genetic information and origin must be provided. For research manuscripts involving rare and non-model plants (other than, e.g., *Arabidopsis thaliana, Nicotiana benthamiana, Oryza sativa*, or many other typical model plants), voucher specimens must be deposited in an accessible herbarium or museum. Vouchers may be requested for review by future investigators to verify the identity of the material used in the study (especially if taxonomic rearrangements occur in the future). They should include details of the populations sampled on the site of collection (GPS coordinates), date of collection, and document the part(s) used in the study where appropriate. For rare, threatened or endangered species this can be waived but it is necessary for the author to describe this in the cover letter.

Editors reserve the rights to reject any submission that does not meet these requirements.

An example of Ethical Statements:

Torenia fournieri plants were used in this study. White-flowered Crown White (CrW) and violet-flowered Crown Violet (CrV) cultivars selected from 'Crown Mix' (XXX Company, City, Country) were kindly provided by Dr. XXX (XXX Institute, City, Country).

Arabidopis mutant lines (SALKxxxx, SAILxxxx,…) were kindly provided by Dr. XXX , institute, city, country).

Clinical Trials Registration

Registration

MDPI follows the International Committee of Medical Journal Editors (ICMJE) **guidelines** which require and recommend registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication.

Purely observational studies do not require registration. A clinical trial not only refers to studies that take place in a hospital or involve pharmaceuticals, but also refer to all studies which involve participant randomization and group classification in the context of the intervention under assessment.

Authors are strongly encouraged to pre-register clinical trials with an international clinical trials register and cite a reference to the registration in the Methods section. Suitable databases include **clinicaltrials.gov**, **the EU Clinical Trials Register** and those listed by the World Health Organisation **International Clinical Trials Registry Platform**.

Approval to conduct a study from an independent local, regional, or national review body is not equivalent to prospective clinical trial registration. MDPI reserves the right to decline any paper without trial registration for further peer-review. However, if the study protocol has been published before the enrolment, the registration can be waived with correct citation of the published protocol.

CONSORT Statement

MDPI requires a completed CONSORT 2010 checklist and flow diagram as a condition of submission when reporting the results of a randomized trial. Templates for these can be found here or on the CONSORT website (http://www.consort-statement.org) which also describes several CONSORT checklist extensions for different designs and types of data beyond two group parallel trials. At minimum, your article should report the content addressed by each item of the checklist.

Sex and Gender in Research

We encourage our authors to follow the 'Sex and Gender Equity in Research – SAGER – guidelines' and to include sex and gender considerations where relevant. Authors should use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Article titles and/or abstracts should indicate clearly what sex(es) the study applies to. Authors should also describe in the background, whether sex and/or gender differences may be expected; report how sex and/or gender were accounted for in the design of the study; provide disaggregated data by sex and/or gender, where appropriate; and discuss respective results. If a sex and/or gender analysis was not conducted, the rationale should be given in the Discussion. We suggest that our authors consult the full guidelines before submission.

Borders and Territories

Potential disputes over borders and territories may have particular relevance for authors in describing their research or in an author or editor correspondence address, and should be respected. Content decisions are an editorial matter and where there is a potential or perceived dispute or complaint, the editorial team will attempt to find a resolution that satisfies parties involved.

MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Publication Ethics Statement

Risks is a member of the Committee on Publication Ethics (**COPE**). We fully adhere to its **Code of Conduct** and to its **Best Practice Guidelines**.

The editors of this journal enforce a rigorous peer-review process together with strict ethical policies and standards to ensure to add high quality scientific works to the field of scholarly publication. Unfortunately, cases of plagiarism, data falsification, image manipulation, inappropriate authorship credit, and the like, do arise. The editors of *Risks* take such publishing ethics issues very seriously and are trained to proceed in such cases with a zero tolerance policy.

Authors wishing to publish their papers in *Risks* must abide to the following:

- Any facts that might be perceived as a possible conflict of interest of the author(s) must be disclosed in the paper prior to submission.
- Authors should accurately present their research findings and include an objective discussion of the significance of their findings.
- Data and methods used in the research need to be presented in sufficient detail in the paper, so that other researchers can replicate the work.
- Raw data should preferably be publicly deposited by the authors before submission of their manuscript. Authors need to at least have the raw data readily available for presentation to the referees and the editors of the journal, if requested. Authors need to ensure appropriate measures are taken so that raw data is retained in full for a reasonable time after publication.
- Simultaneous submission of manuscripts to more than one journal is not tolerated.
- The journal accepts exact translations of previously published work. All submissions of translations must conform with our policies on translations.
- If errors and inaccuracies are found by the authors after publication of their paper, they need to be promptly communicated to the editors of this journal so that appropriate actions can be taken. Please refer to our policy regarding Updating Published Papers.
- Your manuscript should not contain any information that has already been published. If you include already published figures or images, please obtain the necessary permission from the copyright holder to publish under the CC-BY license. For further information, see the Rights and Permissions page.
- Plagiarism, data fabrication and image manipulation are not tolerated.
    - **Plagiarism is not acceptable** in *Risks* submissions.
      Plagiarism includes copying text, ideas, images, or data from another source, even from your own publications, without giving any credit to the original source.
      Reuse of text that is copied from another source must be between quotes and the original source must be cited. If a study's design or the manuscript's structure or language has been inspired by previous works, these works must be explicitly cited. All MDPI submissions are checked for plagiarism using the industry standard software iThenticate. If plagiarism is detected during the peer review process, the manuscript may be rejected. If plagiarism is detected after publication, an investigation will take place and action taken in accordance with our policies.

- **Image files must not be manipulated or adjusted in any way** that could lead to misinterpretation of the information provided by the original image.
  Irregular manipulation includes: 1) introduction, enhancement, moving, or removing features from the original image; 2) grouping of images that should obviously be presented separately (e.g., from different parts of the same gel, or from different gels); or 3) modifying the contrast, brightness or color balance to obscure, eliminate or enhance some information.
  If irregular image manipulation is identified and confirmed during the peer review process, we may reject the manuscript. If irregular image manipulation is identified and confirmed after publication, we may correct or retract the paper.

Our in-house editors will investigate any allegations of publication misconduct and may contact the authors' institutions or funders if necessary. If evidence of misconduct is found, appropriate action will be taken to correct or retract the publication. Authors are expected to comply with the best ethical publication practices when publishing with MDPI.

Citation Policy

Authors should ensure that where material is taken from other sources (including their own published writing) the source is clearly cited and that where appropriate permission is obtained.

Authors should not engage in excessive self-citation of their own work.

Authors should not copy references from other publications if they have not read the cited work.

Authors should not preferentially cite their own or their friends', peers', or institution's publications.

Authors should not cite advertisements or advertorial material.

In accordance with COPE guidelines, we expect that "original wording taken directly from publications by other researchers should appear in quotation marks with the appropriate citations." This condition also applies to an author's own work. COPE have produced a discussion document on citation manipulation with recommendations for best practice.

Reviewer Suggestions

During the submission process, please suggest five potential reviewers with the appropriate expertise to review the manuscript. The editors will not necessarily approach these referees. Please provide detailed contact information (address, homepage, phone, e-mail address). The proposed referees should neither be current collaborators of the co-authors nor have published with any of the co-authors of the manuscript within the last five years. Proposed reviewers should be from different institutions to the authors. You may identify appropriate Editorial Board members of the journal as potential reviewers. You may suggest reviewers from among the authors that you frequently cite in your paper.

English Corrections

To facilitate proper peer-reviewing of your manuscript, it is essential that it is submitted in grammatically correct English. Advice on some specific language points can be found **here**.

MDPI provides minor English editing by native English speakers for all accepted papers, included in the APC. The APC does not cover extensive English editing. Your paper could be returned to you at the English editing stage of the publication process if extensive editing is required. You may choose to use a paid language-editing service, such as MDPI's **Author Services**, before submitting your paper for publication. If you use an alternative service that provides a confirmation certificate, please send a copy to the Editorial Office. Authors from economically developing countries or nations should consider registration with **AuthorAid**, a global research community that provides networking, mentoring, resources and training for researchers.

Preprints and Conference Papers

Risks accepts submissions that have previously been made available as preprints provided that they have not undergone peer review. A preprint is a draft version of a paper made available online before submission to a journal.

MDPI operates ***Preprints***, a preprint server to which submitted papers can be uploaded directly after completing journal submission. Note that *Preprints* operates independently of the journal and posting a preprint does not affect the peer review process. Check the *Preprints* **instructions for authors** for further information.

Expanded and high-quality conference papers can be considered as articles if they fulfill the following requirements: (1) the paper should be expanded to the size of a research article; (2) the conference paper should be cited and noted on the first page of the paper; (3) if the authors do not hold the copyright of the published conference paper, authors should seek the appropriate permission from the copyright holder; (4) authors are asked to disclose that it is conference paper in their cover letter and include a statement on what has been changed compared to the original conference paper. Risks does not publish pilot studies or studies with inadequate statistical power.

Unpublished conference papers that do not meet the above conditions are recommended to be submitted to the Proceedings Series journals.

Authorship

MDPI follows the International Committee of Medical Journal Editors (**ICMJE**) guidelines which state that, in order to qualify for authorship of a manuscript, the following criteria should be observed:

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Those who contributed to the work but do not qualify for authorship should be listed in the acknowledgments. More detailed guidance on authorship is given by the **International Council of Medical Journal Editors (ICMJE)**.

Any change to the author list should be approved by all authors including any who have been removed from the list. The corresponding author should act as a point of contact between the editor and the other authors and should keep co-authors informed and involve them in major decisions about the publication. We reserve the right to request confirmation that all authors meet the authorship conditions.

For more details about authorship please check **MDPI ethics website**.

Reviewers Recommendation

Authors can recommend potential reviewers. Journal editors will check to make sure there are no conflicts of interest before contacting those reviewers, and will not consider those with competing interests. Reviewers are asked to declare any conflicts of interest. Authors can also enter the names of potential peer reviewers they wish to exclude from consideration in the peer review of their manuscript, during the initial submission progress. The editorial team will respect these requests so long as this does not interfere with the objective and thorough assessment of the submission.

Editorial Independence

Lack of Interference With Editorial Decisions

Editorial independence is of utmost importance and MDPI does not interfere with editorial decisions. All articles published by MDPI are peer reviewed and assessed by our independent editorial boards, and MDPI staff are not involved in decisions to accept manuscripts. When making an editorial decision, we expect the academic editor to make their decision based only upon:

- The suitability of selected reviewers;
- Adequacy of reviewer comments and author response;
- Overall scientific quality of the paper.

In all of our journals, in every aspect of operation, MDPI policies are informed by the mission to make science and research findings open and accessible as widely and rapidly as possible.

Editors and Editorial Staff as Authors

Editorial staff or editors shall not be involved in processing their own academic work. Submissions authored by editorial staff/editors will be assigned to at least two independent outside reviewers. Decisions will be made by other Editorial Board Members who do not have a conflict of interest with the author. Journal staff are not involved in the processing of their own work submitted to any MDPI journals.

Conflicts of Interest

According to The International Committee of Medical Journal Editors, “Authors should avoid entering into agreements with study sponsors, both for-profit and non-profit, that interfere with authors’ access to all of the study’s data or that interfere with their ability to analyze and interpret the data and to prepare and publish manuscripts independently when and where they choose.”

All authors must disclose all relationships or interests that could inappropriately influence or bias their work. Examples of potential conflicts of interest include but are not limited to financial interests (such as membership, employment, consultancies, stocks/shares ownership, honoraria, grants or other funding, paid expert testimonies and patent-licensing arrangements) and non-financial interests (such as personal or professional relationships, affiliations, personal beliefs).

Authors can disclose potential conflicts of interest via the online submission system during the submission process. Declarations regarding conflicts of interest can also be collected via the **MDPI disclosure form**. The corresponding author must include a summary statement in the manuscript in a separate section "Conflicts of Interest" placed just before the reference list. The statement should reflect all the collected potential conflicts of interest disclosures in the form.

See below for examples of disclosures:

Conflicts of Interest: Author A has received research grants from Company A. Author B has received a speaker honorarium from Company X and owns stocks in Company Y. Author C has been involved as a consultant and expert witness in Company Z. Author D is the inventor of patent X.

If no conflicts exist, the authors should state:

Conflicts of Interest: The authors declare no conflicts of interest.

Editorial Procedures and Peer-Review

Initial Checks

All submitted manuscripts received by the Editorial Office will be checked by a professional in-house *Managing Editor* to determine whether they are properly prepared and whether they follow the ethical policies of the journal. Manuscripts that do not fit the journal's ethics policy or do not meet the standards of the journal will be rejected before peer-review. Manuscripts that are not properly prepared will be returned to the authors for revision and resubmission. After these checks, the *Managing Editor* will consult the journals' *Editor-in-Chief* or *Associate Editors* to determine whether the manuscript fits the scope of the journal and whether it is scientifically sound. No judgment on the potential impact of the work will be made at this stage. Reject decisions at this stage will be verified by the *Editor-in-Chief*.

Peer-Review

Once a manuscript passes the initial checks, it will be assigned to at least two independent experts for peer-review. A single-blind review is applied, where authors' identities are known to reviewers. Peer review comments are confidential and will only be disclosed with the express agreement of the reviewer.

In the case of regular submissions, in-house assistant editors will invite experts, including recommendations by an academic editor. These experts may also include *Editorial Board Members* and Guest Editors of the journal. Potential reviewers suggested by the authors may also be considered. Reviewers should not have published with any of the co-authors during the past three years and should not currently work or collaborate with any of the institutions of the co-authors of the submitted manuscript.

Optional Open Peer-Review

The journal operates optional open peer-review: *Authors are given the option for all review reports and editorial decisions to be published alongside their manuscript. In addition, reviewers can sign their review, i.e., identify themselves in the published review reports*. Authors can alter their choice for open review at any time before publication, but once the paper has been published changes will only be made at the discretion of the *Publisher* and *Editor-in-Chief*. We encourage authors to take advantage of this opportunity as proof of the rigorous process employed in publishing their research. To guarantee impartial refereeing, the names of referees will be revealed only if the referees agree to do so, and after a paper has been accepted for publication.

Editorial Decision and Revision

All the articles, reviews and communications published in MDPI journals go through the peer-review process and receive at least two reviews. The in-house editor will communicate the decision of the academic editor, which will be one of the following:

- *Accept after Minor Revisions*: The paper is in principle accepted after revision based on the reviewer's comments. Authors are given five days for minor revisions.
- *Reconsider after Major Revisions*: The acceptance of the manuscript would depend on the revisions. The author needs to provide a point by point response or provide a rebuttal if some of the reviewer's comments cannot be revised. A maximum of two rounds of major revision per manuscript is normally provided. Authors will be asked to resubmit the revised paper within a suitable time frame, and the revised version will be returned to the reviewer for further comments. If the required revision time is estimated to be longer than 2 months, we will recommend that authors withdraw their manuscript before resubmitting so as to avoid unnecessary time pressure and to ensure that all manuscripts are sufficiently revised.
- *Reject and Encourage Resubmission*: If additional experiments are needed to support the conclusions, the manuscript will be rejected and the authors will be encouraged to re-submit the paper once further experiments have been conducted.
- *Reject*: The article has serious flaws, and/or makes no original significant contribution. No offer of resubmission to the journal is provided.

All reviewer comments should be responded to in a point-by-point fashion. Where the authors disagree with a reviewer, they must provide a clear response.

Author Appeals

Authors may appeal a rejection by sending an e-mail to the Editorial Office of the journal. The appeal must provide a detailed justification, including point-by-point responses to the reviewers' and/or Editor's comments using an **appeal form**. Appeals can only be submitted following a "reject and decline resubmission" decision and should be submitted within three months from the decision date. Failure to meet these criteria will result in the appeal not being considered further. The *Managing Editor* will forward the manuscript and related information (including the identities of the referees) to a designated *Editorial Board Member*. The Academic Editor being consulted will be asked to provide an advisory recommendation on the manuscript and may recommend acceptance, further peer-review, or uphold the original rejection decision. This decision will then be validated by the *Editor-in-Chief*. A reject decision at this stage is final and cannot be reversed.

Production and Publication

Once accepted, the manuscript will undergo professional copy-editing, English editing, proofreading by the authors, final corrections, pagination, and, publication on the **www.mdpi.com** website.

Promoting Equity, Diversity and Inclusiveness Within MDPI Journals

Our Managing Editors encourage the Editors-in-Chief and Associate Editors to appoint diverse expert Editorial Boards. This is also reflective in our multi-national and inclusive workplace. We are proud to create equal opportunities without regard to gender, ethnicity, sexual orientation, age, religion, or socio-economic status. There is no place for discrimination in our workplace and editors of MDPI journals are to uphold these principles in high regard.