SNS BANK N.V. UNIVERSITY OF TWENTE.

Estimating the Financial Impact of Data Quality Issues

Rolf de Jong

4 April 2016

IEM – Financial Engineering & Management

Classification	Public
Status	Final

Supervisory committee

University of Twente	Drs. ir. A.C.M. de Bakker
	Dr. B. Roorda
SNS Bank N.V.	M.P. van Leeuwen, MSc

Contact details

Telephone	+31 6 168 262 14
Student number	0210641
Employee number	9322322
Email addresses	r.dejong-2@student.utwente.nl
	rolf.dejong@sns.nl

ASN BANK BLG wonen RegioBank SNS Zwitserleven

Management summary

SNS Bank N.V. considers credit risk to be the most important risk that it faces. Credit risk is mitigated by holding adequate amounts of loan loss provisions and equity capital. The required levels of these types of capital should be accurately determined, because an underestimation can leave the bank insufficiently protected against credit risk while an overestimation is costly, impairing the bank's ability to turn a healthy profit that allows it to stay in business. However, there are data quality issues in the credit data that impair the accurate determination of these capital requirements. Data quality issues can be resolved more effectively once their financial impact is known. The present research describes a model that can estimate such financial impacts. In particular, it can help to answer a number of questions from downstream stakeholders. The questions are:

i. What is the impact of every data quality issue?
ii. Which data quality issues should be resolved first?
iii. What is the progress of the resolution efforts that are underway?
iv. Which records are affected by the data quality issue that is currently being resolved?
v. How does the quality of the credit data evolve over time?

The model that must help answer these question is based around a sensitivity analysis of the capital requirement calculation process, which takes credit data on the residential mortgage portfolio and calculates a number of risk measures, capital requirement measures and income measures. These measures are jointly called financial measures in the present research. The model is shown in Figure 1. It estimates what the credit dataset would look like if there were no data quality issues, and it then determines what the output of the capital requirement calculation process would be if there were no data quality issues. Comparing that output with the actual output from the capital requirement calculation process then reveals the impact of data quality issues. The next page contains a brief overview of how the model is implemented. The model output is presented in an Excel dashboard that is made available on an internal SharePoint site. This dashboard offers three main types of insights:

i. **Prevalence and impact of data quality issues**. A table for every in-scope data quality issue shows its number of occurrences and the impact on a selected number of financial measures.
ii. **Development through time**. A graph plots the prevalence and impact of a data quality issue against time, which reveals the effectiveness of data cleansing efforts.
iii. **Record-level analysis**. A separate tab in the Excel dashboard file holds information on every database record that is affected by a data quality issue, such that a data repair team can instantly find which records need fixing.

Each month, a new dashboard with updated insights is created. Chapter 6 includes a brief analysis of the January 2016 dashboard.

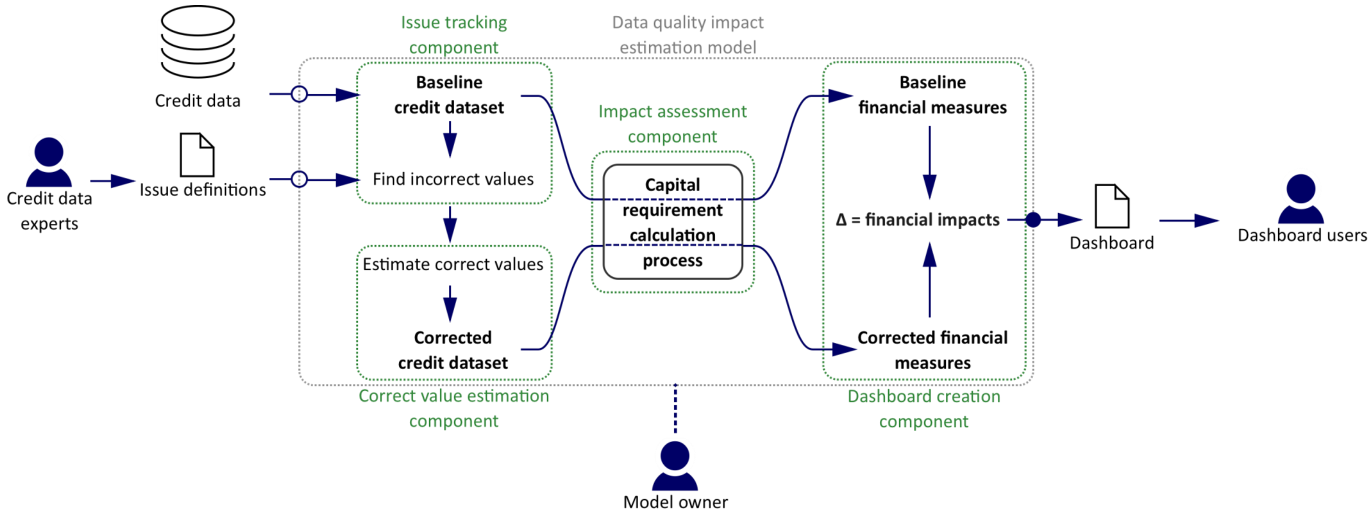

Credit data
Issue definitions
Credit data experts
Issue tracking component
Data quality impact estimation model
Baseline credit dataset
Find incorrect values
Estimate correct values
Corrected credit dataset
Correct value estimation component
Impact assessment component
Capital requirement calculation process
Baseline financial measures
Δ = financial impacts
Corrected financial measures
Dashboard creation component
Dashboard
Dashboard users
Model owner

Figure 1: Model overview. A full-page version of this image is included in Appendix 9.

As mentioned on the previous page, the data quality impact estimation model performs a sensitivity analysis on the capital requirement calculation process. The model is implemented for the greatest part in Matlab and for a smaller part using SQL and Excel. It is split up into four distinct components:

1. The **issue tracking component** gathers the current (baseline) credit data into a single dataset and uses a list of issue definitions to determine which records in the credit data have incorrect values.
2. The **correct value estimation component** uses missing data imputation procedures to estimate the correct values of all affected records in the baseline credit dataset, such that the corrected credit dataset can be constructed. This corrected credit dataset is an estimate of what the 'issue-free' dataset would look like.
3. The **impact assessment component** uses a Monte Carlo simulation approach to calculate the corrected financial measures and compares them to the baseline financial measures. This comparison results in an estimate of the impact of the data quality issues that are being analysed.
4. The **dashboard creation component** gathers the impact information in an Excel dashboard, which makes it easy to analyse and share the impact information.

Table of Contents

Preface

Data is one of the most important assets of a modern bank (Veenman et al., 2015). Large amounts of data are created during customer interactions, generated by internal processes and procured from external parties such as credit rating agencies. It is used for a large variety of activities, such as customer service improvement, product development and credit risk assessment. In the latter, the quality of the data is of special interest (Moges, Dejaeger, Lemahieu, & Baesens, 2013), as credit risk assessments form the basis for regulatory capital calculations as mandated by compliance guidelines such as Basel II and Basel III (BCBS, 2004, 2011a).

The quality of credit risk data is however a current issue at many banks (Roeleven, 2015). This is exemplified by the findings from the recent Asset Quality Review (AQR) conducted by the European Central Bank (ECB) at 130 European banks. The ECB states that "EBITDA and cash flow data for going concern debtors was of conspicuously low quality" and "obtaining collateral information proved difficult for a number of banks" (ECB, 2014, p. 103). Faulty data tends to get into the database due to a variety of causes, such as manual data entry errors, conversion errors after e.g. a merger and real-world changes that are not adequately captured in the database (Moges et al., 2013). The Financial Risk & Modelling department of SNS Bank N.V. has asked me to develop a model that estimates the financial benefit of resolving the bank's most prevalent data issues. Such insights help to prioritise data cleansing efforts, track data quality through time and promote a shared sense of responsibility for correct data from the first line to the bank-office.

The present report documents my attempt to create such a model. Designing, building and implementing it would not have been possible without the invaluable ideas, input and challenging attitude of my company supervisor Paul van Leeuwen and my university supervisors Toon de Bakker and Berend Roorda. Special thanks go to Egbert Simmelink and Floris Harthoorn for maintaining the issue definitions list, for their ideas and feedback during our weekly meetings and for their management of the politics of data quality improvement. And thanks go to the Modelling department for their continued interest in the model and their constant stream of tantalising questions. Finally, thank you Willemijn Cremers for putting up with me during times when I would not be home before 10 PM only to flip open my laptop again and 'finish something up'.

In the foreword to the 2014 Annual Report, Dick Okhuijsen (Chairman of the Board of SNS Bank N.V. at that time) named "raising data quality and data management to a higher level" as a key to improving internal business operations (SNS Bank N.V., 2015a, p. 10). Hopefully, this thesis helps with turning those words into practice.

Rolf de Jong
Amsterdam, 4 April 2016

1 Introduction

SNS Bank N.V. considers credit risk to be the important financial risk that it faces. This risk is a consequence of the retail banking strategy that it pursues. Credit risk is mitigated by ensuring that the bank is adequately capitalised. In practice that means that the bank must hold a loan loss provision and a level of equity capital that are sufficiently sizeable to absorb the expected and unexpected losses that arise from the loans portfolio. Being undercapitalised can make the bank overexposed to credit risk and being overcapitalised is costly, so it is important to accurately assess the amounts of capital that must be held. These amounts are determined using a capital requirement calculation process that takes data on the credit portfolio as its primary input.

The credit data contain data quality issues that may have an adverse effect on the accuracy of the capital requirement calculations. The bank has allocated resources to resolve these data quality issues. These data cleansing efforts could benefit from a model that can estimate the financial impact of the data quality issues that are known to the bank. Such knowledge primarily allows the bank to rank data quality issues by impact, which helps to determine which issues should be resolved first. In addition, this knowledge helps the bank to monitor the progress of data cleansing efforts, develop better credit risk models and promotes a shared responsibility for data quality throughout the bank. This thesis is a documentation of the development and implementation of the aforementioned model. The outline of this thesis is as follows:

Section 1: Introduction outlines the rationale behind the present research and introduces the research question.

Section 2: Stakeholders & Requirements describes the upstream and downstream stakeholders and their requirements.

Section 3: Capital requirement calculation process describes the system under study.

Section 4: Model design describes the model in detail.

Section 5: Performance assesses how the model's most critical components perform.

Section 6: Results displays and analyses the model's outputs.

Section 7: Conclusion & next steps list a number of concrete action that SNS Bank N.V. may consider to further the usage of the model that is developed in the present research.

Section 8: Discussion suggest points of improvement for the model and accompanying research. It also deals with theoretical topics that could not be touched upon in the course of the present research.

1.1 Credit risk is SNS Bank N.V.'s most important financial risk

SNS Bank N.V. is a retail bank with a focus on private individuals and Small and Medium-sized Enterprises (SME) (SNS Bank N.V., 2015a, §1.2). From a financial standpoint, this retail banking strategy involves sourcing capital mainly from shareholders and depositors and lending it out to borrowers, retaining some capital as a reserve (see Figure 2). Depositors supply capital in the form of savings, for which they are compensated with interest paid by the bank. With the exception of long-term deposits, these depositors have the right to withdraw their savings instantly or on short notice. Shareholders supply capital by purchasing shares of equity issued by the bank. They are entitled to dividends and may profit from share value growth, but they have no guarantee that their investment turns into a profit. Borrowers lend from the bank and pay interest over that loan in return. In a retail bank, the majority of loans is extended in the form of residential mortgage loans. The property that is acquired with a mortgage loan serves as collateral and can be seized and foreclosed by the bank if the borrower is not able to repay his loan. The capital that is not lent out to borrowers is retained as a reserve. A part of this reserve is held in the form of cash and cash equivalents and another part is held on a mandatory reserve account at the European Central Bank (ECB, n.d.). The reserve aims to ensure that depositors are able to withdraw cash from their accounts when desired. This banking strategy is called fractional-reserve banking because the amount held in reserve is a fraction of the amount of savings deposited in the bank (Abel & Bernanke, 2001, pp. 520-526).

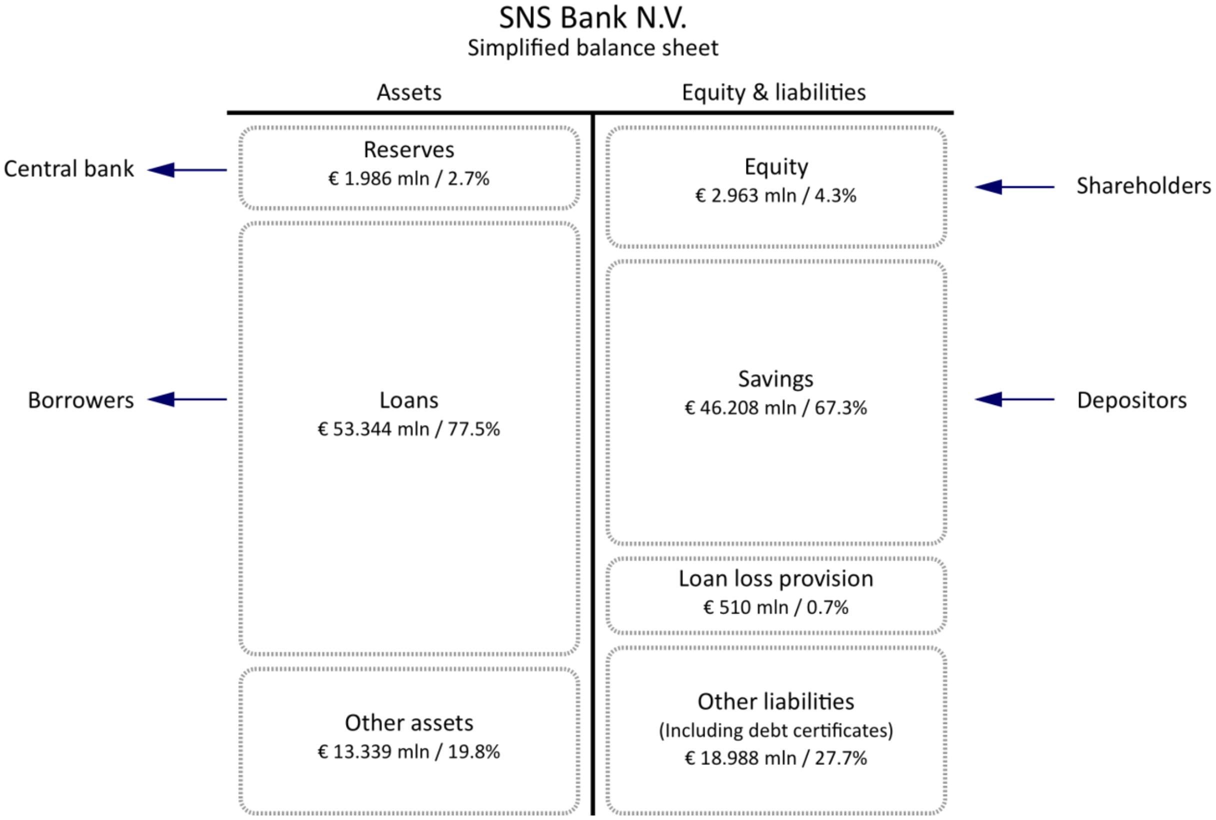


Figure 2: Essential elements of the fractional-reserve balance sheet. (Adapted from Abel & Bernanke, 2001)

To remain viable as a commercial company, a retail bank such as SNS Bank N.V. must turn a net profit in the long term. Apart from attractive marketing and strategic positioning, it can do so by increasing the amount of loans it extends. This increases the total amount of interest received from loans. This is not the only option available, but it is the one wherein a bank has the most flexibly (Fischer, Law, De Demandolx, Broeskamp, & Togashi, 2015). Alternatively, a bank can increase the interest rate it charges to borrowers and decrease the interest rate it offers to depositors. However, in the competitive retail banking market a single bank has only limited room for choosing the interest rates it charges and offers without losing too many customers. Another option is to reduce costs by reducing the number of employees, closing branches and automating processes. Alas, the gains from this option are limited as SNS Bank N.V. positions itself as a bank with a distinct human touch (SNS Bank N.V., 2015a, §1.1), which, apparently, requires a dense network of branches and a customer service with an adequate amount of staff.

When choosing the amount of lending that may be issued, a bank must take into account that an inadequate balance between loans and equity capital can result in unacceptable levels of credit risk. Credit risk is the potential that a borrower will fail to meet his financial obligations in accordance with agreed terms (BCBS, 2000). If this happens and the borrower falls behind on his payments for an extended period of time (90 days at SNS Bank NV), he is said to 'be in default' or 'default on his loan'. Such a default leads to a loss for the bank if the borrower does not catch up on his payments and when the foreclosure value of any property that is posted as collateral is not sufficient to pay off the amount due. Such *loan losses* are charged against the bank's equity capital at the end of the year. If too many borrowers fail in a single year, the cumulative loan losses can completely diminish the bank's equity capital, making the bank in question insolvent. Insolvency can lead to bankruptcy, which is a serious event for a systemically important bank. SNS Bank N.V. notes that it considers credit risk to be the most important financial risk it faces (SNS Bank N.V., 2015a, §5.3.1).

1.2 Credit risk is mitigated by holding adequate amounts of capital

To protect itself against the losses that originate from credit risk, a bank must make sure that it is adequately capitalised. It does so by holding a loan loss provision to cover the losses that it expects to incur in the next 12 months and an adequate amount of equity capital to be able absorb unexpectedly high loan losses. To determine the required levels of the loan loss provision and equity capital, a bank first quantifies the credit risk on each borrower in terms of the risk measures Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD). Those risk measures are then translated to the loss forecast measures Expected Loss (EL) and Unexpected Loss (UL) and a risk impact measure called Risk Weighted Assets (RWA). The bank maintain a loan loss provision at the full EL amount and an equity capital level of at least 15% of the RWA amount (BCBS, 2005b). Each step in this process is discussed into further detail in Chapter 3.

1.3 Capital requirements should be accurately determined

Holding too little capital can make a bank overexposed to credit risk, but holding too much is also suboptimal. Capital is costly and the bank is impacted financially if it holds more capital than it needs, given its risk appetite and the supervisor's capital requirement regulations. Firstly, investors require long-term a Return on Equity (RoE) of around 9.5% in return for the equity capital that they have invested in the bank (SNS Bank N.V. [internal], 2015). Consequently, if the amount of equity capital held by the bank increases then the bank must also reach a higher amount of target profit to keep investors interested. When expressed as a percentage of the amount of equity capital, this target profit is called the Cost of Equity (CoE).

Secondly, there is a cost to holding provisions since these are funded by recognising costs before they are actually incurred. This means that the result of a ceteris paribus year-on-year increase in the loan loss provision requirement is a decrease of the amount of profit that is made in a year. Section 3.3 goes into further detail on the costs of equity and the cost of provisions, as well as the financial measures that affect those costs. Of course, SNS Bank N.V. does not measure its performance solely by the end-year profit, but this measure does allow for an unequivocal quantification of the impacts that are calculated by the model that is described in the present report. Chapter 8 elaborates upon how the model can benefit SNS Bank N.V. and its stakeholders in other ways.

1.4 Data quality issues impair the accurate determination of capital requirements

Credit data is the collection of records in the bank's database that describes the bank's borrowers, including details on their outstanding loans and posted collateral. Credit data is usually presented in tabular form, with one row per month for every borrower. The credit risk models that are used to calculate the PD, LGD and EAD risk measures rely on good quality credit data for making reliable risk assessments. Firstly, because a bank cannot refrain from using its historical credit data during model development since it "must incorporate all relevant, material and available data, information and methods" (BCBS, 2005a, §448; Hanmanth, 2014, p. 27). Secondly, credit risk models are generally developed as statistical regression models. Such models perform better if the data used as input is more complete and accurate. This is colloquially known as the 'Garbage In-Garbage Out'-principle. Unreliable credit risk data can lead to unreliable risk measure estimates.

SNS Bank N.V. has identified data quality issues in the credit data. For the purpose of this research, a data quality issue is defined as a characteristic in the bank's credit data that invalidates a known set of values in a known set of borrower records. An illustration of a data quality issue would be that the effective interest rates for the mortgage loans of a number of borrowers are incorrectly recorded into the credit database. IFRS 9 §B5.5.44 and §B5.5.45 require a bank to discount expected losses on mortgage loans at their effective interest rate (Deloitte; EY, 2014). If the effective interest rate of a mortgage loan is incorrectly registered in the database then the loan's EL and UL estimates will be less inaccurate and/or biased. In summary, the data quality issue impairs the bank's ability to correctly determine its capital requirements on the mortgage loans for the respective set of borrowers. The amount by which a risk measure or capital requirement is misstated due to a data quality issue is hereafter called a financial impact. See Appendix 1 for a list of the data quality issues that have been identified.

1.5 Issues may be resolved more effectively with knowledge on their financial impact

Quantitative insights into the financial impact of data quality issues would benefit the management and resolution of data quality issues. In particular, these insights would enable:

i. Better prioritisation of data cleansing efforts.
ii. Better data quality monitoring.
iii. More credibility when promoting the importance of data quality.

This gives rise to the following research question:

"How can the financial impact of known data quality issues in SNS Bank N.V.'s credit data be estimated?"

1.6 A model that estimates the financial impact of data quality issues

The solution proposed in this report is to develop, implement and put to use a model that takes descriptions of known data quality issues as inputs and produces estimated impacts on financial measures as output. Describing this solution involves answering the sub questions shown below.

1. *What are the model requirements?* (Chapter 2)
   - 1.1. *Who are the relevant stakeholders?*
   - 1.2. *What are the stakeholders' requirements?*
2. *What is the system under study?* (Chapter 3)
3. *How should the model be designed?* (Chapter 4)
   - 3.1. *What framework should be used?* (Section 4.1)
   - 3.2. *What are the components of the model and how should they work?* (Section 4.2)
   - 3.3. *How should the model be implemented?* (Sections 4.3 through 4.6)
4. *What is the model performance?* (Chapter 5)
5. *What is the resulting output of the model?* (Chapter 6)

For the purpose of development, the model uses the IRB-scored part of the residential mortgage portfolio as input data. This portfolio accounts for the majority of SNS Bank N.V.'s credit outlays. However, it should be possible to adapt the model to other data sources, such as the SME loans portfolio, in a later stage.

2 Stakeholders & Requirements

The model that is proposed in Section 1.6 should be embedded in an environment of input data supplies and output data users. Without a continually expanding and improving list of data quality issue definitions, the model cannot be trusted to provide an up-to-date picture of the credit data's quality. And the impact assessments produced by the model will only have use if there are users that leverage this information to actually fix errors in the credit data. Furthermore, a model owner is needed to ensure that the model is continually monitored, maintained and improved after the initial development has been completed. In brief, the present research recognises the following main stakeholders in the direct model environment (also see Figure 3):

i. Credit data experts
ii. Model owner
iii. Dashboard users

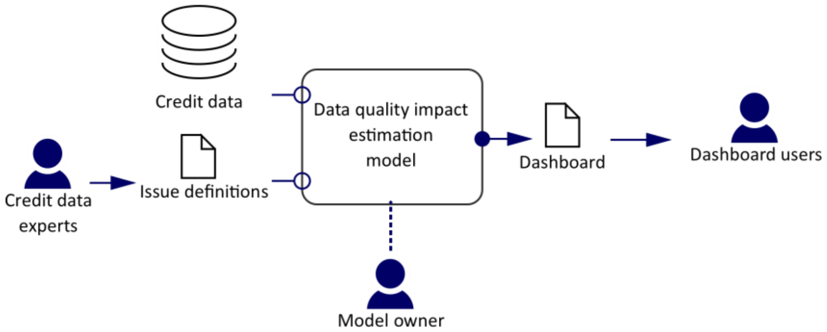


Figure 3: Users in the direct model environment.

These users seek to answer the following questions about the data quality of the credit data:

i. What is the impact of every data quality issue?
ii. Which data quality issues should be resolved first?
iii. What is the progress of the resolution efforts that are underway?
iv. Which records are affected by the data quality issue that is currently being resolved?
v. How does the quality of the credit data evolve over time?

3 Capital requirement calculation process

The system under study is the capital requirement calculation process. This process uses credit data to calculate a number of risk measures and capital requirement measures. The latter measures can be further translated to income measures. This chapter walks through the capital requirement calculation process and shows how the income measures can be calculated from the capital requirement measures. Figure 4 is a schematic representation of this process. All terminology in the figure is elaborated upon in this chapter. Table 1 shows the financial measures that are in scope.

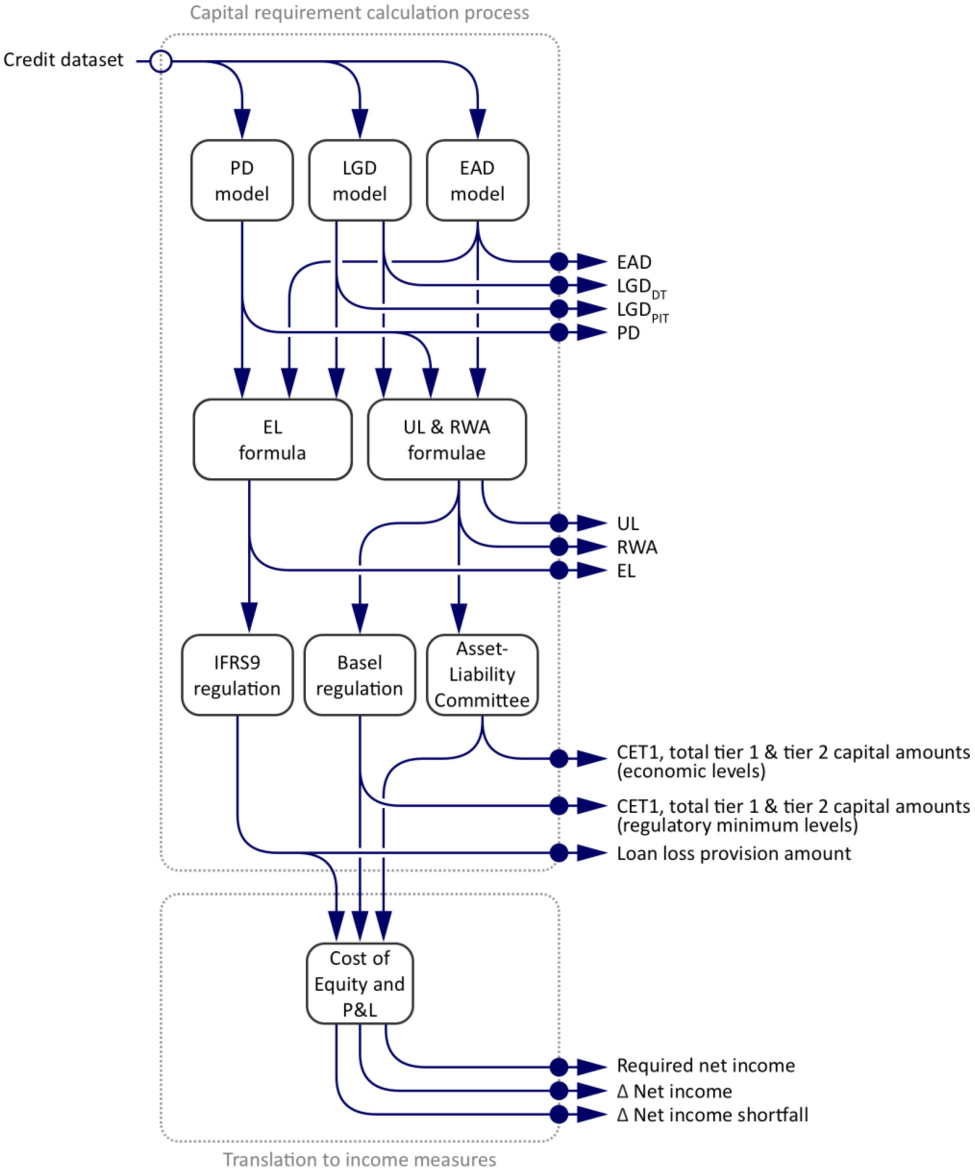

Capital requirement calculation process
Credit dataset
PD model
LGD model
EAD model
EAD
LGD_DT
LGD_PIT
PD
EL formula
UL & RWA formulae
UL
RWA
EL
IFRS9 regulation
Basel regulation
Asset-Liability Committee
CET1, total tier 1 & tier 2 capital amounts (economic levels)
CET1, total tier 1 & tier 2 capital amounts (regulatory minimum levels)
Loan loss provision amount
Cost of Equity and P&L
Required net income
Δ Net income
Δ Net income shortfall
Translation to income measures

Figure 4: Overview of the capital requirement calculation process.

Table 1: Financial measures that are in scope.

Risk measures	Capital requirement measures	Income measures
• Probability of Default (PD) • Exposure At Default (EAD) • Loss Given Default (LGD) • Expected Loss (EL) • Unexpected Loss (UL) • Risk Weighted Assets (RWA)	• Regulatory capital amounts ○ Common equity tier 1 ○ Total tier 1 ○ Tier 2 capital • Economic capital amounts ○ Common equity tier 1 ○ Total tier 1 ○ Tier 2 capital • Loan loss provision amount	• Net income • Required net income • Net income shortfall

3.1 Determining risk measures

SNS Bank N.V. first quantifies the credit risk on the set of loans held by every individual borrower along the risk measures Probability of default (PD), exposure at default (EAD) and loss given default (LGD). These measures are calculated used credit risk models.

i. **Probability of default** is the probability that a borrower goes into default within the next 12 months.

ii. **Exposure at default** is the part of the borrowed sum that is not yet repaid at the moment of a default.

iii. **Loss given default** is the loss that the bank expects to sustain when a default occurs. LGD is usually expressed as a fraction of EAD. In practice, a bank must use two different versions of LGD risk measure in the expected loss (EL) and unexpected loss (UL) formulae that are introduced in the next paragraph. The EL formula requires the Point-in-Time (PiT) version ($\mathrm{LGD}_{\mathrm{PiT}}$), which is calculated based on the current economic environment. The UL formula requires the Downturn version ($\mathrm{LGD}_{\mathrm{DT}}$), which is calculated by assuming a downturn economic environment.

Basel II allows banks to either use prescribed models or internally developed credit risk models to determine their PD, EAD and LGD risk measures. A bank that uses the prescribed models is said to apply the Standard Approach (SA). A bank that uses internally developed models is said to apply the Internal Ratings Based (IRB) approach. Such bespoke credit risk models may be used if the bank can substantiate that these models play an "essential role in [the bank's] credit approval, risk management, internal capital allocation and corporate governance function" (Allen & Overy, 2014; BCBS, 2005a, §444; 2006). The premise is that using internally developed credit risk models can lead to more accurately determined provision and equity capital requirements. That accuracy is rewarded with the prospect of milder minimum provision and equity capital requirements than when a bank uses the standard models that are prescribed by Basel II.

The PD, EAD and LGD risk measures are used to determine two types of loss forecasts: Expected Loss (EL) and Unexpected Loss (UL). UL is further translated to Risk Weighted Assets (RWA).

i. **Expected loss** is the average level of credit loss that a bank expects to incur on the loans extended to a borrower during the normal course of business. There will always be a probability of borrowers going into default and thereby incurring a loss to the bank. The losses that make up the total amount of expected loss tend to be of the high-frequency low-severity type (Hull, 2012). The expected loss that a bank expects to incur on a borrower in the next 12 months is determined as follows:

$$\mathrm{EL} = \mathrm{PD} \cdot \mathrm{LGD}_{\mathrm{PiT}} \cdot \mathrm{EAD} \tag{1}$$

(BCBS, 2005b)

ii. **Unexpected loss** is the level of loss that is incurred by a bank above the expected level of loss. Whereas expected losses are charged against the loan loss provision, unexpected losses are charged against Profit & Loss, directly reducing the amount of equity capital held by a bank. Furthermore, unexpected losses tend to be of the low-frequency high-severity type, which makes them hard to predict but potentially very disruptive (Hull, 2012, p. 436). The UL amount is determined using a more complex formula based on Merton's model. For brevity, it is deferred to Appendix 2.

$$\mathrm{UL} = f(\mathrm{PD}, \mathrm{LGD}_{\mathrm{DT}}, \mathrm{EAD}) \tag{2}$$

(BCBS, 2005b)

The EL and UL formula are designed such that when summing up the EL and UL for every loan in a portfolio, the result equals the Value-at-Risk (VaR) on the loan portfolio. VaR is the total amount of credit loss from the portfolio in a year that a bank expects not to exceed with at a certain confidence level α. In the Basel II Accords, $\alpha = 99.9\%$. An underlying assumption is that the total credit loss incurred on the portfolio within a year follows the probability distribution similar to the one illustrated in Figure 5.

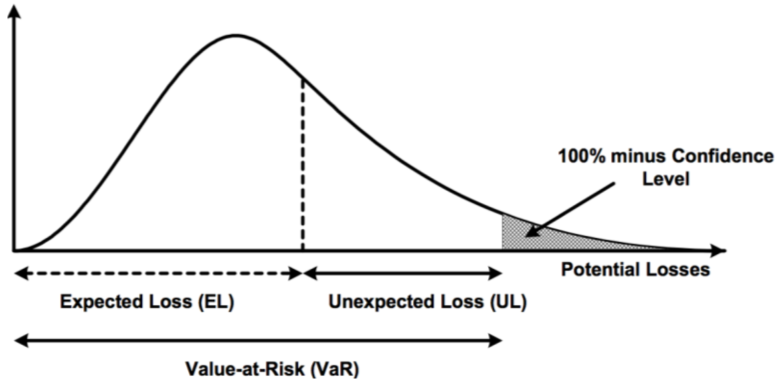


Figure 5: Assumed shape of the total credit loss probability distribution in the Basel II EL and UL formulae.
(BCBS, 2005b, p. 3)

iii. **Risk weighted assets** measures for the amount of assets on a bank's balance sheet, weighted according to risk instead of book value. It is determined as follows:

$$\mathrm{RWA} = \mathrm{UL} \cdot \mathrm{MoC} \cdot \mathrm{C} \cdot 12.5 \tag{3}$$

(BCBS, 2005b)

This formula comprises two calibration factors, which are used by supervisors to fine-tune the RWA figure. MoC is the Margin of Conservatism calibration factor. This factor is applied to account for model risks in the credit risk models used by individual banks. In particular, the MoC is intended to compensate for difficult to quantify risk that surround the credit risk models, such as credit data quality risks, model risks, misinterpretation risks and process risks. C is an additional calibration factor of 1.06 that all banks must apply. This factor was introduced when it was found that most banks would underestimate the RWA figure without the calibration factor *C* (BCBS, 2005a, §44).

3.2 Determining capital requirement measures

A bank uses the EL an RWA measures to determine the capital requirements that it must adhere to. These requirements are meant to make the bank resilient against both expected losses and unexpected losses. Broadly speaking, a bank must hold four partly overlapping types of capital: a loan loss provision, common equity tier 1 capital, total tier 1 capital and tier 2 capital (BCBS, 2011a; 2011b, §52-58).

i. **Loan loss provisions** cover expected losses. When a loan turns out to be loss-making, this loss is first charged to the loan loss provision that is recognised on the balance sheet (EY, 2014). The required amount of this provision is regulated by the IAS 39 standard, which is developed by the International Accounting Standards Board (IASB) in London, UK. The more stringent IFRS 9 standard will supersede the IAS 39 standard on 1 January 2018. Since the SNS Bank N.V. is already working towards the implementation of IFRS 9, the model follows the IFRS 9 standard when needed. In a global survey by Deloitte among 59 major banks, most respondents expect loan loss provisions to increase by as much as 50% when IFRS 9 has been implemented compared to the situation under IAS 39 (Rhys & Mickeler, 2015).
ii. **Common Equity Tier 1 capital** covers unexpected losses. When the losses on loans incurred in a financial year exceed the loan loss provision amount, the losses are charged to the profit & loss statement directly. This reduces the amount of equity held by the bank. A very large amount of loan losses in a year can completely diminish a bank's equity, which in turn can lead to insolvency and bankruptcy. To prevent such an event from happening, banks are required by the Basel II and Basel III Accords to hold a certain minimum level of high-quality equity capital. The highest quality type of equity capital is called 'common equity tier 1' (CET1) capital. CET1 capital excludes preferred shares and non-controlling interest because these types of equity cannot be sufficiently depended upon by the bank during financial turmoil. be depended on by a bank equally well in times of financial turmoil. For instance, the preferred shares that many banks issue entitle the shareholder to a fixed dividend. This debt-like characteristic makes preferred shares more attractive to conservative investors such as pension funds, but the mandatory dividend payments can become a burden for a bank in times of large credit losses.
iii. **Additional tier 1 capital and tier 2 capital** cover further unexpected losses. CET1 capital plus additional tier 1 capital is total tier 1 capital. Additional tier 1 capital and tier 2 capital impose less stringent conditions on the types of equity or even debt that may be recognised as capital. This allows a bank to further improve its resilience to unexpected losses by attracting 'cheaper' types of capital such as preferred shares in addition to the relatively expensive CET1 capital.

The minimum amounts of equity capital that a bank must hold are outlined in Pillar I of the Basel II Accords, which are developed by the Bank for International Settlements (BIS) in Basel, Switzerland. These levels of capital are often referred to as regulatory capital levels, since they are required by regulation. The proportion between the level of certain type of capital and the bank's total RWA is called a capital ratio. As a basic rule, a bank must in any event maintain a CET1 capital ratio of at least 4.5%, a Total Tier 1 capital ratio of at least 6% and a Tier 2 capital ratio of at least 8%.

However, the European Banking Authority (EBA), which supervises the capital adequacy of the European banking sector, requires banks to maintain more conservative, internally determined capital ratios (EBA, 2013, CRD IV, Title VII, Chapter 4, §128-142). Table 2 outlines the regulatory and current capital ratios for each class of equity capital as currently held by SNS Bank N.V.

Table 2: Types of equity capital and the applicable capital requirements. (BCBS, 2011a; 2011b, §52-58; SNS Bank N.V., 2015a, 2015b)

Equity capital class	Regulatory capital ratios	Current capital ratios (2015Q4)
Common equity tier 1 capital	4.5% (€ 604 M)	25.7% (€ 3,450 M)
Total tier 1 capital (CET1 + additional tier 1 capital)	6.0% (€ 805 M)	25.7% (€ 3,450 M)[1]
Tier 2 capital (total tier 1 capital + tier 2 capital)	8.0% (€ 1,074 M)	30.0% (€ 4,027 M)[2]
Capital ratios are expressed as a percentage of RWA and in absolute amounts. Numbers are based on the total RWA reported the half-year financial report of 2015, which was € 13,423 M.		

3.3 Determining income measures

Attracting and retaining capital in the form of provisions and equity is costly since provision increases are changed against net income and since investors expect a return on their investment in equity. This paragraph unifies these costs in a single measure called the 'net income shortfall'.

i. **Net income** is impacted by changes in the loan loss provisions. A bank must recognise a loan loss provision for the amount of credit loss that is expects to incur in the next twelve months. The provision amount is increased or decreased by respectively debiting or crediting the profit & loss account of a bank. In other words, a ceteris paribus increase in provisions results in a lower net income for the year and a ceteris paribus decrease in provisions means a higher net income for the year.

ii. **Required net income** is impacted by changes in the equity capital requirements. Shareholders provide a bank's equity capital by acquiring equity shares. The return on these shareholders' investment is called the 'return on equity' (RoE), which is the net income in a year divided by the average amount of equity capital held by the bank (Investopedia). Net income is partially retained by the bank for reinvestment and partially distributed to shareholders in the form of dividend. Shareholders are, in principle, not guaranteed to receive a return on their investment. However, investors will not provide their capital if the bank cannot give them the prospect of a certain level of return. Therefore, there is an implicit cost to attracting and holding equity capital. This expected level of return is called the 'cost of equity' (CoE).

[1] SNS Bank N.V. held no additional tier 1 capital in 2015Q4.

[2] Includes the € 500 M Tier 2 subordinated debt issue of October 2015.

iii. **Net income shortfall** equals required net income less net income. The CoE is not a cost in the accounting sense of the word; it is rather a benchmark RoE that a company must aim to generate in order the attract investors. When the RoE is lower than the CoE, there is a 'RoE shortfall' (Daniel, Denis, & Naveen, 2008). RoE, CoE and RoE shortfall are expressed relative to the average amount of equity capital held by the bank during a year. The absolute versions of the RoE shortfall is the net income shortfall.

$$\mathrm{RoE} = \frac{\text{Net income}}{\text{Average equity capital level}} \tag{4}$$
$$\mathrm{CoE} = \frac{\text{Required net income}}{\text{Average equity capital level}}$$

$$\begin{aligned}\text{RoE shortfall} &= \mathrm{CoE} - \mathrm{RoE} \\ \text{Net income shortfall} &= (\mathrm{CoE} - \mathrm{RoE}) \cdot \text{Average equity capital leve} \\ &= \text{Required net income} - \text{net income}\end{aligned} \tag{5}$$

A positive net income shortfall is not necessarily detrimental to investor confidence (i.e. the investors' willingness to invest), as long as the shortfall is expected to diminish in the future through an increased RoE or decreased CoE. The net income shortfall can be seen as the additional amount of net income a bank should generate to continue attracting investors and it is thus a measure of income adequacy.

3.4 Summary

The 'Capital requirement calculation process' section has explained that credit risk models calculate the PD, LGD and EAD risk measures for every record in a loans portfolio. These measures are uses to calculate the EL and UL and RWA measures for the portfolio. These measures, together with regulatory and economic capital ratios, determine the loan loss provision and equity capital amounts that the bank should hold to be able to withstand expected and unexpected credit losses. The costs of holding these amounts can be expressed using the net income, required net income and net income shortfall measures.

The main financial impacts from changes in risk measures can be summarised as visualised in Figure 6. In this figure, an *f* in a circle means that a relation is governed by a formula. A number indicates a multiplication factor for the effect of one measure on another. Each of these financial measures is potentially important to internal and external stakeholders, since a bank's financial governance is not only based on net income adequacy but also on the bank's capital adequacy and risk appetite (Babel et al., 2012).

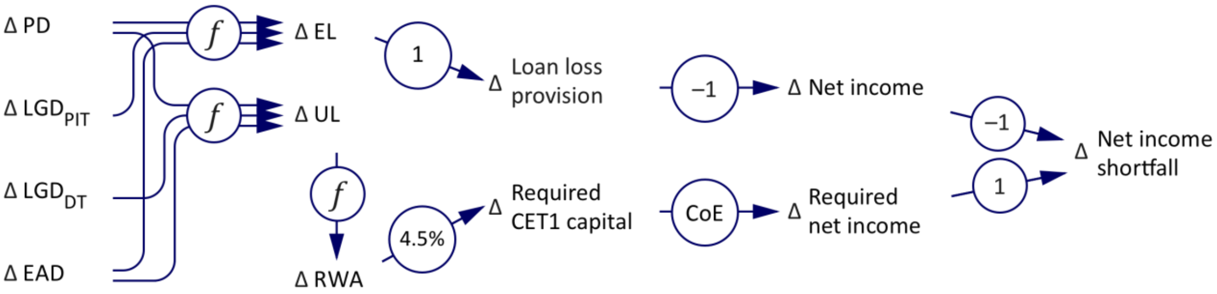


Figure 6: Interdependency of the impacts on financial measures.

4 Model design

The model is designed as a sensitivity analysis on the capital requirement calculation process. Such an analysis studies how changes in the input variables of a system result in changes in the output variables of that system, whereby the system under study is the capital requirement calculation process. That process has been described in Chapter 3. In this chapter, Section 4.1 first shows how a sensitivity analysis works. Section 4.2 then introduces the four model components that make up the implementation of the model. Section 4.3 through 4.6 discuss every component in detail.

4.1 Sensitivity analysis

A sensitivity analysis model studies how changes in the input variable values of a system[3] result in changes in the output variable values of that system (Oakley, 2004). For the present data quality impact estimation model, it is of interest how the financial measures change when the credit dataset is replaced by a corrected, 'issue-free' credit dataset. Calculating that change involves running the capital requirement calculation process twice; once with the current, baseline credit dataset as input and once with the corrected credit dataset as input. Both runs result in a set of financial measures and the difference between these two sets represents the financial impact of the data quality issues that were removed from the baseline credit dataset. This concept is depicted in Figure 7.

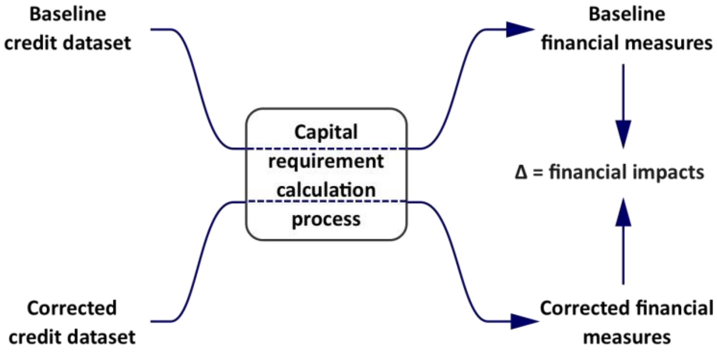

Baseline credit dataset
Capital requirement calculation process
Corrected credit dataset
Baseline financial measures
Δ = financial impacts
Corrected financial measures

Figure 7: The sensitivity analysis approach.

The baseline credit dataset is readily available but the corrected credit dataset is not. If it were available, then there would be no data quality issues in the first place. Moreover, it is unfeasible to resolve all data quality issues in the baseline credit dataset to obtain the corrected credit dataset. Instead, the corrected credit dataset is estimated using statistical procedures.

[3] Literature often speaks of 'a model' instead of 'a system' when discussing sensitivity analysis. However, using the word 'model' at this point would lead to confusion with the impact estimation model that is described in this report. Furthermore, 'variables' are also referred to as 'parameters' in the context of sensitivity analysis.

Example

As application example of a sensitivity analysis, assume the following fictive formula for calculating the expected loss on a borrower's loan. This formula is referred to as 'the system'. $\mathbf{X}$ is the set of input variables, namely the loan's maturity and exposure and the borrower's income. $\mathbf{Y}$ comprises only one output variable, namely the expected loss on the borrower's loan.

$$\begin{aligned}
\mathbf{X} &= (\text{Maturity}, \text{Exposure}, \text{Income}) \\
\mathbf{Y} &= \text{EL} \\
\mathbf{Y} &= f(\mathbf{X}) \\
\text{EL} &= \frac{\text{Exposure}}{\sqrt{\text{Maturity}}} \cdot \max\left(0\,, 1 - \frac{0.5 \cdot \text{Income} \cdot \text{Maturity}}{\text{Exposure}}\right)
\end{aligned}$$

Now assume that for a certain borrower, $\mathbf{X} = (4\ \text{years}, €100000, €25000)$. Evaluating the system with these input values yields an expected loss of $\mathbf{Y} = €25000$. However, say that there is a known data quality issue that results in a misstated maturity value. The model estimates that the maturity of this borrower's loan is actually 6 years instead of 4 years, i.e. $\widehat{\mathbf{X}} = (6\ \text{years}, €100000, €25000)$. These estimates of the correct values yield $\widehat{\mathbf{Y}} = €10206$ and thereby an estimated impact $\Delta\mathbf{Y} = \mathbf{Y} - \widehat{\mathbf{Y}} = €25000 - €10206 = €14794$. In other words, the estimated EL impact of the data quality issue that affects the maturity variable is $€14794$ for this borrower.

4.2 The four model components

The model has been implemented in four components that each deal with a part of the sensitivity analysis. Figure 8 shows that the corrected credit dataset is obtained by taking the baseline credit dataset, determining which values have been affected by a data quality issue and then replacing those values by correct value estimates.

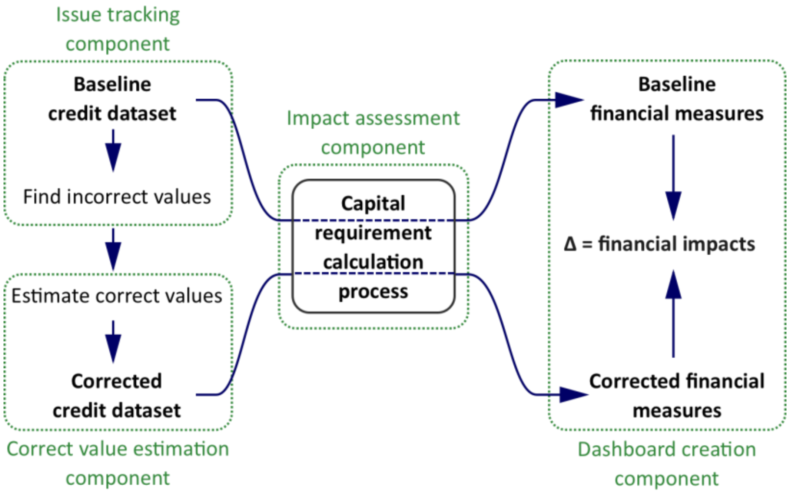

Issue tracking component
Baseline credit dataset
Find incorrect values
Estimate correct values
Corrected credit dataset
Correct value estimation component
Impact assessment component
Capital requirement calculation process
Baseline financial measures
Δ = financial impacts
Corrected financial measures
Dashboard creation component

Figure 8: The four model components each cover a part of the sensitivity analysis.

1. The **issue tracking component** gathers the current credit data into a single dataset and uses a list of issue definitions to determine which records in the resulting baseline credit dataset have incorrect values.
2. The **correct value estimation component** uses missing data imputation procedures to estimate the correct values of all affected records in the baseline credit dataset, such that the corrected credit dataset can be constructed.
3. The **impact assessment component** uses a Monte Carlo simulation approach to calculate the corrected financial measures and compares them to the baseline financial measures. A Monte Carlo approach is used because the correct value estimates from the previous component are stochastic values, whereas the system only accepts deterministic values.
4. The **dashboard creation component** gathers the impact information in an Excel dashboard, which makes it easy to analyse and share the impact information.

Figure 3 can be combined with Figure 8 to reveal an overview of the full data quality impact estimation model and how it is embedded in the model environment, as shown in Figure 9. The remainder of this chapter goes into further detail on every individual model component.

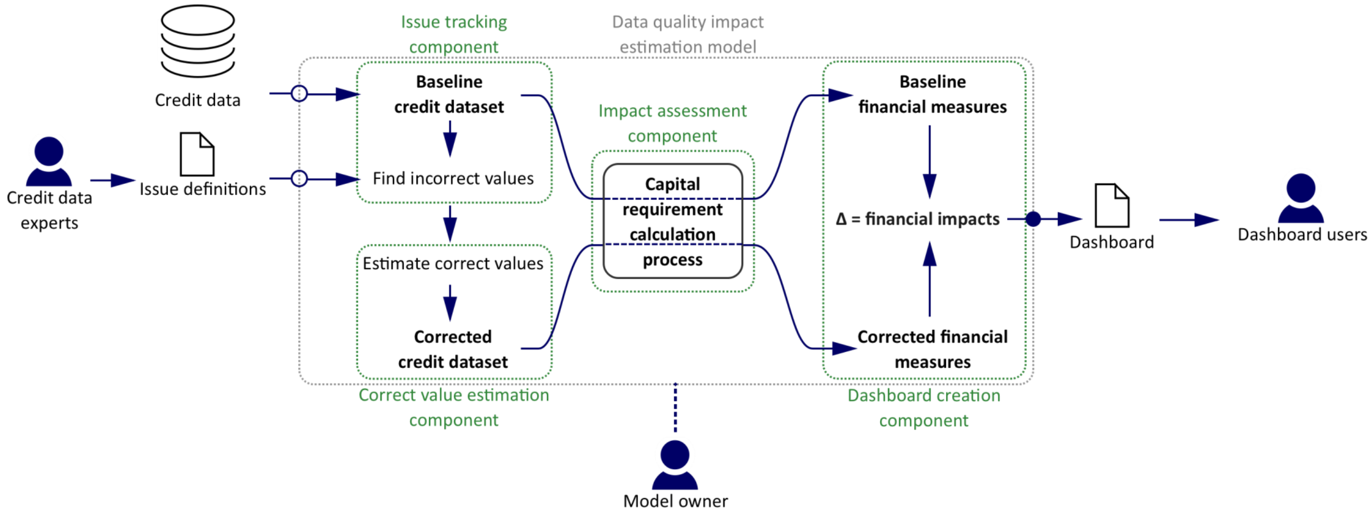

Credit data
Credit data experts
Issue definitions
Issue tracking component
Data quality impact estimation model
Baseline credit dataset
Find incorrect values
Estimate correct values
Corrected credit dataset
Correct value estimation component
Impact assessment component
Capital requirement calculation process
Baseline financial measures
Δ = financial impacts
Corrected financial measures
Dashboard creation component
Dashboard
Dashboard users
Model owner

Figure 9: The model embedded in the model environment.

4.3 Issue tracking component

The issue tracking component takes the list of issue definitions and a connection to the credit database. It creates or updates the baseline credit dataset and uses the issue definitions to compile a list of occurrences of the data quality issues in the baseline credit dataset.

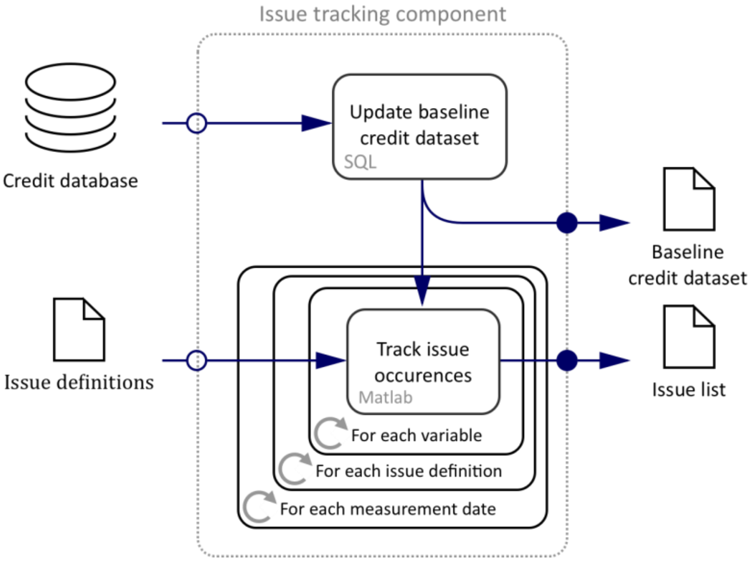


Figure 10: Schematic overview of the issue tracking component.

4.3.1 Implementation

The issue tracking component carries out its task in the two steps shown in Figure 10. These steps are explained in more detail in the paragraphs that follow.

1. **Update baseline credit dataset**. The credit data is collected in a set of tables in a dedicated credit risk database. For the model architecture it is however more suitable to use a single table as input. That is why the component first collects all credit data in a single table on a separate database.
2. **Track issue occurrences**. In this step, the component uses the issue definitions list to search for occurrences of data quality issues in the baseline credit dataset. These occurrences are recorded in the issue list.

Example (continued)

The actual baseline credit dataset contains too many variables to use it for the purpose of illustration. As an alternative, consider the simplified set of customer, loan and collateral data from a fictive mortgage loan portfolio shown in the table below. Suppose that the bank is aware of a (also fictive) data quality issue in the credit data that affects the income variable for several borrowers. The data quality issue description reads as follows:

```
Variable 'Income' for borrower N is 'Below minimum' if 'Income < 10000'
```

Borrower ID	Income	Exposure	Maturity	Foreclosure value
1	€ 6,000	€ 180,000	10 years	€ 120,000
2	€ 21,000	€ 30,000	5 years	€ 120,000
3	€ 12,000	€ 100,000	16 years	€ 80,000
4	€ 4,000	€ 40,000	2 years	€ 100,000
5	€ 50,000	€ 375,000	15 years	€ 400,000
6	€ 51,000	€ 180,000	6 years	€ 200,000
7	€ 70,000	€ 50,000	1 year	€ 500,000

Clearly, borrowers 1 and 4 match this description, and therefore their values for the income variables are considered incorrect. This information is recorded in the issue list:

Borrower ID	Issue name	Variable name	Incorrect value
1	Below minimum	Income	€ 6,000
4	Below minimum	Income	€ 4,000

4.3.2 Updating the baseline credit dataset

[Omitted in the public version]

4.3.3 Tracking issue occurrences

The second input for this component is the data quality issue definitions list, or issue definitions for short. Among other fields, this list contains a field called ‘SQL filter’ (see also Appendix 1). This field contains a SQL-formatted filter that can be used to find the records in the baseline credit dataset that have been affected by a data quality issue. The issue tracking component uses these filters to create the issue list. It is essentially a map to the incorrect values in the baseline credit dataset, containing the following information:

i. The primary keys of the affected record. Primary keys are the set of variables whose values can be used to uniquely identify any record in the dataset.
ii. The name of the affected variable in the affected record.
iii. The name of data quality issue of which an occurrence has been found.
iv. The incorrect value of the affected variable in the affected record.

Every item in this list is called an ‘occurrence’ of a data quality issue. The issue list does not contain the correct value of the affected variables because these correct values cannot be readily observed. See Figure 11 for an example of how a SQL query is used to build a part of the issue list. The issue tracking component runs this query for every in-scope permutation of measurement date (‘peil_dt’), variable name and issue name. The results from running these queries are combined to form one complete issue list.

`SELECT  MeasurementDate as MeasurementDate`
`      , BorrowerId      as BorrowerId`
	Primary keys. These variables uniquely identify a record in the baseline credit dataset.

`      , 'LTFV'          as VariableName`
`      , 'Above maximum' as IssueName`
	Records that are found using this SQL query have a value on the ‘LTFV’ variable that has been affected by the ‘Above maximum’ data quality issue.

`      , LTFV            as IncorrectValue`	Stores the value as currently registered in the credit risk dataset.
`FROM    CreditDataTable`	Reference to the baseline credit risk dataset.
`WHERE   MeasurementDate = '2016-01-31'`	Scope settings. In this case, only records pertaining to active, residential borrowers on measurement date 2016-01-31 are analysed.
`  AND   LTFV > 150`	SQL filter for finding the records that are affected by respective data quality issue.

Figure 11: Example of a SQL query for building a part of the issue list.

4.4 Correct value estimation component

This component takes the issue list and the baseline credit dataset as inputs. It uses correct value estimation procedures to estimate the correct value of the affected variable in each in occurrence that is recorded in the issue list. When done, the component outputs the estimates list, which is essentially the issue list augmented with correct value estimates. §4.4.1 shows how the component is implemented. §4.4.2 briefly introduces the concept of missing data imputation, which forms the basis for the correct value estimation procedures that are implemented in this component. §4.4.3 deals with measurement levels, which is a preparation for the explanation of the correct value estimation procedures that have been implemented in the component, in §4.4.4. Finally, §4.4.5 suggests a method for determining which correct value estimation procedures is best and §4.4.6 touches upon alternative estimation approaches.

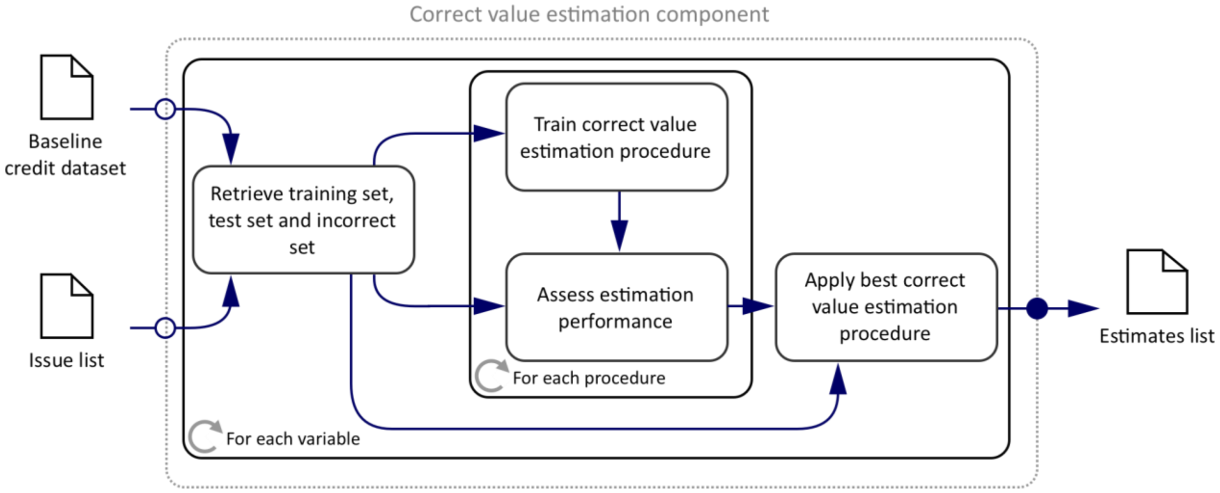

Correct value estimation component
Baseline credit dataset
Issue list
Retrieve training set, test set and incorrect set
Train correct value estimation procedure
Assess estimation performance
For each procedure
Apply best correct value estimation procedure
For each variable
Estimates list

Figure 12: Schematic overview of the correct value estimation component.

4.4.1 Implementation

The component is implemented using the four steps depicted in Figure 12:

1. **Retrieve training set, test set and incorrect set**. The component first splits the credit dataset into a correct set and an incorrect set. The former set includes all records that do not occur in the issue list and the latter set includes all records that do occur in the issue list. The component then randomly samples two mutually exclusive subsets of equal size from the correct set, which are called the training set and the test set. The training set is used to train the correct value estimation procedures and the test set is used to assess each procedure’s estimation performance.
2. **Train correct value estimation procedure**. Correct value estimation procedures must be trained on an unaffected part of the credit dataset before they can estimate what the correct values in affected records should be. A procedure uses the training set for this purpose and the information it extracts from the training sample is passed on to the next step.

3. **Assess estimation performance**. There should only be one correct value estimate for every incorrect value in the credit data. To determine which procedure should generate these estimates, this step of the component first compares the estimation performance of each procedure. To do so, it first applies every procedure to the test set. The estimates that are generated by each procedure are compared to the actual values that are available in the test set. The procedure that achieves the smallest error between the estimated values and the actual values in the test set is considered the best procedure. This assumes that when a procedure performs well on the test set, it will also be perform well on the incorrect set.
4. **Apply best correct value estimation procedure**. The best procedure is applied to the incorrect set, yielding a set of correct value estimates. Those estimates are appended to the issue list. The resulting list is called the estimates list.

Example (continued)

The issue list that is compiled in the previous example shows which records and variables are affected by which data quality issue. However, the list does not contain estimates for the correct values of the affected variables. Estimating these values is the task of the correct value estimation component. Assume that the correct value estimation component uses a simple procedure for estimating the correct values: it takes the arithmetic mean and sample standard deviation of all values of the affected variable in the training set and creates a normal distribution $\mathrm{N}(\mu, \sigma)$ with these parameters. The unaffected records that form the training set are the 2nd, 3rd, 5th, 6th and 7th record. The mean value of the income variable in these records is €36,200 and the standard deviation is € 15,418. It follows that the correct value estimate for the income variable in the 1st and 4th records is a stochastic variable with a $\mathrm{N}(36200,15418)$ distribution. These correct value estimates are recorded in the estimates list shown below.

Borrower ID	Issue name	Variable name	Incorrect value	Estimated value
1	Below minimum	Income	€ 6,000	$\mathrm{N}(36200,15418)$
4	Below minimum	Income	€ 4,000	$\mathrm{N}(36200,15418)$

This information can be used to create the corrected credit dataset as shown below. However, only the estimates list is passed along to the next step because it contains all the necessary information while being much smaller in size.

Borrower ID	Income	Exposure	Maturity	Foreclosure value
1	$\mathrm{N}(36200,15418)$	€ 150,000	10 years	€ 120,000
2	€ 21,000	€ 30,000	5 years	€ 120,000
3	€ 19,000	€ 100,000	16 years	€ 80,000
4	$\mathrm{N}(36200,15418)$	€ 40,000	2 years	€ 100,000
5	€ 50,000	€ 375,000	15 years	€ 400,000
6	€ 51,000	€ 180,000	6 years	€ 200,000
7	€ 40,000	€ 50,000	1 year	€ 500,000

4.4.2 Missing data imputation

The procedures that are used to make correct value estimates are based on so-called missing data imputation procedures. Missing data imputation is originally developed to address the problem of missing values in datasets when statistical tests and analysis techniques require complete datasets as input. Missing data imputation procedures impute estimates of correct values into an incomplete dataset to make it complete (see Figure 13). The imputed values are designed to resemble the statistical characteristics of the non-missing part of a dataset as closely as possible to minimise the probability of drawing incorrect statistical inferences based on the completed dataset (Schafer & Olsen, 1998).

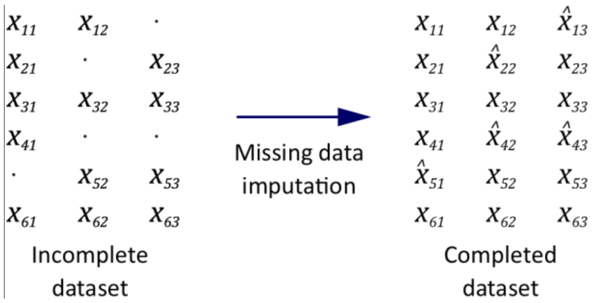

x_{11} x_{12} ·
x_{21} · x_{23}
x_{31} x_{32} x_{33}
x_{41} · ·
· x_{52} x_{53}
x_{61} x_{62} x_{63}
Incomplete dataset
Missing data imputation
x_{11} x_{12} $\hat{x}_{13}$
x_{21} $\hat{x}_{22}$ x_{23}
x_{31} x_{32} x_{33}
x_{41} $\hat{x}_{42}$ $\hat{x}_{43}$
$\hat{x}_{51}$ x_{52} x_{53}
x_{61} x_{62} x_{63}
Completed dataset

Figure 13: Missing data imputation procedures impute correct value estimates to make a dataset complete.

Many of these missing data imputation procedures work on a variable-by-variable basis. A procedure first 'learns' how to make predictions for the correct value of a certain variable using a part of the dataset that does not contain missing values. The variable in question is called the 'target variable'. The procedure then uses that knowledge to estimate the correct values of that variable to replace the missing values in the dataset. In fact, missing data imputation procedures can also be used to cleanse datasets with other types of incorrect values than missing values. For instance, a common problem with survey datasets is that some values in them are obviously incorrect, such as an age of 250 or a non-existent address. Missing data imputation procedures can replace such incorrect values by correct values estimates as if they were missing values. This makes missing data imputation procedures a good candidate for inclusion in the model. Since the procedures are used to create estimates of the correct value of variables that have been effected by a data quality issue, these procedures are from now on referred to as 'correct value estimation procedures'. A study on suitable procedures resulted in the following list. These procedures are described into more detail in the course of this chapter.

i. **Overall mean imputation**. Takes the sample mean value of the target variable values in the training set and suggests this value as the correct value estimate for every record.
ii. **Overall median imputation**. Takes the sample median of the values of the target variable in the training set and uses this for the correct value estimates.
iii. **Overall mode imputation**. Takes the sample mode of the values of the target variable in the training set and uses this for the correct value estimates.
iv. **Normal distribution imputation**. Fits a normal distribution on the target variable values in the training set and suggests a stochastic variable with this distribution for the correct value estimates.
v. **Empirical distribution imputation**. Determines the empirical frequency distribution of the target variable values in the training set and suggests a stochastic variable with the same distribution for the correct value estimates.

vi. **Linear regression imputation**. Fits a simple linear model on the target variable values using the other variables in the training set as potential covariates. Uses the linear model and the covariates in the incorrect set to make correct value estimates.

vii. **Bounded linear regression imputation**. Applies linear regression but then with a link function that bounds the possible values that the target value can assume.

viii. **Logistic linear regression imputation**. Applies logistic regression to make correct value estimates for nominal and ordinal values that can assume two values.

ix. **Multinomial logistic regression imputation**. Applies multinomial logistic regression to make correct value estimates for nominal and ordinal values that can assume a countable and reasonably small number of distinct values.

x. **K-nearest neighbours imputation**. Matches each record in the incorrect set with K similar records in the training set and suggests the mode, median or mean value of the values of the target variable in those K records as the correct value estimate.

4.4.3 Measurement levels

1. Correct value estimation procedures can only estimate correct values for a variable if that variable is of a compatible measurement level. For instance, if a procedure works by taking the arithmetic mean of every unaffected target variable value in the training set then that procedure can only be used on target variables for which the concept of a 'mean' is actually defined. In practice, this implies that the variable in question must be numeric, ordered and the difference between two values must be quantifiable. Psychologist Stanley Smith Stevens (1946) has formalised this concept onto the measurement level scale that follows.

Table 3 shows the levels of measurement of the target variables that are accepted by each correct value estimation procedures.

2. **Nominal**. Variables at the nominal measurement level differentiate values only by their name. There is no natural ordering to these values. An example of a nominal variable is the 'City of Residence' of a borrower or a variable that can assume the values 'Yes' and 'No'. The latter example is also called an *indicator* variable. Nominal values can be represented with numbers but these numbers will have their usual numerical properties.
3. **Ordinal**. Variables at the nominal measurement level have all the properties of nominal variables and there is a natural way to order the values that can be assumed. However, it is not possible to quantify the difference between two variables. An example is a variable that captures borrower satisfaction from 'Poor' to 'Excellent'.
4. **Interval**. Variables are the interval measurement level have a natural ordering in the values that can be assumed and it is possible to quantify the difference between two values. For instance, the difference between 1 and 3 is twice the difference between 1 and 2. There is no natural zero-point, which makes it invalid to state that a value of 6 is three times higher than a value of 2.
5. **Ratio**. Variables at the ratio measurement level have all the properties of interval variables and there is a natural zero-point. An example is the 'Foreclosure Value' variables. Clearly, a foreclose value of € 200,000 is twice as high as a foreclose value of € 100,000.

Table 3: Accepted levels of measurement.

Correct value estimation procedure name	Nominal	Ordinal	Interval	Ratio
Overall mean			X	X
Overall median		X	X	X
Overall mode	X	X	X	X
Empirical distribution	X	X	X	X
Normal distribution			X	X
Linear regression			X	X
Bounded linear regression			X	X
Logistic linear regression	X	X		
Multinomial logistic regression	X	X		
K-nearest neighbours	X	X	X	X

4.4.4 Correct value estimation procedures

This section uses the notation that follows to describe the correct value estimation procedures. Correct value estimation procedures work in two steps. The first step is to train the procedure on a sample of records from the part of the credit risk data that is not affected by a data quality issue. This sample is called the training set. The next step is then to apply the information that has been extracted from the training sample to the part of the dataset that is actually affected by the data quality issue. This part of the dataset is called the incorrect set.

$\mathbf{v}$ is the set of all variables in the credit data. v_j is the j^{th} variable. $\tilde{v}$ is the current target variable.

$\mathbf{y}$ represents the vector of all values of the target variable in the complete credit dataset. The subset of these values that is affected by a data quality issue is $\tilde{\mathbf{y}}$ and the set of correct value estimates for these affected values is $\hat{\mathbf{y}}$. The respective values for an individual borrower with record number i are y_i, $\tilde{y}_i$ and $\hat{y}_i$.

$\mathbf{x}$ represents the matrix of all values on variables other than the target variable. $x_{i,j}$ is the value of variable j in record i and $\mathbf{x}_{i,\cdot}$ is vector of values of all variables except the target variable, on record i.

$\mathbf{T}$ represents the training set. The expression $i \in \mathbf{T}$ means that the record for borrower number i is part of the training set. $N_{\mathbf{T}}$ is the number of records in the training set.

$\mathbf{I}$ represents the incorrect set. The expression $i \in \mathbf{I}$ means that the record for borrower number i is part of the incorrect set. $N_{\mathbf{I}}$ is the number of records in the incorrect set.

Figure 14 shows how the notation that is introduced above describes a dataset with six records and three variables. The third variable is the target variable, which means that the correct value estimation procedures are being trained and applied to create correct value estimates for all incorrect values of the third variable. Two records contain incorrect values of the target variable and are therefore part of the incorrect set. A random subset of the four remaining unaffected records has been selected to form the training set. The incorrect values are $\tilde{y}_5$ and $\tilde{y}_6$. These values are replaced by correct value estimates $\hat{y}_5$ and $\hat{y}_6$ in the corrected credit dataset.

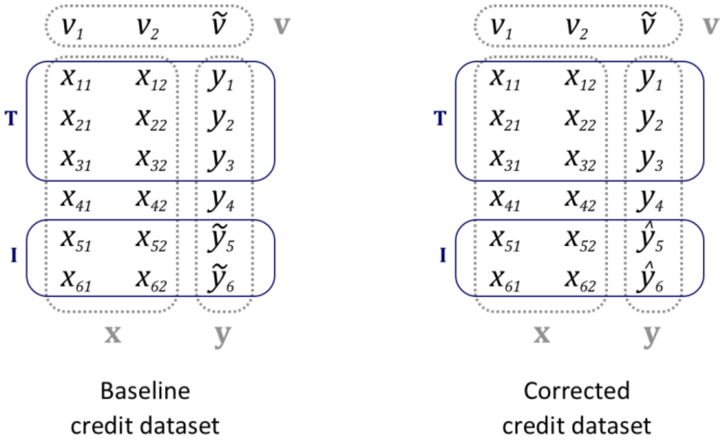


Figure 14: Graphic representation of the notation that is used in Section 0.

Overall mean

This procedure takes the mean value of the target variable $\tilde{v}$ in the training set and uses this mean as the correct value estimate. The aim of this procedure is simplicity and zero bias in the imputed values. The arithmetic mean of the values of variable $\tilde{v}$ in the training set is:

$$m_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{\forall\, i \in \mathbf{T}} y_i \tag{6}$$

However, this mean is dependent on the random selection of records that forms the training set. If one takes a different sample of records as the training set instead, then one may find a different mean. Hence, the mean of a sample is a stochastic variable. By the Central Limit Theorem (CLT), the sample mean follows a normal distribution[4] with the following variance:

$$s_\mu^2 = \frac{s_{\mathrm{T}}^2}{N_{\mathrm{T}}} \tag{7}$$

(Adapted from Larsen & Marx, 2012, p. 246)

The CLT assumes a sufficiently large training set and independence between every value in the training set. s_{T}^2 is the sample variance of the training set:

$$s_{\mathrm{T}}^2 = \frac{1}{N_{\mathrm{T}} - 1} \sum_{\forall\, \mathrm{k} \in \mathrm{T}} (y_{\mathrm{k}} - m_{\mathrm{T}})^2 \tag{8}$$

(Adapted from Larsen & Marx, 2012, p. 246)

It is now possible to fully specify the overall mean imputation procedure:

[4] This should not be confused with the theoretical distribution of the difference between the sample mean and the true mean of a sample, which follows a Student-t distribution. See Appendix 6 for a background analysis.

Training step

$$m_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{\forall\, i \in \mathbf{T}} y_i \tag{9}$$

$$s_{\mathrm{T}}^2 = \frac{1}{N_{\mathrm{T}} - 1} \sum_{\forall\, \mathrm{k} \in \mathbf{T}} (y_{\mathrm{k}} - m_{\mathrm{T}})^2$$

Application step

$$\hat{y}_i \sim \text{Normal}\left[m_{\mathrm{T}}\,, \frac{s_{\mathrm{T}}}{\sqrt{N_{\mathrm{T}}}}\right] \qquad \forall\, i \in \mathbf{I} \tag{10}$$

Overall median

A variation on the overall mean procedure uses the sample median value of the training set instead of the sample mean value of the target variable in the training set. A major advantage is that this procedure also works with variables at the ordinal measurement level. It is also less sensitive to outliers in the training set compared to the mean. To define the median of $\mathbf{y_T}$ formally, let $y_{\mathbf{T}}^{(i)}$ be the i^{th} lowest value in $\mathbf{y_T}$, such that $\min[\mathbf{y_T}] = y_{\mathbf{T}}^{(1)}$ and $\max[\mathbf{y_T}] = y_{\mathbf{T}}^{(N_{\mathrm{T}})}$. Then the median is:

$$\text{median}[\mathbf{y_T}] = \begin{cases} y_{\mathbf{T}}^{((N_{\mathrm{T}}+1)/2)} & \text{if } N_{\mathrm{T}} \text{ is odd} \\ \frac{1}{2}\left(y_{\mathbf{T}}^{(N_{\mathrm{T}}/2)} + y_{\mathbf{T}}^{((N_{\mathrm{T}}/2+1))}\right) & \text{if } N_{\mathrm{T}} \text{ is even} \end{cases} \tag{11}$$

(Weisstein, n.d.)

The median value is also sensitive to the sample of records that forms the training set. A different random selection of training records may result in a different mean. Hence, the correct value estimates that are produced by the overall median procedure are stochastic. The distribution of these values be approximated using bootstrapping, which involves randomly sampling with replacement b individual vectors of N_{T} values from $\mathbf{y_T}$ and determining the median of every one of these bootstrapped vectors. The set of b bootstrapped medians can then be used to construct an empirical random distribution of the overall median.

Overall mode

The overall mode procedure is a third variation on the overall mean imputation procedure. It takes the sample mode value of target variable in the training set. The mode value is the value that occurs most often in a vector of values. It can be calculated for every measurement level, including the nominal level.

$$\begin{aligned} &\text{mode}[\mathbf{y_T}] = \; y_k \\ &\text{s.t.} \quad N_{\mathbf{y_T}=y_k} \geq \; N_{\mathbf{y_T}=y_i} \qquad \forall\, i \neq k \end{aligned} \tag{12}$$

It is conceivable that there is more than one mode value. For practical reason the procedure will then use the value with the lowest index k, i.e. the first mode value encountered. Like the sample mean and sample median, the sample mode is sensitive to the random selection of records that is included in the training set, making the sample mode stochastic. Its random distribution can be approximated using bootstrapping.

Normal distribution

The overall mean, overall median and overall mode procedures by definition try to reproduce the central tendency of the correct values that are used as input. However, as Schafer and Olsen (1998) argue, it may also be useful to capture the variance of the correct values in the estimated values. This is what the normal distribution aims to achieve. The normal distribution procedure estimates the arithmetic mean and sample standard deviation of the target variable in the training set and uses a normally distributed stochastic variable with those parameters for every impute value.

Training step
$$m_{\mathrm{T}} = \frac{1}{N_{\mathrm{T}}} \sum_{\forall\, i \in \mathbf{T}} y_i \qquad (13)$$
$$s_{\mathrm{T}}^2 = \frac{1}{N_{\mathrm{T}} - 1} \sum_{\forall\, \mathrm{k} \in \mathbf{T}} (y_{\mathrm{k}} - m_{\mathrm{T}})^2$$

Application step
$$\hat{y}_i \sim \mathrm{Normal}[m_{\mathrm{T}}\, , s_{\mathrm{T}}] \qquad \forall\, i \in \mathbf{I} \qquad (14)$$

Note that the application step in Equation (14) is not exactly the same as in Equation (10), because the normal distribution procedure uses a standard deviation of s_{T} whereas the overall mean imputation procedure uses $s_{\mathrm{T}}/\sqrt{N_{\mathrm{T}}}$. Figure 15 shows the effect of this difference. The histogram represents the training values of a certain target variable. The continuous line represents the probability density function (pdf) of a correct value estimate that is produced by the normal distribution procedure, while the dotted line represents the pdf of a correct value estimate that is produced by the mean value imputation procedure. The latter pdf is clearly much narrower. Moreover, the histogram of the data is also a representation of the pdf of a correct value estimate that is produced by the empirical distribution procedure, which is described next.

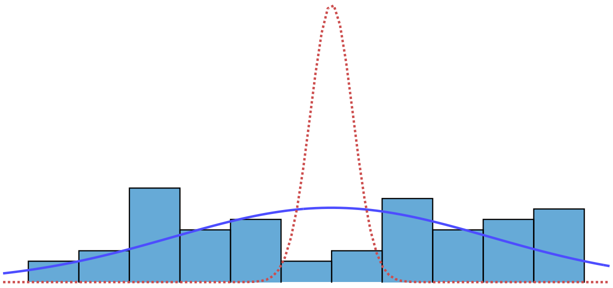


Figure 15: The normal distribution procedure compared to the overall mean procedure.

Empirical distribution

The empirical distribution imputation procedure counts the occurrences of every unique value in $\mathbf{y}_{\mathrm{T}}$ to create an empirical distribution. This distribution is used for every impute value. Let $N_{\mathbf{y}_{\mathrm{T}}=\mathrm{t}}$ be the number of times that a value in $\mathbf{y}_{\mathrm{T}}$ equals t.

Training step $$f(t) \quad \text{s.t.} \quad P[\mathbf{y}_\mathrm{T} = t] = \frac{N_{\mathbf{y}_\mathrm{T}=\mathrm{t}}}{N_\mathrm{T}} \tag{15}$$

Application step $$\hat{y}_i = f(t) \qquad \forall\, i \in \mathbf{I} \tag{16}$$

The number of unique values in the training set is liable to be very large for continuous variables. In such cases, it may be efficient to reduce the number of unique values in $\mathbf{y}_\mathrm{T}$ before applying the empirical distribution procedure. One way to so is by bucketing the values in $\mathbf{y}_\mathrm{T}$, which involves splitting up the range of $\mathbf{y}_\mathrm{T}$ into a limited set of sub ranges. These sub ranges are called buckets. All values in $\mathbf{y}_\mathrm{T}$ that fall in the same bucket are then assigned a common value. That value is usually the midpoint of the bucket in question. There are numerous approaches to choosing the sub ranges. One approach is split the full range of $\mathbf{y}_\mathrm{T}$ into N_buckets sub ranges of equal width and another approach is to choose the sub ranges such that the number of values that falls in the each bucket is the same for every bucket.

Linear regression

When a data quality issue affects a borrower's records, it will usually only invalidate the values of one or at most several variables. However, the credit data held by SNS Bank N.V. stores values on more than 100 variables for each borrower record. The values on these records may contain valuable information for estimating the correct value of the invalided variable value in a record. A technique for using that information is linear regression. This paragraph shows the basics of fitting a linear model, generating estimates, selecting covariates, handling qualitative covariates and handling missing values in the covariates.

Fitting a model

The linear regression procedure chooses a subset $\mathbf{c}$ from the set of non-target variables in $\mathbf{v}$, such that the variables in $\mathbf{c}$ have enough explanatory power to help predict the correct value estimate $\hat{y}_i$. These variables are called covariates or predictors. The number of covariates is $N_\mathbf{c}$ and c_i is the i^{th} covariate. Linear regression assumes the following linear relationship between the target variable value y_i and the row vector of covariate values $\mathbf{x}_{i,\mathbf{c}}$ (Verbeek, 2004, pp. 8-10):

$$y_i = \beta_0 + \beta_1 x_{1,c_1} + \beta_2 x_{1,c_2} + \cdots + \beta_{n_\mathbf{c}} x_{1,c_{N_\mathbf{c}}} + \epsilon_\mathrm{i} = \begin{bmatrix}1 & \mathbf{x}_{i,\mathbf{c}}\end{bmatrix}' \boldsymbol{\beta} + \epsilon_\mathrm{i} \tag{17}$$

$\boldsymbol{\beta}$ is the $N_\mathbf{c} + 1$ vector of linear coefficients in this linear model, whereby β_0 is the constant term or *intercept* and ϵ is a normally distributed error term with the same variance for every i and an expected value of 0. $\boldsymbol{\beta}$ is unknown and must therefore be estimated. The corresponding estimated vector is denoted $\widehat{\boldsymbol{\beta}}$, such that:

$$\begin{aligned} y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{1,c_1} + \hat{\beta}_2 x_{1,c_2} + \cdots + \hat{\beta}_{n_\mathbf{c}} x_{1,c_{N_\mathbf{c}}} + \hat{e}_i = \begin{bmatrix}1 & \mathbf{x}_{i,\mathbf{c}}\end{bmatrix}' \widehat{\boldsymbol{\beta}} + \hat{e}_i \\ \hat{y}_i &= \begin{bmatrix}1 & \mathbf{x}_{i,\mathbf{c}}\end{bmatrix}' \widehat{\boldsymbol{\beta}} \end{aligned} \tag{18}$$

$e_i =$ is an error term that contains the difference between the correct value y_i and the actual correct value estimate $\hat{y}_i$. A smaller squared e_i implicates a better prediction for observation i. $\mathbf{e}$ is the vector of error terms for all borrowers. $\begin{bmatrix}1 & \mathbf{x}_{i,\mathbf{c}}\end{bmatrix}' \widehat{\boldsymbol{\beta}}$ is the linear model that creates the correct value estimate $\hat{y}_i$. Assuming that the covariates have already been selected, $\widehat{\boldsymbol{\beta}}$ can be estimated ('fitted') on the training sample using the ordinary least squares (OLS) estimator. Let $\mathbf{X} = \mathbf{x}_{\mathrm{T},\mathbf{c}}$ be the matrix with the values for all borrowers in the training set on all covariates.

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}_{\mathbf{S}} \tag{19}$$

See Appendix 4 for a derivation of the univariate version of this estimator using the likelihood function that optimises $\widehat{\boldsymbol{\beta}}$.

The OLS estimator makes several assumptions about the model and the data it is fitted on. These assumption are referred to as the Gauss-Markov assumptions:

1. The expected value of the error term is zero (no bias).
2. The covariates and error terms are independent (no multicollinearity).
3. All errors terms have the same variance (homoscedasticity).
4. There is zero correlation between the error terms (no autocorrelation).

In practice, the assumptions are seldom all fulfilled. Especially assumption 2 is hard to fulfil when using credit data, since the variables in this data tend to be correlated. When two covariates are perfectly or strongly correlated, then the OLS estimator yields unreliable results because the matrix $\mathbf{X}'\mathbf{X}$ becomes (close to) singular, which means that the matrix inverse $(\mathbf{X}'\mathbf{X})^{-1}$ cannot be calculated. However, when covariates are only slightly correlated, then the main effect is that the variance of the regression coefficients $\hat{\beta}_i$ increases (O'brien, 2007, p. 673). This does not necessarily lead to a biased linear regression model.

Generating correct value estimates

If the assumptions 1 to 4 are fulfilled, the error terms e_i of a linear model are normally distributed with the same standard deviation in every error term and zero bias. Therefore, the expected value of every impute value is:

$$\mathrm{E}[\hat{y}_i] = \begin{bmatrix} 1 & \mathbf{x}_{i,\mathbf{c}} \end{bmatrix}' \widehat{\boldsymbol{\beta}} \tag{20}$$

The variance of this impute value depends on the variance of the error term and the variance of the coefficients. Assuming independent covariates, these variances can be estimated as follows:

$$s_{\mathrm{e}}^2 = \frac{1}{N_{\mathrm{T}} - 1} \sum_{\forall\, k \,\in\, \mathrm{T}} (\hat{y}_k - y_k)^2 \tag{21}$$

$$s_j^2 = \frac{\mathrm{s}_{\mathrm{e}}^2}{s_{\mathbf{x}_{\mathrm{T},j}}^2} \tag{22}$$

Assuming that the covariates are mutually independent, the variance or the correct value estimate can be found by scaling and summing these variances:

$$s_{\mathrm{\hat{y}_i}}^2 = \sum_{\mathrm{j=1}}^{\mathrm{N_c}} (\mathrm{x_{i,j}^2} \cdot \mathrm{s_j^2}) + \mathrm{s_e^2} \tag{23}$$

It is now possible to specify the linear regression imputation procedure as follows:

Training step

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_{\mathrm{T}} \tag{24}$$
$$s_e^2 = \frac{1}{N_T - 1} \sum_{\forall\, i \in \mathrm{T}} (\hat{y}_i - y_i)^2$$

Application step

$$\mathrm{s}_{\hat{y}_i}^2 = \sum_{j=1}^{N_{\mathbf{c}}} \left(x_{i,j}^2 \cdot s_j^2\right) + s_{\mathrm{e}}^2 \tag{25}$$
$$\hat{y}_i = \mathrm{Normal}\left[\left[1 \;\; \mathbf{x}_{i,\mathbf{c}}\right]' \hat{\boldsymbol{\beta}} \,,\, s_{\hat{y}_i}\right] \qquad \forall\, i \in \mathbf{I}$$

Selecting covariates

The performance of a model does not necessarily get better as more covariates are added, because a model with too many covariates is likely to be overfitted. The consequence of over fitting is a model that gives good predictions when estimating the dependant variable in the training set (i.e. in-sample performance) but poor results when it is applied to records that are not in the training set (i.e. out-of-sample performance). A complex model with many covariates should only be chosen over a simple model with few covariates if it has significantly better predictive power (Verbeek, 2004, pp. 56-58). To make this judgement, Akaike (1973) and Schwartz (1978) have developed the Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC), respectively. Both model selection criteria balance model performance with the number of included variables.

$$\mathrm{AIC} = N_{\mathrm{T}} \cdot \log\left[\frac{1}{N_{\mathrm{T}}} \sum_{\forall\, \mathrm{i} \in \mathbf{T}} e_i^2\right] + 2 \cdot N_{\mathbf{c}} \tag{26}$$

$$\mathrm{BIC} = N_{\mathrm{T}} \cdot \log\left[\frac{1}{N_{\mathrm{T}}} \sum_{\forall\, \mathrm{i} \in \mathbf{T}} e_i^2\right] + N_{\mathbf{c}} \cdot \log N_{\mathrm{T}} \tag{27}$$

(Verbeek, 2004, p. 58)

A model with a lower information criterion value is preferred. The BIC awards a larger penalty to the number of included covariates and thus prefers models with a lower number of covariates compared to the AIC. The model with a low information criterion based on a set of possible covariates can be found using a stepwise linear regression algorithm. The following heuristic algorithm builds a linear model by adding or removing one covariate and a time and determining whether the addition or removal of a covariate results in a better model. This is done until no more improvements can be made:

1. Start with a model with no covariates (only an intercept term).
2. Try out adding every covariate that is not yet in the model and determine which covariate addition resulted in the best criterion improvement. Permanently add that covariate, or none if no addition results in a criterion improvement.
3. Try out removing every covariate that is already in the model and determine which covariate removal resulted in the best criterion improvement. Permanently remove that covariate, or none if no removal results in a criterion improvement.

4. An optimal set of covariates has been found if none were permanently added nor removed in this iteration. Otherwise proceed from step 2.

Handling qualitative covariates

The covariates included in a linear model need to be of the interval or ratio measurement level. However, nominal and ordinal variables can be easily adapted to fit this requirement by splitting the variable up into several sub indicator variables. A variable with k possible options can be represented by $k-1$ indicator variables. As an example, take the qualitative variable j containing values $q \in (A, B, C, D)$. Then define the indicator variables $x_{i,B}, x_{i,C}, x_{i,D}$, such that:

$$x_{i,j,q} = \begin{cases} 1 & \text{if } x_{i,j} = q \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

$$x_{\cdot,j} = \begin{bmatrix} A \\ C \\ D \\ A \\ B \\ D \end{bmatrix} \quad \rightarrow \quad x_{\cdot,j,B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, x_{\cdot,j,C} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, x_{\cdot,j,D} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

The set of three indicator variables now fully represents all states in the original variable despite having one variable less than the number of states assumed by original variable because the model also includes an intercept coefficient β_0:

$$y_i = \beta_0 + \beta_1 x_{i,j,B} + \beta_2 x_{i,j,C} + \beta_3 x_{i,j,D} + e_j \tag{29}$$

Handling covariates with missing values

A variable can only be used as a covariate in a linear model if it does not contain any missing values in the training sample nor the incorrect set since the OLS estimator does not accept missing values. As a remedy, one can substitute all missing values in a variable by a number, e.g. a zero. However, such a substitution may result in valuable information to be thrown away, because the fact that a value is missing is information in itself. To avoid discarding such information when substituting missing values an additional indicator variable is added to the dataset of explanatory variables with a 1 if a substitution is done and a 0 otherwise, as exemplified below. The OLS estimator estimates a separate coefficient for the 'isNaN' variable. If the fact that a value is missing does not have significant predictive value, then the stepwise covariate selection procedure will not include the 'isNaN' column.

$$x_{\cdot,j} = \begin{bmatrix} 53712{,}67 \\ NaN \\ 0 \\ NaN \\ 894{,}50 \\ 1044{,}07 \end{bmatrix} \quad \rightarrow \quad x_{\cdot,j} = \begin{bmatrix} 53712{,}67 \\ 0 \\ 0 \\ 0 \\ 894{,}50 \\ 1044{,}07 \end{bmatrix}, x_{\cdot,j,isNaN} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{30}$$

Bounded linear regression

The previously described ordinary linear regression model can make estimates for target variables that are continuous and unbounded. However, most variables (i.e. Age, Maturity, Exposure) cannot assume values outside a certain range and are thus bounded. Moreover, many variables are not continuous but discrete with two or more possible values. All these cases can be handled quite straightforwardly using an adaptation of linear regression models. The key is to use a *link function* that transforms the values of the dependent variable to an unbounded, continuous range that can be used in an ordinary linear regression model. This function is denoted $z_i = G(y_i)$ and the inverse is $y_i = G^{-1}(z_i)$. The linear model and OLS estimator then become:

$$\begin{aligned} z_i &= \left[1 \ \mathbf{x}_{(\mathrm{i,cov})}\right]' \widehat{\boldsymbol{\beta}} \\ G(y_i) &= \left[1 \ \mathbf{x}_{(\mathrm{i,cov})}\right]' \widehat{\boldsymbol{\beta}} \\ y_i &= G^{-1}\left(\left[1 \ \mathbf{x}_{(\mathrm{i,cov})}\right]' \widehat{\boldsymbol{\beta}}\right) \end{aligned} \tag{31}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'G(\mathbf{y}_{\mathbf{S}}) = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{z}_{\mathbf{S}} \tag{32}$$

An interesting choice for G and G^{-1} is the following pair of link functions:

$$\begin{aligned} z_i &= G(y_i) = \log\left(\frac{y_i - a}{b - y_i}\right) \\ y_i &= G^{-1}(z_i) = \frac{a + b \cdot \exp(z_i)}{1 + \exp(z_i)} \end{aligned} \tag{33}$$

These functions can map a continuous value y_i that lies between a and b to a value z_i that lies between $-\infty$ and ∞ (see Appendix 5). That way, it is possible to fit a model that only generates values that lie between two bounds. This procedure does not work for variables that are bounded on just only side, because the link functions do not work when a or b is infinite. Once the target variable is mapped, the bounded linear regression imputation procedure works identically to the linear regression imputation procedure. The resulting distribution of the resulting stochastic correct value estimates is then a transformation of the underlying $\hat{z}_i$.

$$\begin{aligned} &\text{Training step} && \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{\mathbf{-1}}\mathbf{X}'G(\mathbf{y}_{\mathrm{T}}) \\ & && s_e^2 = \frac{1}{N_T - 1}\sum_{\forall\, i \in \mathrm{T}} (\hat{y}_i - y_i)^2 \end{aligned} \tag{34}$$

$$\begin{aligned} &\text{Application step} && \mathrm{s}_{\hat{y}_i}^2 = \sum_{j=1}^{N_{\mathbf{c}}} \left(x_{i,j}^2 \cdot s_j^2\right) + s_{\mathrm{e}}^2 \\ & && \hat{z}_i = \mathrm{Normal}\left[\left[1 \ \mathbf{x}_{i,\mathbf{c}}\right]' \widehat{\boldsymbol{\beta}} \,,\, s_{\hat{y}_i}\right] \qquad \forall\, i \in \mathbf{I} \\ & && \hat{y}_i = G^{-1}(\hat{z}_i) \end{aligned} \tag{35}$$

Logistic regression

Indicator variables can only assume two values, e.g. 'true'/'false', 'yes'/'no', 1/0. An example such a variable is whether a customer has even been in default in the last twelve months. Such variables can be estimated using logistic regression. The logistic link function is designed to estimate a probability p that the value is 1 (which usually means 'true' or 'yes'). The estimated probability for the value 0 is the complement: $1 - p$.

$$z_i = G(y_i) = \log\left(\frac{y_i}{1 - y_i}\right) \tag{36}$$

(Cox, 1958)

$$\begin{aligned} \mathrm{P}[y_i = 1] = G^{-1}(z_i) = p_i &= \frac{\exp(z_i)}{1 + \exp(z_i)} \\ \mathrm{P}[y_i = 0] = 1 - p_i &= \frac{1}{1 + \exp(z_i)} \end{aligned} \tag{37}$$

(Cox, 1958)

The logistic link function resembles the bounded linear regression link function with $a = 0$ and $b = 1$. However, it is applied differently because the logistic regression model cannot be used in combination with the OLS estimator. To see why, consider that the bounded linear regression procedure only accepts continuous variables. By definition, the probability of a continuous variable assuming any particular value is zero. This implies that:

$$\mathrm{P}[y_i = a] = 0 \quad \text{and} \quad \mathrm{P}[y_i = b] = 0 \tag{38}$$

This is an important property because the link functions yield infinite values on these bounds. For instance, consider $y_i = a$:

$$z_i = \log\left(\frac{y_i - a}{b - y_i}\right) = \log\left(\frac{a - a}{b - a}\right) = \log(0) = -\infty \tag{39}$$

The transformed OLS estimator only works with real values for $\mathbf{z}$ and as a consequence y_i cannot be a or b. In the logistic regression case y_i is always either 0 or 1, which implies that z_i is always $-\infty$ or ∞. An alternative estimator for β is therefore needed. That estimator must maximise following likelihood equation (Verbeek, 2004, p. 193):

$$\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) = \sum_{\forall\ \mathrm{i} \in \mathbf{S}} \mathrm{P}\left[y_i = 1 \middle| \mathbf{x}_{i,\mathbf{c}}; \boldsymbol{\beta}\right)^{y_i} \mathrm{P}\left[y_i = 0 \middle| \mathbf{x}_{i,\mathbf{c}}; \boldsymbol{\beta}\right)^{1-y_i} \tag{40}$$

This maximisation problem does not have a closed-form solution for $\hat{\beta}$. Instead the logistic regression procedure iteratively approximates the following first-order condition:

$$\frac{\partial \log L(\hat{\beta})}{\partial \hat{\beta}} = \sum_{\forall\ \mathrm{i} \in \mathbf{S}} \left[\mathrm{y_i} - \frac{\exp(x_i \hat{\beta})}{1 - \exp(x_i \hat{\beta})}\right] x_i = 0 \tag{41}$$

Moreover, logistic regression is also referred to as logit regression and the regression model that is fitted is called a logit model. The Bernoulli random distribution of $\hat{y}_i$ follows naturally from the definition of a logit model.

Training step Approximate: (42)

$$\frac{\partial \log \mathrm{L}(\hat{\beta})}{\partial \hat{\beta}} = \sum_{\forall\, \mathrm{i} \in \mathbf{S}} \left[\mathrm{y_i} - \frac{\exp(\mathrm{x_i}\hat{\beta})}{1 - \exp(\mathrm{x_i}\hat{\beta})} \right] \mathrm{x_i} = 0$$

Application step (43)

$$p_i = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)} \qquad \forall\, i \in \mathbf{E}$$
$$\hat{y}_i = \mathrm{Bernoulli}[p_i] \qquad \forall\, i \in \mathbf{E}$$

Multinomial logistic regression

The link function of the logistic regression model can be extended to support variables with more than two possible values, such as the 'Housing type' variable. This variable is relevant to the credit risk models because housing prices are indexed at a different rate for every type of house. Before running the regression y_i has to be decomposed into indicator variables $y_{i,1}, \dots y_{i,k}$. Then let $z_{i,q}$ be the transformed variable representing value q for borrower i. Each $z_{i,q}$ except for $z_{i,1}$ requires a separate linear model to be fitted. The link functions are as follows:

$$z_{i,q} = G(y_{i,q}) = \log\left(\frac{y_{i,q}}{1 - y_{i,q}}\right) \quad (44)$$
$$\text{where } z_{i,1} = 0$$

$$\mathrm{P}[y_i = 1] = G^{-1}(z_{i,2}, \dots, z_{i,k}) = \frac{1}{1 + \exp(z_{i,2}) + \dots + \exp(z_{i,k})} \qquad \text{if } q = 1 \quad (45)$$
$$\mathrm{P}[y_i = q] = G^{-1}(q, z_{i,2}, \dots, z_{i,k}) = \frac{\exp(z_{i,q})}{1 + \exp(z_{i,2}) + \dots + \exp(z_{i,k})} \qquad \text{if } q > 1$$

The multinomial regression imputation procedure then works analogously to the logistic regression imputation procedure. The correct value estimates follow a Multinomial stochastic distribution. This procedure works best for variables at the nominal and ordinal measurement level. It is theoretically possible but not practically feasible to apply multinomial logistic regression to interval and ratio variables because these variables tend to assume a large number of distinct values.

K-nearest neighbour imputation

The previous procedures are based on the assumption that relationships between variables in the credit data can be described using a mathematically defined model. Such approaches generate good results in an efficient manner as long as the credit data does indeed fit in the model imposed on it. However, it is conceivable that the relationships between variables in the credit data are simply too complex to be fitted in a model. As an alternative, this paragraph introduces a missing value imputation technique that does not try to impose a model on the credit data but instead searches for already existing values to suggest as correct value estimates. This procedure is K-nearest neighbour imputation, or k-NN imputation for short.

Given a record i with an incorrect value on a target variable and a training set of records with correct values, a k-NN procedure finds the K records in the training set that are the most similar to record i. These records are called the 'nearest neighbours'. It then takes the mean or mode of the valid values in the nearest neighbours and imputes it into record i. A rule of thumb for the number of selected neighbours K is the square root of the number of records in the training sample, i.e. $\sqrt{N_\mathrm{T}}$. The 'nearness' of a record i to another record k is defined by how close the variable values or records i are to the corresponding variable values in record k. That closeness is quantified using a distance measure such as the Euclidian distance measure. The Euclidian distance between record $\mathbf{x}_{i,\cdot}$ and record $\mathbf{x}_{k,\cdot}$ is:

$$d(\mathbf{x}_{i,\cdot}, \mathbf{x}_{k,\cdot}) = \sqrt{\sum\nolimits_{\forall\ v \neq \hat{v}} \left(\mathbf{x}_{i,v} - \mathbf{x}_{k,v}\right)^{\mathbf{2}}} \tag{46}$$

This distance measure makes intuitive sense because it essentially calculates the shortest geometric distance between two points in a space with $(n_{\mathbf{v}} - 1)$ dimensions. However, in its current form the variables with a high variance have a much larger impact on the distance measure than variables with a small variance. This is resolved by standardising all variables $\mathbf{v}$ by their standard deviation to form the standardised Euclidean distance measure:

$$\underline{d}(\mathbf{x}_{i,\cdot}, \mathbf{x}_{k,\cdot}) = \sqrt{\sum\nolimits_{\forall\ v \neq \hat{v}} \left(\underline{\mathbf{x}}_{t,v} - \underline{\mathbf{x}}_{i,v}\right)^{\mathbf{2}}} \tag{47}$$

$$\text{where } \underline{\mathbf{x}}_{t,v} = \frac{\mathbf{x}_{t,v}}{\sqrt{\mathrm{var}[\mathbf{x}_{\cdot,v}]}}$$

$$\text{and } \quad \underline{\mathbf{x}}_{i,v} = \frac{\mathbf{x}_{i,v}}{\sqrt{\mathrm{var}[\mathbf{x}_{\cdot,v}]}}$$

The k-NN imputation procedure does not fit a model on the training set. Instead, it simply stores the entire training set for use in the application step. When creating correct value estimates, it determines the standardised Euclidian distance between every incorrect record and every record in the training set. For every incorrect record, it takes the K closest neighbours and it takes the mean or mode value of the target variable values in those neighbours as the correct value estimate.

Training step — Store $\mathbf{x}_{\mathrm{T}}$ and $\mathbf{y}_{\mathrm{T}}$ (48)

Application step — For every $i \in \mathbf{I}$ (49)

1. Determine $\underline{d}_{i,k} = \underline{d}(\mathbf{x}_{i,\cdot}, \mathbf{x}_{k,\cdot}) \quad \forall\, k \in \mathbf{T}$
2. Take the K target variable values from $\mathbf{y}_{\text{training sample}}$ for which $\underline{d}_{i,k}$ is the smallest.
3. If the target variable is of the interval or ratio measurement level, let $\hat{y}_i = \bar{\mathbf{y}}_{\text{nearest neighbours}}$.

   Otherwise, $\hat{y}_i = \mathrm{mode}[\mathbf{y}_{\text{nearest neighbours}}]$.

Selecting covariates

A problem that is commonly encountered when applying k-NN imputation is the ‘curse of dimensionality’. The problem entails that in a dataset with a large number of variables there may be only a limited number of variables that are actually useful for determining the distance between two records. Using more variables as covariates than needed may at best make the distance measurement more complex and less efficient and at worse yield unexpected results. This problem is similar to the problem of overfitting a linear model and a common solution to the problem is analogous to stepwise linear regression:

1. Divide the training set into a training subset and a test subset.
2. Start with a distance measure that has only one randomly selected covariate. Use this distance measure to make predictions for every value of the target variable in the test subset. Use the BIC to determine the quality of these predictions.
3. Try out adding every covariate that is not in the distance measure, make new predictions and determine which covariate addition results in the best BIC improvement. Permanently add that covariate, or none if no addition resulted in a criterion improvement.
4. If there are currently two or more covariates in the distance measure, then try out removing every covariate that is in the distance measure, make new predictions and determine which covariate removal results in the best BIC improvement. Permanently remove that covariate, or none if no removal resulted in a criterion improvement.
5. An optimal set of covariates has been found if none were added or removed in this iteration. Otherwise return to step 3.

4.4.5 Selecting the best correct value estimation procedure

Every procedure produces a model for every individual target variable if that variable's measurement level is compatible with the procedure. However, there can be only one correct value estimate for every record in the estimates list. That means there needs to be a method to determine which procedure has made the 'best' correct value estimates. The mean squared error (MSE) goodness-of-fit measure can accomplish that.

To perform the assessment, every procedure is applied on a sample of records from a part of the unaffected set of records that was not used to train the procedure. This is called the test sample and testing the procedure on this data is also referred to as an out-of-sample test. Such a test measures how well a procedure's performance generalises to other datasets than the one that is used for training. The application of the procedure on the test set results in a vector of estimates $\hat{\mathbf{y}}_{\text{test set}}$. That vector is used in two ways:

1. To determine how well the estimates $\hat{\mathbf{y}}_{\text{test set}}$ approach the correct values $\mathbf{y}_{\text{test set}}$ in the test set.
2. To determine how well the corrected financial measures for the records in the test set $\hat{\mathbf{Y}}_{\text{test set}}$ approach the correct[5] financial measures $\mathbf{Y}_{\text{test set}}$. The correct financial measures $\mathbf{Y}_{\text{test set}}$ are obtained by evaluating the system with the test set as input. Then the expected values of the estimates $\mathrm{E}[\hat{\mathbf{y}}_{\text{test set}}]$ are imputed into the test set and the system is evaluated again to obtain $\hat{\mathbf{Y}}_{\text{test set}}$.

Out-of-sample goodness-of-fit of the correct value estimates

One way to measure the closeness of $\hat{\mathbf{y}}_{\text{test set}}$ to $\mathbf{y}_{\text{test set}}$ is to determine the biasedness of the estimated vector compared to the correct vector and the difference between the individual values $\hat{y}_i - y_i$. Based on work by Neyman and Pearson (1933) and Neyman (1937), these two properties can be united in a single measure called the 'mean squared error' (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{50}$$

Where N is the number of records in the test set. The 'best' procedure is the one that attains the lowest MSE when making estimates on the same test set for every procedure. The measure is primarily suited to testing the

[5] 'Correct' in this sense means 'not affected by a known data quality issue'.

performance on variables at the interval and ratio levels. It can be easily adapted to work on variables at the nominal and ordinal measurement levels as well. Let q be one of the $N_{\mathbf{Q}}$ values that variable $\hat{v}$ can assume and define:

$$y_{i,q} = \begin{cases} 1 & \text{if } y_{\mathrm{i}} = q \\ 0 & \text{otherwise} \end{cases} \quad , \quad \hat{y}_{i,q} = \begin{cases} 1 & \text{if } \hat{y}_{\mathrm{i}} = q \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

Then a version of the MSE that is adapted to such discrete values is:

$$\begin{aligned} \mathrm{MSE}_{\mathrm{discrete}} &= \frac{1}{N}\sum_{i=1}^{N} y_{i,1}\left(y_{i,1} - \hat{y}_{i,1}\right)^2 + \cdots + y_{i,1}\left(y_{i,n_{\mathbf{Q}}} - \hat{y}_{i,n_{\mathbf{Q}}}\right)^2 \\ \mathrm{MSE}_{\mathrm{discrete}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{q=1}^{N_{\mathbf{Q}}} y_{i,q}\left(1 - \hat{y}_{i,q}\right)^2 \end{aligned} \tag{52}$$

In the current form this measure essentially counts the number of incorrectly predicted values, where a smaller value is preferred. The measure may be refined using the fact that the correct value estimation procedures estimate a multinomial stochastic variable. Such a correct value estimate defines for every possible value q a certain probability $\hat{p}_{i,q}$ that it is the correct value of the target variable in record i. Redefining the MSE:

$$\begin{aligned} \hat{p}_{i,q} &= \widehat{\mathrm{P}}[y_{i,q} = 1] \\ \mathrm{MSE}_{\mathrm{multinomial}} &= \frac{1}{n}\sum_{i=1}^{n}\sum_{q=1}^{n_{\mathbf{Q}}} y_{i,q}\left(1 - \hat{p}_{i,q}\right)^2 \end{aligned} \tag{53}$$

This gives a more accurate assessment of the performance of a procedure on a nominal or ordinal variable. Firstly, a correct value estimation procedure can now 'spread it bets' for the correct value estimate over several possible values, assigning a higher probability to the values that it deems more likely to be the correct value.

Out-of-sample goodness-of-fit of the financial measures

The correct value estimation procedures work by replicating the statistical characteristics of the training data and thus making correct value estimates that are close to the training values. However, the ultimate goal of the correct value estimation component is to suggest correct value estimates that lead to correctly estimated financial measures. These are two different goals, because the system under study is not linear and therefore a good estimate of the correct value of the target variable does not necessarily imply a good estimate of the correct financial measure. Hence, it makes sense to test how closely the estimated financial measures for the records in the test set $\hat{\mathbf{Y}}_{\text{test set}}$ approach the correct financial measures $\mathbf{Y}_{\text{test set}}$. Every financial measure that is in the scope of this research is either at the interval or ratio measurement level. The MSE can be applied without adaption once the financial measures for the original and the imputed test set are calculated:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \tag{54}$$

Once again, a refinement is order to more accurately determine the performance of procedures that output multinomial correct value estimates. Let $\hat{Y}_{i,q}$ be a financial measure that is calculated based on record i with value q imputed for the target variable[6]:

[6] Note that the credit risk models used by SNS Bank N.V. calculate risk measures at record-level rather than portfolio level.

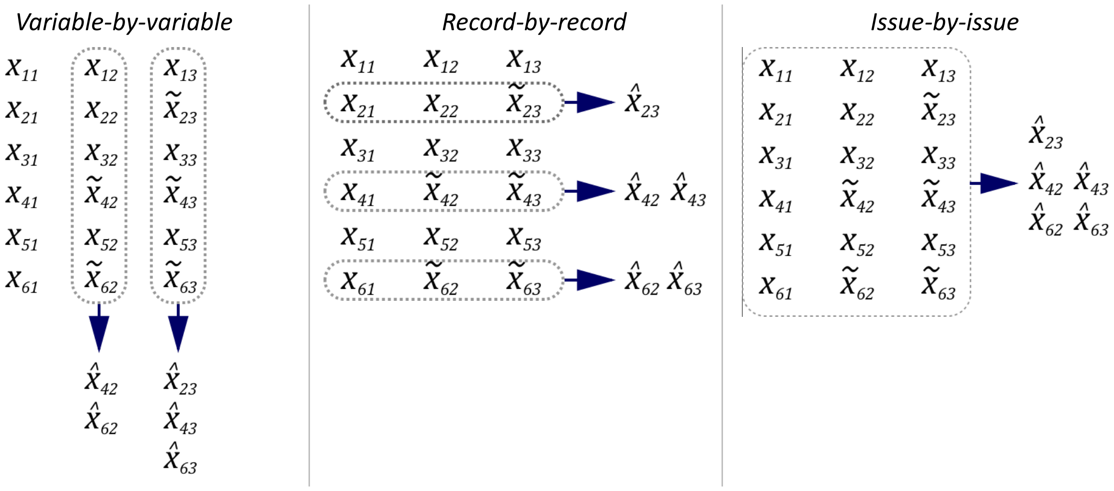
$$\mathrm{MSE}_{\mathrm{multinomial}} = \sum_{i=1}^{N}\sum_{q=1}^{N_{\mathbf{Q}}} \hat{p}_{i,q}\left(Y_i - \hat{Y}_{i,q}\right)^2 \tag{55}$$

To use this measure, the model evaluates the system several times such that every value q that has $\hat{p}_{i,q} > 0$ is imputed once for every record in the test set. The resulting MSE is the probability-weighted sum of the squared errors that result from imputing each individual value q.

4.4.6 Alternative estimation approaches

Figure 12 reveals that the component trains and applies the correct value estimation procedures on a variable-by-variable basis. This is, however, just one possible approach for estimating the correct value of incorrect variables in a dataset.

Figure 16 compares this approach and several other approaches.



Variable-by-variable
Record-by-record
Issue-by-issue
X_{11} X_{12} X_{13}
X_{21} X_{22} $\tilde{X}_{23}$
X_{31} X_{32} X_{33}
X_{41} $\tilde{X}_{42}$ $\tilde{X}_{43}$
X_{51} X_{52} X_{53}
X_{61} $\tilde{X}_{62}$ $\tilde{X}_{63}$
$\hat{X}_{23}$
$\hat{X}_{42}$ $\hat{X}_{43}$
$\hat{X}_{62}$ $\hat{X}_{63}$

Figure 16: Correct value estimation approaches.

i. **Estimating correct values variable-by-variable**. This approach first determines which variables in the dataset have at least one incorrect value. Hereby, it is useful to look at a variable as a column in the dataset. The variable has one value for every record and some of these values are incorrect due to a data quality issue. Not every record that is affected by a data quality issue needs to have an incorrect value in this column. The approach takes one variable at a time and estimates the correct value for all affected values of that variable simultaneously, even if the incorrectness those values is caused by more than one data quality issue. A major advantage of this approach is that is allows one to use many commonplace statistical estimation techniques such as linear regression for the correct value estimation. Another advantage is that this approach can easily capture the characteristics of a variable, such as its measurement level and probability distribution. A disadvantage is that is it hard to capture exact relationships between the variables in an individual record. For instance, if the value of variable A must always be the sum of variables B and C, then the variable-by-variable approach is liable to give invalid predictions because it cannot easily take such relationships into account.

ii. **Estimating correct values records-by-record**. The disadvantage of the former method can resolved by simultaneously producing estimates for all incorrect values in each individual record. Hereby, each record is essentially a row in the dataset. A correct value estimation procedure would be programmed or trained to recognise the relations between the variables in a record. It can then use that information to make accurate predictions of the correct values in each record. However, such relations are seldom straightforward to learn precisely. Programming them by hand is time-consuming and makes it harder to reuse the model for a different system and dataset. Furthermore, techniques like linear regression can also do a good job of capturing the approximate relationships between variables.

iii. **Estimating correct values issue-by-issue**. A third approach is to estimate the correct values for all affected records of a single data quality issue at a time. This approach assumes that the correct value estimation procedure has a way of learning about the characteristics of every individual data quality issue. That would mean that the data quality issue has already been analysed in-depth or that it has already been partially resolved. However, the present model is meant make fair estimations of data quality issue impacts before the bank has decided where to allocate the human resources that are needed to analyse and resolve an issue. As a consequence, the issue-by-issue approach lacks the required information to be effective.

4.5 Impact assessment component

The impact assessment component takes the estimates list and the baseline credit dataset. It then uses Monte Carlo simulation to assess the impact of the in-scope data quality issues on financial measures. This chapter first explains how the component is implemented. It then presents three approaches to handling the stochastic variables in the estimates list and the choice for a simulation approach, which is introduced afterwards.

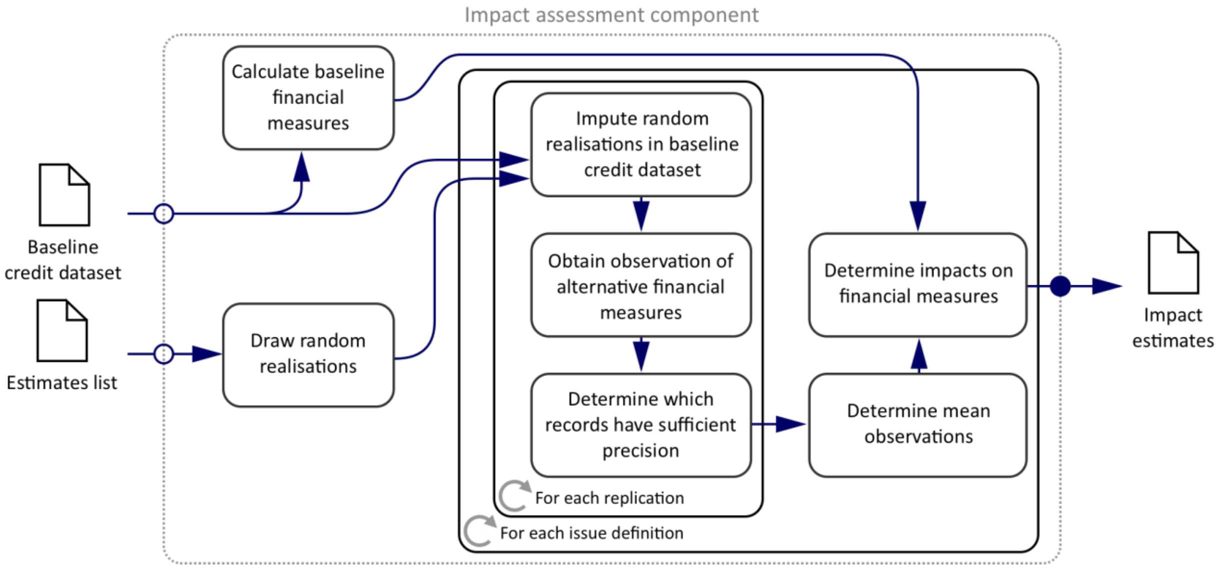

Impact assessment component
Calculate baseline financial measures
Baseline credit dataset
Estimates list
Draw random realisations
Impute random realisations in baseline credit dataset
Obtain observation of alternative financial measures
Determine which records have sufficient precision
For each replication
For each issue definition
Determine impacts on financial measures
Determine mean observations
Impact estimates

Figure 17: Schematic overview of the impact assessment component.

4.5.1 Implementation

The component carries out its task in the following steps:

1. **Calculate baseline financial measures**. The baseline financial measures can be readily determined by evaluating the system with the baseline credit dataset as input. The system is described in more detail in Chapter 3. It results in a set of values of all financial measures for each record in the baseline credit dataset.
2. **Draw random realisations**. The estimates in the estimates list are stochastic, but the system requires deterministic values as input. Using a Monte Carlo simulation simulation approach, the component draws a sample of realisations from each stochastic correct value estimate in the estimates list. A Monte Carlo simulation involves evaluating a system a large number of times, each time with different realisations of the input variables. Each evaluation is called a ‘replication’. The number of replications in a simulation run is called the ‘run length’. Since a new set of realisations is needed for every replication, the number of realisations that needs to be sampled from every stochastic value in the estimates list is equal to the run length.

3. **Impute sample realisations in baseline credit data set**. This step and the following two steps are repeated for every issue definition and for every replication. Here, the component takes the random realisations from the previous step and imputes them into the baseline credit risk dataset at the positions of the values that have been identified to be incorrect. This essentially creates a version of the corrected credit dataset that is discussed in Section 4.1. It only contains the borrower records that have been affected by the data quality issue in question.
4. **Obtain observation of corrected financial measures**. This steps evaluates the system once, using as input the version of the corrected credit dataset that is created in the prior step. The evaluation produces a set of values on the financial measures. Every value is called an 'observation' of a corrected financial measure for an affected record.
5. **Determine which records have enough precision**. With each replication that is performed, each affected record in the estimates list gets an additional observation of every financial measure. At some point, enough of these observations have been collected to make an accurate estimation of the actual corrected value of the most important financial measures. This step assesses for every individual record in the estimates list whether enough observations have been collected. The next replication then excludes the records for which that is the case, which saves performance. The run is done when all records have enough precision or when the maximum number of replications per run is reached.
6. **Determine means of the observations**. The corrected value of a financial measure on a certain record is estimated by taking the mean of the Monte Carlo observations of that measure. This is done for all financial measures on all records.
7. **Determine impacts on financial measures**. The impact on a financial measure is defined in the present research as its corrected value less its baseline value.

Example (continued)

The issue tracking components has determined which records and variables are affected by a data quality issues. The correct value estimate component has estimates what the correct values of the affected variables should be. In this running example, the records of borrowers 1 and 4 are affected by a data quality issue that invalidates the income variable. To reiterate, these are the baseline credit dataset and the correct credit dataset for borrowers 1 and 4:

Baseline credit dataset

Borrower ID	Income	Exposure	Maturity	Foreclosure value
1	€ 6,000	€ 180,000	10 years	€ 120,000
4	€ 4,000	€ 40,000	2 years	€ 100,000

Correct credit dataset with correct value estimates for the income variable

Borrower ID	Income	Exposure	Maturity	Foreclosure value
1	$\mathrm{N}(36200{,}15418)$	€ 150,000	10 years	€ 120,000
4	$\mathrm{N}(36200{,}15418)$	€ 40,000	2 years	€ 100,000

The next step is to calculate the financial measures for borrower 1 and 4 using both the baseline and the correct credit dataset. It is then possible to see the impact of the data quality issue. Since the real capital requirement calculation process is too complex to use as an example, an earlier example introduced the following fictive system as an alternative:

$$\mathrm{EL} = \frac{\mathrm{Exposure}}{\sqrt{\mathrm{Maturity}}} \cdot \max\left(0\,,1 - \frac{0.5 \cdot \mathrm{Income} \cdot \mathrm{Maturity}}{\mathrm{Exposure}}\right)$$

This system is not a linear function of the income variable due to the $\max(\cdot)$ function. It is therefore not valid to take the expected values of the correct value estimates and use those as input to calculate the corrected financial measures. If one would do so anyways, then that would yield the following result:

	Baseline values		Corrected values		Impact on EL
Borrower ID	Income	EL	E[Income]	EL	(baseline – corrected)
1	€ 6,000	€ 47,434	€ 32,600	€ 5,376	€ 42,058
4	€ 4,000	€ 25,456	€ 32,600	€ 5,233	€ 20,223

And now for the simulation approach. As this section defends, that approach yields theoretically valid results even if the the system under study is not linear. The simulation approach draws 200 random realisations from $\mathrm{N}(32600{,}15418)$ and imputes each realisation for the income value of borrower 1. The system is re-evaluated after each imputation. That results in 200 observations of the corrected EL measure for borrower 1, as shown below. The corrected EL measure is then obtained by taking the mean of all these observations.

Replication	Income (realisation from N(36200,15418))	Observation of corrected EL
1	€ 34,048	€ 3,086
2	€ 2,303	€ 53,280
:	:	:
200	€ 35,921	€ 125
	Mean corrected EL measure	**€ 11,562**

The same is done for borrower number 4, yielding the results shown below. The interpretation of these fictive numbers is that the model estimates that the bank can save € 35,872 worth of EL on borrower 1 and € 12,765 worth of EL on borrower 4 by fixing the data quality issue that leads to incorrect values of the income variable.

	Baseline values		Corrected values		Impact on EL
Borrower ID	Income	EL	Income	EL	(baseline – corrected)
1	€ 6,000	€ 47,434	N(32600,15418)	€ 11,562	€ 35,872
4	€ 4,000	€ 25,456	N(32600,15418)	€ 12,691	€ 12,765

4.5.2 Three approaches to handling stochastic input variables

As demonstrated in the example previous example, the correct value estimates that are produced by the correct value estimation component must be imputed in the baseline credit dataset to form the corrected credit dataset. Both datasets are then used as input for the system, resulting in the baseline financial measures and the corrected financial measures. The difference between the two is the impact in the financial measures. However, this is less straightforward than it seems at first, because the correct value estimates are stochastic whereas the system only accepts deterministic values. This research considers three approaches that can be used to evaluate the system anyway; the expected value approach, the analytical approach and the simulation approach.

i. **The expected value approach**. In this approach, one first takes the expected value of every correct value estimate, such that the entire correct credit dataset consists of deterministic values. That dataset is then used as input for the system. This assumes that the following is true:

$$\mathrm{E}[\hat{\mathbf{Y}}] = \mathrm{E}[f(\hat{\mathbf{X}})] = f(\mathrm{E}[\hat{\mathbf{X}}]) \tag{56}$$

Where $\hat{\mathbf{X}}$ is the corrected credit dataset, f is the system and $\hat{\mathbf{Y}}$ is the estimated set of corrected financial measures. The difficulty is that this is only true if the system f is a linear function of its inputs, whereby the inputs must also be independent. More formally, the following transformations of expected values are valid for independent X_1 and X_2:

$$\begin{aligned} &\mathrm{E}[a + b\mathrm{X}] = a + b\mathrm{E}[\mathrm{X}] && \text{for deterministic } a, b \text{ (linearity)} \\ &\mathrm{E}[\mathrm{X}_1 + \mathrm{X}_2] = \mathrm{E}[\mathrm{X}_1] + \mathrm{E}[\mathrm{X}_2] && \text{for independent } \mathrm{X}_1, \mathrm{X}_2 \text{ (additivity)} \end{aligned} \tag{57}$$

However, the system is not linear for most of its inputs, as Appendix 9 shows. Furthermore, several input variables of the system are correlated to some degree. Hence, this approach does not yield theoretically valid correct financial measure estimates.

ii. **The analytical approach**. As second way of evaluating a deterministic system is to find an analytical solution to the expected value formula with $\hat{\mathbf{X}}$ as the input and f and the integrand:

$$\mathrm{E}[\hat{\mathbf{Y}}] = \int_{\forall\, \mathbf{x} \in \hat{\mathbf{X}}} p_{\mathbf{x}} f(\mathbf{x}) d\mathbf{x} \tag{58}$$

This formula integrates f over every possible combination of values in $\hat{\mathbf{X}}$, whereby $p_{\mathbf{x}}$ is the probability mass of a certain combination $\mathbf{x}$. Unfortunately, this integral proves to be impractical to solve for $\hat{\mathbf{X}}$ as the input and f as the integrand because $\hat{\mathbf{X}}$ comprises a multitude of both deterministic and stochastic values with both continuous and discrete distributions. Also, f comprises many discontinuous features such as bucketing which hinder the calculation of the integral. In short, this approach is unfeasible.

iii. **The simulation approach**. Also called the Monte Carlo approach, this approach is a type of numeric analysis. It involves evaluating the system a sufficiently large number of times, each time with different random realisations from the stochastic variables in the corrected credit dataset. Each evaluation is called a ‘replication’ and each replication results in one set of observations of the corrected financial measures. The collection of observations from all replications can then be analysed using statistical tools such as the sample mean and sample variance estimators. This approach does not assume that the system is linear, nor does it require any algebraic transformations of the system.

4.5.3 The simulation approach

Let $N_{\mathbf{R}}$ be the number of replications used in the simulation and let $\widehat{\mathbf{X}}^r$ be the r^{th} realisation of the corrected credit dataset $\widehat{\mathbf{X}}$. Such a realisation is obtained by replacing every stochastic value in the corrected credit dataset by a random realisation of itself. As a result, $\widehat{\mathbf{X}}^r$ solely contains deterministic values that can be used as input for the system. Let f be the system and $\widehat{\mathbf{Y}}^r$ the r^{th} observation of the corrected financial measures, calculated as $\widehat{\mathbf{Y}}^r = f(\widehat{\mathbf{X}}^r)$. The simulation approach assumes that by taking the average of a sufficiently large number of observations of the output variables of a system with stochastic inputs, it is possible to approach the actual expected values of those output variables:

$$\begin{aligned} \mathrm{E}[\widehat{\mathbf{Y}}] = \mathrm{E}[f(\widehat{\mathbf{X}})] &\approx \frac{1}{N_{\mathbf{R}}} \cdot \left(f(\widehat{\mathbf{X}}^1) + \cdots + f(\widehat{\mathbf{X}}^{N_{\mathbf{R}}})\right) \\ &\approx \frac{1}{N_{\mathbf{R}}} \cdot \left(\widehat{\mathbf{Y}}^1 + \cdots + \widehat{\mathbf{Y}}^{N_{\mathbf{R}}}\right) \end{aligned} \tag{59}$$

In the case of SNS Bank N.V., it is possible to evaluate the system for each borrower individually. Letting N_i be the number of replications used to calculate the financial measures for borrower i:

$$\begin{aligned} \mathrm{E}[Y_i] = \mathrm{E}[f(\mathbf{x}_{i,\cdot})] &\approx \frac{1}{N_i} \cdot \left(f(\mathbf{x}_{i,\cdot}^1) + \cdots + f(\mathbf{x}_{i,\cdot}^{N_i})\right) \\ &\approx \frac{1}{N_i} \cdot \left(Y_i^1 + \cdots + Y_i^{N_i}\right) \end{aligned} \tag{60}$$

The simulation approach is computationally intensive since the system has to be evaluated a sufficient number of times to obtain accurate results. A ‘sufficient number of times’ is defined by Law (2007) as the number of replications at which the variance in the target output variable is sufficiently small. This poses two new problems: What is the target output variable and what is ‘sufficiently small’?

The system outputs a large number of financial measures, which are listed in Table 1. Each of these measures serves its own informational purpose to the downstream stakeholders outlined in Chapter 2, but in the context of simulation it is useful to select a limited number of output variable as target output variables and optimise the simulation for those variables. The EL and RWA output variables have been selected to fulfil the role of target output variables because the majority of the financial measures is linearly dependent on either the EL measure or the RWA measure.

The threshold of ‘sufficiently small’ may be expressed using the relative error measure (Law, 2007, pp. 500-503). This is the standard deviation of the sample mean of the observations on the target output variable divided by the absolute sample mean of the observations on that variable. The relative error measure assumes that this sample mean follows a Student t distribution, based on the work of Gossett (1908). As more replications are added, this relative error should decrease until it falls below a threshold value. Law (2007) recommends a threshold value $c \leq 0.15$ and $N_i \geq 10$. The number of replications N_i should be the smallest number that satisfies the following equation for every target variable:

$$\frac{t_{(N_i-1),1-\alpha/2}\sqrt{\frac{s^2}{N_i}}}{|m|} < c \qquad \text{where } m = \frac{1}{N_i} \cdot \sum_{r=1}^{N_i} Y_i^r \quad \text{and } s^2 = \frac{1}{N_i - 1} \cdot \sum_{r=1}^{N_i} (Y_i^r - m)^2 \tag{61}$$

(Law, 2007, p. 504)

4.6 Dashboard creation component

This component takes the baseline credit dataset, the impact estimates list and an Excel template file. It then processes and combines the information into a single Excel file with a number of graphs and tables to visualise, analyse and share the information that is generated by the model.

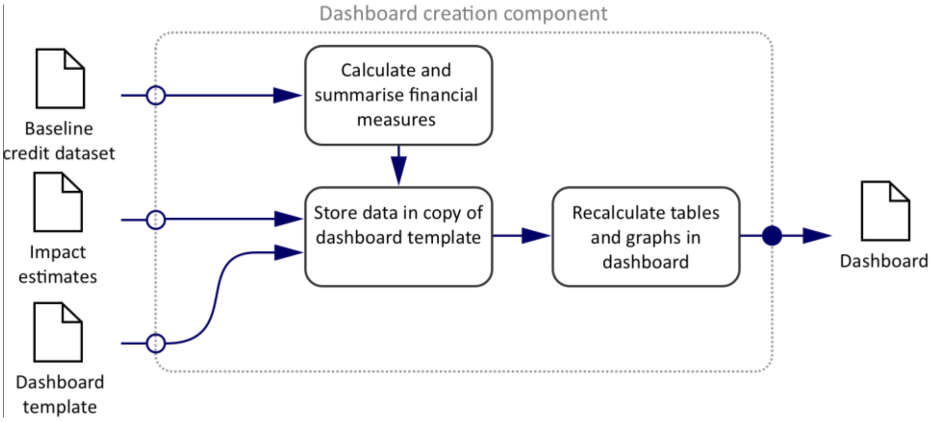

Dashboard creation component
Baseline credit dataset
Impact estimates
Dashboard template
Calculate and summarise financial measures
Store data in copy of dashboard template
Recalculate tables and graphs in dashboard
Dashboard

Figure 18: Schematic overview of the dashboard creation component.

4.6.1 Implementation

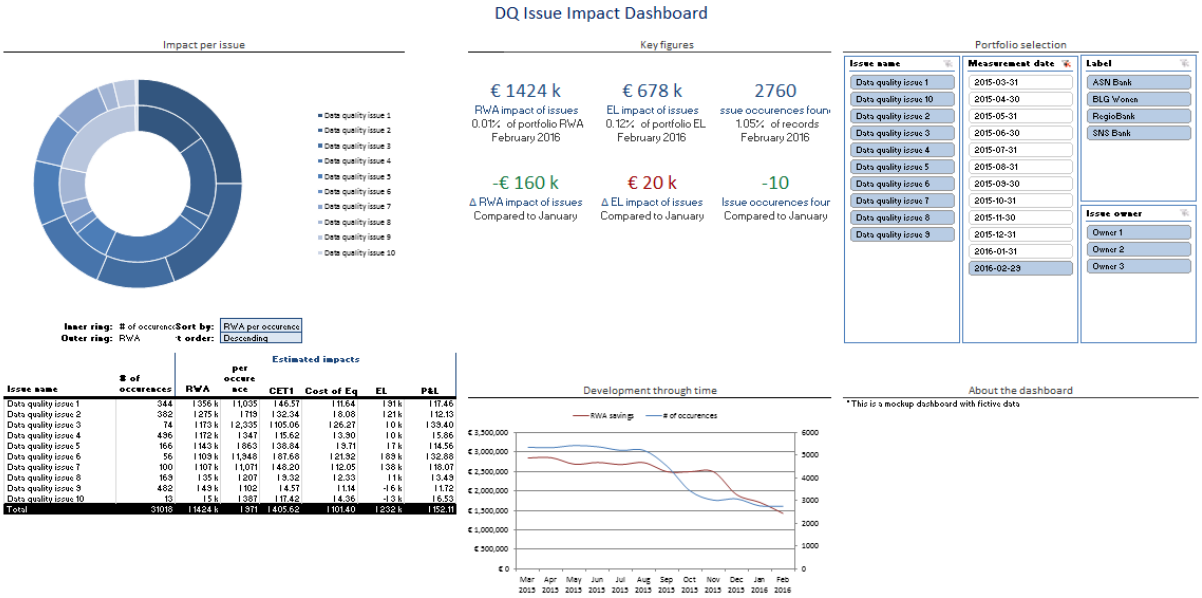
The component executes the following steps:

1. **Calculate and summarise financial measures**. The dashboard uses these summary statistics on the entire portfolio to put the impact numbers in the dashboard into perspective.
2. **Store data in copy of dashboard template**. Most dashboard functionality has been implemented in an 'empty' Excel file that contains only the graphs, tables and formula's that are needed for a proper presentation. Once the model has generated all data, a Matlab function makes a copy of that template and loads the data into the template, resulting in a complete dashboard.
3. **Recalculate tables and graphs in dashboard**. After the data has been loaded into a fresh copy of the dashboard template, Matlab uses an Excel ActiveX call to update the dashboard file. This recalculates all the graphs and tables in the Excel file, such that is immediately presentable to the downstream stakeholders.

When the new dashboard is ready it is stored on a shared file location. The model then uploads a status report of the model run to a SharePoint site. The model owner automatically gets an email to signal that a new model run has been completed and that a new dashboard is ready for approval and distribution.

4.6.2 The dashboard

The model creates a large amount of data, that is hard to quickly comprehend by a user. The information needs to be summarised and visualised to become insightful. This is mainly the job of the dashboard, which has the form of an Excel workbook. The main tab of this workbook is designed using the requirements that have been collected through meetings with the upstream and downstream stakeholders. See Section 2 for a list of the requirements that have been collected. During these meetings, the researcher showed the most recent results from the model in a tentative version of the overview tab in the Excel dashboard. This always resulted in a useful list of possible improvements. After several iterations, the dashboard got the appearance in Figure 19. A larger version of this image is included in Appendix 11. The dashboard comprises five main elements.


DQ Issue Impact Dashboard

Key figures

€ 1424 k
RWA impact of issues
0.01% of portfolio RWA
February 2016

€ 678 k
EL impact of issues
0.12% of portfolio EL
February 2016

2760
ssue occurences foun
1.05% of records
February 2016

-€ 160 k
Δ RWA impact of issues
Compared to January

€ 20 k
Δ EL impact of issues
Compared to January

-10
Issue occurences four
Compared to January

Portfolio selection

Issue name	Measurement date	Label
Data quality issue 1	2015-03-31	ASN Bank
Data quality issue 10	2015-04-30	BLG Wonen
Data quality issue 2	2015-05-31	RegioBank
Data quality issue 3	2015-06-30	SNS Bank
Data quality issue 4	2015-07-31	
Data quality issue 5	2015-08-31	
Data quality issue 6	2015-09-30	
Data quality issue 7	2015-10-31	
Data quality issue 8	2015-11-30	
Data quality issue 9	2015-12-31	
	2016-01-31	
	2016-02-29	

Issue owner
Owner 1
Owner 2
Owner 3

Inner ring: # of occurencSort by: RWA per occurence
Outer ring: RWA 't order: Descending

		Estimated impacts					
Issue name	# of occurences	RWA	per occurence	CET1	Cost of Eq	EL	P&L
Data quality issue 1	344	I 356 k	I 1,035	I 46.57	I 11.64	I 91 k	I 17.46
Data quality issue 2	382	I 275 k	I 719	I 32.34	I 8.08	I 21 k	I 12.13
Data quality issue 3	74	I 173 k	I 2,335	I 105.06	I 26.27	I 0 k	I 39.40
Data quality issue 4	496	I 172 k	I 347	I 15.62	I 3.90	I 0 k	I 5.86
Data quality issue 5	166	I 143 k	I 863	I 38.84	I 9.71	I 7 k	I 14.56
Data quality issue 6	56	I 109 k	I 1,948	I 87.68	I 21.92	I 89 k	I 32.88
Data quality issue 7	100	I 107 k	I 1,071	I 48.20	I 12.05	I 38 k	I 18.07
Data quality issue 8	169	I 35 k	I 207	I 9.32	I 2.33	I 1 k	I 3.49
Data quality issue 9	482	I 49 k	I 102	I 4.57	I 1.14	-I 6 k	I 1.72
Data quality issue 10	13	I 5 k	I 387	I 17.42	I 4.36	-I 3 k	I 6.53
Total	31018	I 1424 k	I 971	I 405.62	I 101.40	I 232 k	I 152.11

About the dashboard

* This is a mockup dashboard with fictive data

Figure 19: The dashboard. A larger version of this image is included in Appendix 11. The numbers are fictive.

i. **Impact per issue (graph)**. These rings give a quick impression of the prevalence and impact of data quality issues. The inner ring represents the number of occurrences of a the data quality issues that are in scope and the outer ring represents the RWA impact of these issues. The stakeholders indicated that this allowed them to quickly see which data quality issues have a relatively low number of occurrences and simultaneously a relatively high impact on RWA. At first sight, these are this issues that are the most attractive to resolve first.

ii. **Impact per issue (table)**. This table shows several key financial impacts for all the data quality issues that are in scope. Not all financial measures from Chapter 3 are shown. Only the measures that have been judged by the downstream stakeholders as being the most insightful have been added

iii. **Key figures**. This area is designed give an overview of the credit data's quality in the most recent month and the prior month using three figures: 'RWA impact of issues', 'EL impact of issues' and 'Issue occurrences found'. These absolute number are put in perspective by comparing them with the portfolio's total RWA, EL and number of records, respectively.

iv. **Development through time** is a graph that plots the number of data quality issue occurrences and their total RWA impact against a time axis. This graph is especially useful with the filtering options that are introduced next.

v. **Portfolio selection** allows a user to filter the data in the dashboard to a certain set of issues, dates, labels, sources and/or issue owners. The stakeholders have found these filtering options to be a useful tool for 'zooming in' on specific data quality issues.

4.7 Summary

The 'Model design' section has introduced the sensitivity analysis approach that is used by the model. It has then shown how the model implementation comprises four consecutive components whereby each component fulfils a distinct task. These components are:

1. The issue tracking component, which uses a list of issue definitions to determine which values in the baseline credit dataset are incorrect due to a data quality issue.
2. The correct value estimation component, which estimates the correct value of the credit data values that have been invalidated by a data quality issue.
3. The impact assessment component, which uses a simulation approach to determine the financial impacts of data quality issues.
4. The dashboard creation component, which organises and visualises all the information that is generated by the model in an Excel dashboard for sharing with downstream stakeholders.

5 Performance

The previous chapter has outlined the model and its components. It is straightforward to assert that the issue tracking component and the dashboard creation component work as intended. That is because they do not make use of any randomness and therefore their output is always the same for every run. In contrast, the correct value estimation component and the impact assessment component make extensive use of random sampling. The output from these components might vary from run to run. Moreover, these two components are much more complex than the first and fourth component, which makes them more prone design or implementation errors. It is therefore sensible to look into the performance of these components. This chapter does so in three sections:

i. Comparison of the correct value estimation procedures.
ii. Assessment of the best correct value estimation procedures.
iii. Significance of the simulation approach for impact assessment.

5.1 Comparison of the correct value estimation component

Section 4.4.4 revealed that the model uses ten distinct correct value estimation procedures to make prediction for every target variable that contains incorrect values. However, only one procedure can be selected as the best procedure for each target variable. That selection is made by letting every procedure make estimates for the target variable in a test set (see Section 4.4.5). A test set contains records that have not been used to train the model, but do contain correct values of the target variable. This makes it possible to see how well the predicted values of a correct value estimation procedure approach the correct values that are available in the test set. That is done using a goodness-of-fit measure. Section 4.4.5 proposed the MSE measure for determining the goodness-of-fit between the observed and predicted values of the target variable and the resulting financial measures. The variable-level goodness-of-fit results are available in the confidential version of this report.

The linear regression, bounded linear regression and overall mean procedures are the most successful at estimating the correct value of the target variable and financial measures. In some cases the MSE is zero and the R^2 is one. These instances suggest that the target variable values in the original test set could be replicated with perfect precision, which raises suspicion. There are several explanations of how that can happen:

i. **The target variable does not affect the financial measure in question**. Some variables are not used as a risk driver and therefore do not have a direct effect on the financial measures that are calculated using the capital requirement calculation process.
ii. **The PD risk measure, which underlies the RWA financial measure, makes use of bucketing**. As a consequence, a correct value estimation procedure does not need to get the target variable value exactly right. Instead, it only needs to estimate the target variable value with enough precision such that the PD value remains in the same bucket.
iii. **The target variable can be expressed as a linear combination of two other variables**.

5.2 Assessment of the best correct value estimation procedures

The previous section revealed which correct value estimation procedures shows the best estimation performance. This section analyses how good 'the best performance' actually is. This performance analysis can be done for every combination of estimation procedure and target variable. For the sake of brevity, the section is limited to two procedures, the bounded linear regression and normal distribution procedures, and one target variable, the LTFV variable. This variable is chosen because it is an important risk driver in the capital requirement calculation process and because it is affected by a relatively large number of data quality issues.

The MSE and the R^2 goodness-of-fit measures are mere summary statistics; they are too coarse for a detailed analysis of the relationship between the observed values in the test set and the predicted values from the correct value estimation procedure. Gelman and Hill (2007, p. 558) outlines two plots that are commonly used to a more detailed analysis:

i. **A scatter plot with the estimated values on the x-axis and the observed values on the y-axis**. A procedure that performs well generates a graph with markings that follow the diagonal line in the centre. Poor performance results in a 'cloud' of markings. Furthermore, this plot allows an analyst to see how well a procedure performs in the tails of a variable.

ii. **A histogram of the residuals**. The residuals are the differences between the estimated and the observed value of the target variable for every record in the test set. This graphs shows how the residuals are distributed and it gives an indication of whether the mean residual is zero (implying zero bias), whether there are extraordinarily large residuals and whether the residuals are normally distributed. Ideally, the residuals have a mean of zero, a limited dispersion and a normal distribution.

Figure 20 shows the aforementioned plots for the out-of-sample performance of the bounded linear regression procedure and normal distribution procedure on the LTFV variable. The red line in the left plot signifies a perfect fit and the red lines in the right plot signify the mean value[7] and the pdf of a fitted normal distribution. As the graphs show, the bounded linear regression procedure provides a much better performance with generally much smaller residuals. In contrast, the normal distribution procedure provides estimates that are close to random, such that this procedure can serve as a baseline for the estimation performance of other procedures.

[7] The OLS estimator that is used to fit a linear regression model by definition results in a mean residual value of zero. However, this is only so for the training set, i.e. the in-sample set, that was used to fit the model. This chapter solely makes use of the test set, i.e. the out-of-sample set, for the purpose of performance measurement.

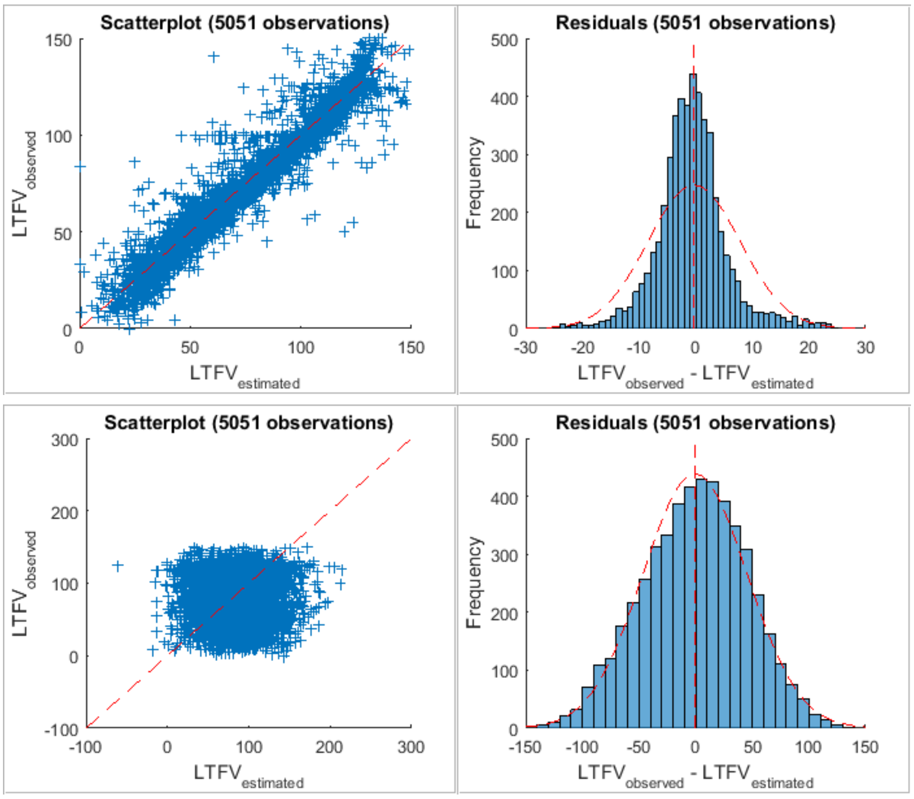


Figure 20: Out-of-sample performance of the bounded linear regression procedure (top) and normal distribution procedure (bottom) on the LTFV variable.

The real test of estimation performance is, however, the goodness-of-fit on the financial measures. The plots for the fits on the RWA financial measure are represented in Figure 21. The fits on the EL financial measure are very similar to the those on the RWA financial measure and have been omitted for brevity. In general, linear regression shows a fair performance since most points in the left plot are on the ideal (red) line. However, there are a few extreme outliers that might signify that the linear regression procedure does not give good results for certain subgroups in the dataset. The normal distribution procedure shows a more dispersed scatterplot and generally larger residuals, which intuitively confirms that the more sophisticated linear regression procedure provides better estimation performance.

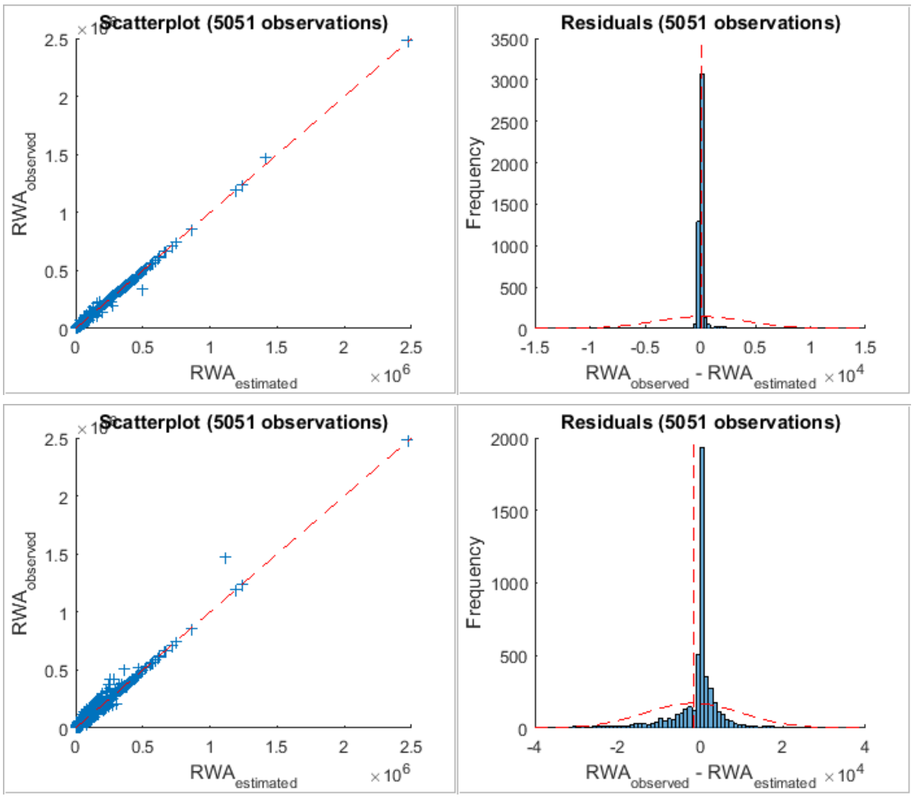


Figure 21: Out-of-sample performance of the bounded linear regression procedure (top) and normal distribution procedure (bottom) on the RWA financial measure.

In summary, the linear regression procedure shows the best estimation performance for almost all target variables, both on the target variable and on the outputted RWA and EL financial measures. However, the plots that show the relationship between the observed values in the test set and the estimated values from the correct value estimation procedure do warrant a further analysis of the cases when linear regression does not provide good estimates.

5.3 Significance of the simulation approach for impact assessment

As section 4.4.5 discussed, the present research considers two feasible approaches for assessing the impact of data quality issues on the financial measures. In short, these approaches are the expected value approach and the simulation approach. The latter approach is theoretically more sound that the former approach, but it is also more complex and computationally intensive. However, when one of the two following conditions is true, it is conceivable that the expected value approach gives the same results as the simulation approach:

i. When the system is linear or close to linear for a certain input variable, then the expected values approach will give the same results as the simulation approach. Given enough replications, the mean observed output values from the simulation approach will tend to the output values from the expected value approach.

ii. When the correct value estimates that are contained in the corrected credit dataset have very narrow confidence intervals, then there will be no practical difference between the correct value estimates as a stochastic value and their respective expected values. In that case, applying the simulation approach would essentially be the same applying the expected value approach an unnecessary large number of consecutive times, leading to the same output for both approaches.

If either or both of the above conditions is true for a certain target variable, then one would expect that there are at most a few differences, or 'mismatches', between the financial measures calculated with the expected value approach and the same measures calculated with the simulation approach. If other words, one would hypothesise that the number of mismatches is below a certain critical value. Appendix 7 treats this analysis with more formal rigour, concluding that it makes sense to keep using the simulation approach over the expected value approach in the model.

5.4 Summary

In short, the 'Performance' chapter has covered the following:

i. Out of the ten correct value estimation procedures that are implemented, the linear regression procedure and bounded linear regression procedure show the best estimation performance for nearly all target variables.

ii. The linear regression procedure shows satisfactory estimation performance on both the target variable and the RWA and EL financial measures, but the performance analysis does warrant a further analysis of the cases when linear regression provides estimates that are off by an extreme amount.

iii. The simulation approach is preferred over the expected value approach for impact assessment.

6 Results

[This chapter is omitted in the public version of this report]

7 Conclusion & next steps

Chapter 1 has introduced the need for a model that can estimate the financial impact of data quality issues in the credit data of SNS Bank N.V. Chapter 2 has identified the users of this model. Chapter 3 introduced the system that is studied by the model, namely the capital requirement calculation process. Chapter 4 showed how the model works and Chapter 5 showed the performance of the model design that is used in the present research. Finally, Chapter 6 showed how the model's output can be used to gain insight into how data quality issues have an impact on financial measures. That insight becomes useful if it reaches those that can actually improve the bank's credit data quality. This is more of a political effort than an academic one and it is beyond the scope of this research. Nonetheless, it is possible to suggest the following steps that can help turn the data quality impact estimation model into a widely accepted source of insights:

i. Make sure that the data quality issue definition list remains up to date.
ii. Present the model's results to higher management.
iii. Make the dashboard more lightweight.

The results from the present research may also find their use outside SNS Bank N.V., perhaps even at other institutions than banks. That is because an organisation's data quality is nowadays often assessed using expert opinion or using elaborate questionnaires (Carlo, Cinzia, Chiara, & Andrea, 2009). These types of assessments can be time-consuming and hard to automate. There are efforts to build models that can quantitatively assess data quality. For instance, Borek, Parlikad, Woodall, and Tomasella (2014) have developed a risk-based model that quantifies the business impact of data quality issues at a manufacturing company. They have received encouraging feedback from the six companies at which they have applied their model. However, there is no one-size fits all model for quantifying data quality (Woodall, Borek, & Parlikad, 2013). The model that has been developed in the present research can serve as an expansion of the choice of the models for those who seek a way to measure the impact of data quality issues on their system. It can be used for cases wherein the input data, the system and output variables are well specified, wherein it is clear how costly a change in the output variables is, and wherein a mock-up version of the system can be run automatically for a large number of times at no significant cost. The capital requirement calculation process in this research is just one case that fits that description. As an example, some other cases that also fit this description are a delivery service's vehicle routing system, an airline's revenue management system and a hospital staff planning system.

8 Discussion & further research

The model that has been developed in this research has been assessed theoretically by employees of the Modelling department. The model outputs have been verified empirically by employees of the Credit Risk Retail department. The code of the Matlab implementation of the model has been reviewed by one of this research's supervisors. However, there are few checks and balances that could not be done due to the constraints of time. Three desirable checks are to assess the model performance with cleansed data, to estimate the correct financial measures directly and to use cross-validation to determine the training set size. These checks are elobarately discussed below.

i. **Estimating the correct financial measures directly**. The current model implementation first estimates the correct value of every affected value in the baseline credit dataset. The correct value estimates are imputed in the credit dataset to form the corrected credit dataset. That set is used in turn to determine the corrected financial measures. It may be possible to bypass a few steps if the correct value estimation component is trained to estimate the correct value of the financial measures directly. In particular, this would bypass the computationally intensive simulation approach in the impact assessment step. It is currently unknown whether this approach yields a satisfactory estimation performance.
ii. **Using cross-validation to determine the training set size**. The correct value estimation component uses a training set size of 5,000 rows by default. This set size is an educated guess that keeps the required computation time within reasonable bounds. One way to measure the appropriateness of this training set size is to apply cross-validation. This analysis trains the same model a large number of times, each time with a newly sampled training set of the same set size. The analysis determines whether the coefficients that are estimated by the correct value estimation procedure remain sufficiently stable from training instance to training instance. If that is the case, then the training set size is sufficiently large.

Moreover, the present report is based on a choice to only include those risk measures, capital measures and financial measures that could be directly calculated from the credit risk data. There may be other benefits to monitoring and improving the quality of credit data, both monetary and non-monetary. Five possible benefits are as follows:

i. **Better regulatory compliance**. The ECB requires banks to maintain a certain data quality standard for their credit data quality. That quality is examined by the ECB at irregular intervals and failure to comply to the quality standards can result in costly fines. Furthermore, banks will soon be required to adhere to the Principles for Effective Risk Data Aggregation and Risk Reporting (PERDARR) (BCBS, 2013). This is a set of principles that among other things stipulates that a bank must be able to quickly compile reports on the data quality of its credit data. The present model can be used as a tool for compiling such reports.
ii. **Reduced margin of conservatism**. Conversely, if the ECB finds that the credit data quality at the bank is of good quality or if the bank can show that it is in control of the flaws in data quality, then the ECB may choose to allow a lower Margin of Conservatism. This directly reduces capital requirements. See Section 3.1 for an explanation of this margin and how it increases the amount of capital that must be held.

iii. **Improved odds of refuting legal claims**. Like most large organisations, SNS Bank N.V. often faces legal claims, e.g. from dissatisfied customers. The Legal Affairs department has indicated that some cases are lost because the department is not able to collect the required data, e.g. on the behaviour or a customer, to serve as proof in court.

iv. **More targeted product development**. As a company, the bank is interested in knowing which customer groups earn the best profit and which customer groups only incur expenses. Such knowledge can be used for targeted marketing and product development. Within the confines of the privacy law, of course.

v. **Easier model development**. During the development of models such as credit risk models, much time is spent on investigating the data quality issues in the model's source data and devising ways to cope with those issues. Conversely, many of these costly man hours can be saved when the data that is used as input for the bank's models is of better quality.

The benefits that are stated above are much less trivial to quantify than the financial measures that are used in the present report. However, they do show that the financial impacts that have been considered in this research are actually just the tip of the iceberg.

References

Abel, A., & Bernanke, B. (2001). *Macroeconomics* (4th ed.). Boston: Addison-Wesley.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Paper presented at the 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR.

Allen & Overy. (2014, January). Capital Requirements Directive IV Framework Internal Ratings Based Approach to Credit Risk in the Banking Book. *Allen & Overy Client Briefing, 4*.

Babel, B., Gius, D., Gräwert, A., Lüders, E., Natale, A., Nilsson, B., & Schneider, S. (2012, November). Capital Management: Banking's new imperative. *McKinsey Working Papers on Risk, 38*.

BCBS. (2000). *Principles for the Management of Credit Risk - final document*. Retrieved from Basel: http://www.bis.org/publ/bcbs75.htm

BCBS. (2004). International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

BCBS. (2005a). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

BCBS. (2005b). An Explanatory Note on the Basel II IRB Risk Weight Functions.

BCBS. (2006, September). The IRB Use Test: Background and Implementation. *Basel Committee Newsletters, 9*.

BCBS. (2011a). Basel III: A global regulatory framework for more resilient banks and banking systems.

BCBS. (2011b). International regulatory framework for banks (Basel III). Retrieved from http://www.bis.org/bcbs/basel3.htm

BCBS. (2013). *Principles for effective risk data aggregation and risk reporting*. Retrieved from Basel: http://www.bis.org/publ/bcbs239.pdf

Borek, A., Parlikad, A. K., Woodall, P., & Tomasella, M. (2014). A risk based model for quantifying the impact of information quality. *Computers in Industry, 65*(2), 354-366. doi:10.1016/j.compind.2013.12.004

Carlo, B., Cinzia, C., Chiara, F., & Andrea, M. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv., 41*(3), 1-52. doi:10.1145/1541880.1541883

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 20*(2), 215-242.

Daniel, N. D., Denis, D. J., & Naveen, L. (2008). Do firms manage earnings to meet dividend thresholds? *Journal of Accounting and Economics, 45*(1), 2-26. doi:http://dx.doi.org/10.1016/j.jacceco.2007.11.002

Deloitte. IFRS 9 — Financial Instruments. *IASPlus.* Retrieved from http://www.iasplus.com/en/standards/ifrs/ifrs9

EBA. (2013). *Capital Requirement Regulation (CRR)*. Retrieved from https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/-/interactive-single-rulebook/toc/2

ECB. (2014). *Aggregate Report on the Comprehensive Assessment*. Retrieved from Frankfurt: https://www.bankingsupervision.europa.eu/banking/comprehensive/html/index.en.html

ECB. (n.d.). Minimum reserves. Retrieved from https://www.ecb.europa.eu/mopo/implement/mr/html/index.en.html

EY. (2014). Impairment of Financial instruments under IFRS 9. *Applying IFRS*.

Fischer, R., Law, S., De Demandolx, O., Broeskamp, U., & Togashi, N. (2015). *Low interest rates: Six actions for retail banks to overcome the impasse*. Retrieved from New York:

http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/2015/oct/2015_Oliver_Wyman_Low_interest_rates.pdf

Gelman, A., & Hill, J. (2007). *Data analysis using regressin and multilevel/hierarchical models*. Cambridge New York: Cambridge University Press.

Gossett, W. S. (1908). The Probable Error of a Mean. *Biometrika, 6*, 1-25.

Hanmanth, M. N., Shivaji, Waghamare. (2014). *Risk Mangement in Banks*. Mumbai: Ashok Yakkaldevi.

Hull, J. C. (2012). *Risk Management and Financial Institutions* (3rd Ed.). Hoboken, New Jersey: Wiley.

Investopedia. (n.d.). Return on Equity. Retrieved from http://www.investopedia.com/terms/r/returnonequity.asp

Larsen, R. J., & Marx, M. L. (2012). *An Introduction to Mathematical Statistics and Its Applications (International Edition)* (5th ed.): Prentice Hall.

Law, A. (2007). *Simulation modelling and analysis*. Boston: McGraw-Hill.

Moges, H.-T., Dejaeger, K., Lemahieu, W., & Baesens, B. (2013). A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management, 50*(1), 43-58. doi:http://dx.doi.org/10.1016/j.im.2012.10.001

Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333-380.

Neyman, J., & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*, 289-337.

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity, 41*(5), 673-690. doi:10.1007/s11135-006-9018-6

Oakley, J. E. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66*(3), 751-769. doi:10.1111/j.1467-9868.2004.05304.x

Rhys, M., & Mickeler, J.-M. (2015). *Finding Your Way*. Retrieved from http://www.iasplus.com/en/publications/global/surveys/fifth-global-ifrs-banking-survey/file

Roeleven, G. (2015, 2015, March 27) *Kwaliteit data van banken onder de loep/Interviewer: D. N. Bank*.

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research, 33*(4), 545-571. doi:10.1207/s15327906mbr3304_5

Schwartz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics, 6*, 461-464.

SNS Bank N.V. (2015a). *2014 Annual Report*. Retrieved from http://www.snsreaal.nl/investors/reports/annual-reports-sns-bank.html

SNS Bank N.V. (2015b). SNS Bank issues € 500 million subordinated debt [Press release]. Retrieved from https://www.snsbanknv.nl/en/news/2015/10/29/sns-bank-issues-500-million-subordinated-debt

SNS Bank N.V. [internal]. (2015). *Cost of Equity*. Retrieved from

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, 103*(2684), 677-680. doi:doi:10.1126/science.103.2684.677

Veenman, F., Steenmeijer, C., Smeets, R., & van Poucke, A. (2015). *The state of Dutch banks in 2015*. Retrieved from http://www.kpmg.com/nl/nl/issuesandinsights/articlespublications/pages/the-state-of-dutch-banks-in-2015.aspx

Verbeek, M. (2004). *A guide to modern econometrics* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Weisstein, E. W. (n.d.). Statistical Median.   Retrieved from http://mathworld.wolfram.com/StatisticalMedian.html

Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: The Hybrid Approach. *Information & Management, 50*(7), 369-382. doi:http://dx.doi.org/10.1016/j.im.2013.05.009

Index

Table of tables

Table of figures

Appendix 1 Data quality issue definition list

[This appendix is omitted in the public version of this report]

Appendix 2 Unexpected loss formula

Basel II prescribes formulae for the calculation of Unexpected Loss (UL) on residential mortgage loans and SME loans for the purpose of capital requirement calculation. The formula for the UL on residential mortgage loans that are not in default is (BCBS, 2005a, §328):

$$\mathrm{UL} = \mathrm{LGD}_{\mathrm{DT}} \cdot \left(\sqrt{\frac{1}{1-\rho}} \cdot \Phi^{-1}(\mathrm{PD}) + \sqrt{\frac{\rho}{1-\rho}} \cdot \ \Phi^{-1}(99.9\%) - \mathrm{PD} \right) \cdot \mathrm{EAD}$$

$$\text{where } \rho = 0.15$$

In this formula, Φ is the cumulative normal distribution function and Φ^{-1} is the inverse cumulative normal distribution function.

Appendix 3 Description of the baseline credit dataset

[This appendix is omitted in the public version of this report]

Appendix 4 Derivation of the OLS estimator

Under the model assumptions of a linear regression model one must find the $\hat{\beta}$ that maximises the following likelihood (Verbeek, 2004, p. 164):

$$\max_{\hat{\beta}} L\left(\hat{\beta}\middle|y_i, x, \sigma^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{\left(y_i - x_i\hat{\beta}\right)^2}{\sigma^2}\right\}$$

$$\max_{\hat{\beta}} \log L\left(\hat{\beta}\middle|y_i, x, \sigma^2\right) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N} \frac{\left(y_i - x_i\hat{\beta}\right)^2}{\sigma^2}$$

$$\min_{\hat{\beta}} \log L\left(\hat{\beta}\middle|y_i, x\right) = \sum_{i=1}^{N} \left(y_i - x_i\hat{\beta}\right)^2 = \sum_{i=1}^{N} y_i^2 - 2y_i x_i\hat{\beta} + \hat{\beta}^2$$

$$\min_{\hat{\beta}} \log L\left(\hat{\beta}\middle|y_i, x\right) = \sum_{i=1}^{N} x_i^2\hat{\beta}^2 - \sum_{i=1}^{N} 2y_i x_i\hat{\beta}$$

Which involves solving the first order equation:

$$\frac{\partial \log L\left(\hat{\beta}\middle|y_i, x\right)}{\partial\hat{\beta}} = 2\sum_{i=1}^{N} x_i^2\hat{\beta} - \sum_{i=1}^{N} 2y_i x_i = 0$$

$$2\sum_{i=1}^{N} x_i^2\hat{\beta} = 2\sum_{i=1}^{N} y_i x_i$$

$$\hat{\beta} = \left(\sum_{i=1}^{N} x_i^2\right)^{-1} \sum_{i=1}^{N} y_i x_i$$

The derivation above assumes a univariate model without an intercept. The derivation works analogously for the case multivariate case whereby one covariate is 1 for all i to form the intercept.

Appendix 5 Bounded linear regression link functions and distribution

The pair of link functions

$$z_i = G(y_i) = \log\left(\frac{y_i - a}{b - y_i}\right) \quad , \quad y_i = G^{-1}(z_i) = \frac{a + b \cdot \exp(z_i)}{1 + \exp(z_i)}$$

maps a value $(y_i \in \mathbb{R} \mid a \leq y_i \leq b)$ to a value $(z_i \in \mathbb{R} \mid -\infty < \mathrm{z_i} < \infty)$ and vice versa, since the functions are strictly increasing and:

$$\log\left(\frac{y_i - a}{b - y_i}\right)\Bigg|_{y_i=a} = \log(0) = -\infty$$

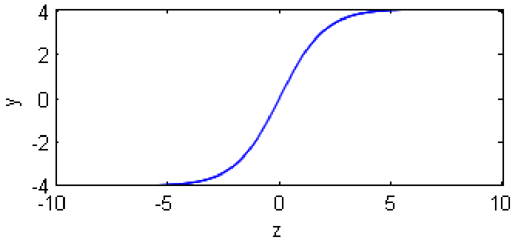
$$\log\left(\frac{y_i - a}{b - y_i}\right)\Bigg|_{y_i=b} = \log(\infty\,) = \infty$$

and vice versa:

$$\lim_{z_i \to -\infty} \frac{a + b \cdot \exp(z_i)}{\exp(z_i) + 1} = \frac{a + b \cdot 0}{0 + 1} = \frac{a}{1} = a$$

$$\lim_{z_i \to \infty} \frac{a + b \cdot \exp(z_i)}{\exp(z_i) + 1} = \frac{a + b \cdot \infty}{\infty + 1} = \frac{b \cdot \infty}{\infty} = b$$

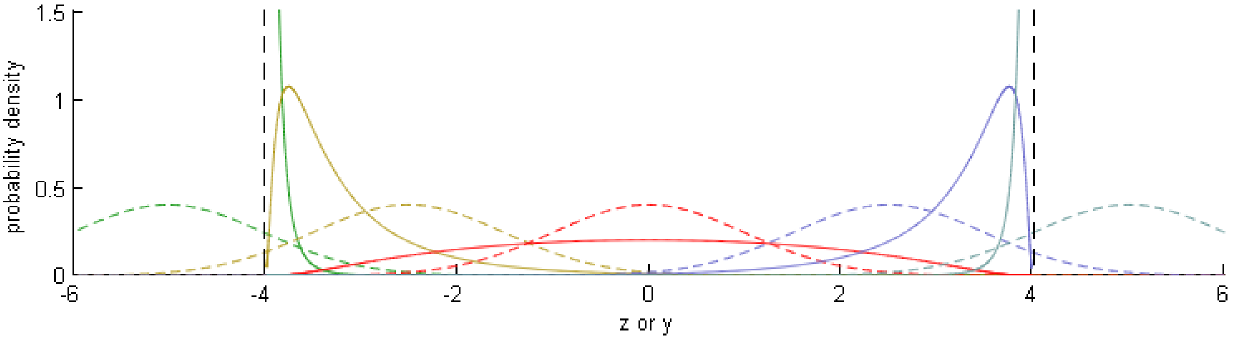
The following figure displays how the link functions map x to z and vice versa when $a = -4$ and $b = 4$.



When the underlying unbounded linear regression model predicts a value $\hat{z}_i$ with mean $\mu_{\hat{z}_i}$ and standard deviation $\sigma_{\hat{z}_i}$, then the probability distribution of $\hat{y}_i$ is a transformation of $\hat{z}_i$:

$$\hat{y}_i \sim \frac{a + b \cdot \exp(\hat{z}_i)}{1 + \exp(\hat{z}_i)} \quad \text{where} \quad \hat{z}_i \sim N(\mu_{\hat{z}_i}, \sigma_{\hat{z}_i})$$

A program can sample from this distribution by sampling from $\hat{z}_i$ and converting the sampled value using G^{-1}. The figure below shows this principle. The dotted line shown the pdf's of five normally distributed $\hat{z}_i$ with$\sigma_{\hat{z}_i} = 1$ and $\mu_{\hat{z}_i} = -5, -2.5, 0, 2.5$ and 5 respectively. The continuous line in the same colour shows the transformation of $\hat{z}_i$ to $\hat{y}_i$.



Furthermore, the derivative of $G(y_i)$ with respect to y_i is:

$$\frac{d}{dy_i} G(y_i) = \frac{b-a}{(y-a)\cdot(b-y)}$$

Appendix 6 Distribution of the sample mean

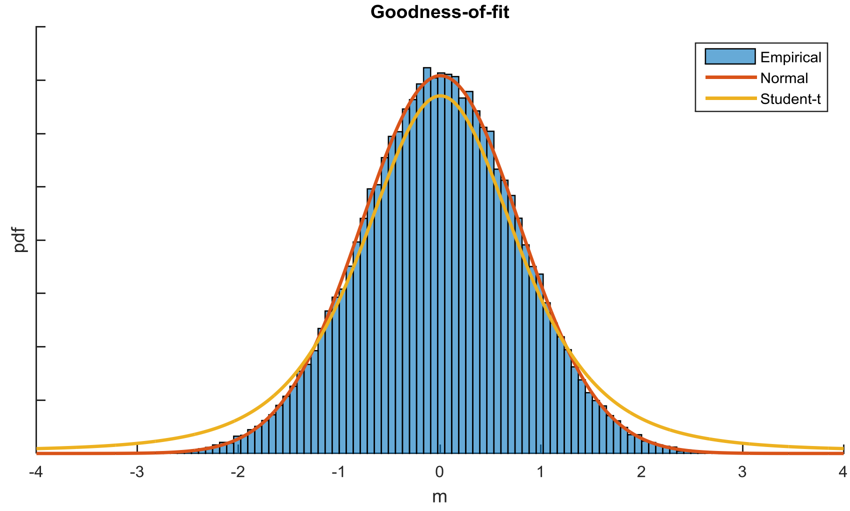
Introduction Gossett (1908), working under his pseudonym ‘Student’, has shown that the difference between the true mean of a stochastic variable and the sample mean of sample of n observations from that stochastic variable follows a Student-t distribution with $n-1$ degrees of freedom. This should not be confused with the distribution of the sample mean itself, which is asymptotically Normal by the CLT, as this empirical analysis shows.

Method Let $X \sim \mathrm{Normal}(\mu_X, \sigma_X)$ be a standard Normal distributed stochastic variable with mean 0 and standard deviation 2. Then let $X_1, \ldots, X_5$ be a sample of five independent realisations from X. The sample mean, its theoretical expected value and its theoretical variance are:

$$m = \frac{1}{5} \cdot \sum_{i=1}^{5} X_i \quad , \quad \mathrm{E[m]} = \mathrm{E}[X_1 + \cdots + X_5] = 0 \quad , \quad \mathrm{Var}[m] = \frac{1}{5^2} \cdot \mathrm{Var}[X_1 + \cdots + X_5] = \frac{5}{5^2} \cdot \mathrm{Var}[X]$$

Under the null hypothesis, m has a Normal distribution and under the alternative hypothesis, m has a different distribution, e.g. the Student-t distribution. This is tested by drawing the sample $X_1, \ldots, X_5$ and determining the sample mean m a very large number of times. This makes it possible to create a histogram of the realisations of m. The pdf of the Student-t distribution and of the Normal distribution are then plotted over the histogram using the theoretical expected value and theoretical variance of the sample mean. The null hypothesis is rejected if the histogram does not follow the Normal distribution plot.

Analysis Using 100,000 realisation of m:



Conclusion The histogram clearly shows a better fit with the Normal distribution than with the Student-t distribution. This may be verified using a χ^2 goodness-of-fit test, but it is safe to visually infer from this graph that the null hypothesis is not rejected.

Appendix 7 The significance of the simulation approach

Introduction The system under study, which is the capital requirement calculation process, is not linear for all of its input variables. Also, the correct values estimates that are used as input are stochastic instead of deterministic. This poses the problem of how to system should evaluated, since the system only accepts deterministic values. Two of the approaches that are described into detail in Section 0 are as follows:

i. **The expected value approach**. Take the expected value of every input value and use that as input for the system. This will not yield theoretically valid output values if the system is not linear.
ii. **The simulation approach**. Sample a large number of realisations from every stochastic input value and evaluate the system with every realisation, yielding one observation of the output value per system evaluation. Every system evaluation is called a 'replication'. The estimated output value is then the mean of all observations of the output value. This approach is theoretically valid if sufficient replications are used.

The simulation approach is preferred to over the expected value approach because it is theoretically more sound. However, the simulation approach is more complex to implement and more computationally intensive than the expected value approach, since the system is evaluated many times. It should therefore only be used if it yields significantly different outputs compared to the expected value approach. This analysis aims to determine whether the outputs are indeed different at a significant level.

Method The most important target variable that is calculated by the model is the RWA impact variable, which measures the increase in Risk Weighted Assets on a borrower's loans as an effect of a data quality issue. The expected value approach yields points estimates of the RWA impact and the simulation approach yields stochastic estimates with a Student t-distribution, which makes it possible to construct a 95% confidence interval for every estimate of the RWA impact. The system is evaluated for every borrower record individually and the system yields one RWA impact figure for every records. If the expected value approach and the simulation approach yield similar values, then one would expect that around 95% of all estimates from the expected value approach falls in within the 95% confidence interval bounds of the estimates that are produced by the simulation approach. Conversely one would expected that around 5% of the estimates from the expected value approach does not fall within the confidence interval bounds of the simulation approach.

Formalising that expectation, the number of 'mismatches' k between the expected value approach's outputs and the simulation approach's output follows a Binomial distribution where n is the number of records for with the RWA impact is calculated and $p = 0.05$. This distribution can be approached by a Poisson distribution with $\lambda = np = 0.05n$:

$$k \sim \text{Binomial}(n, 0.05) \approx k \sim \text{Poisson}(0.05n)$$

Formalising the null hypothesis H_0, the alternative hypothesis H_a and the test statistic T:

$$H_0: k \leq 0.05p$$
$$H_a: k > 0.05p$$

$$\alpha = 0.05$$

$$\mathrm{T} = e^{-0.05p} \sum_{i=k}^{\infty} \frac{(0.05p)^i}{i!} = 1 - e^{-0.05p} \sum_{i=0}^{k-1} \frac{(0.05p)^i}{i!}$$

In words, the null hypothesis is that the RWA impact estimates from the expected value approach are similar to the RWA impact estimates from the simulation approach. The alternative hypothesis is that estimates are not similar. If the null hypothesis is rejected, then it make sense to use the simulation approach instead of the expected value approach. The test statistic is the likelihood that k mismatches or more are observed under the null hypothesis. Reject the null hypothesis if $T < \alpha$.

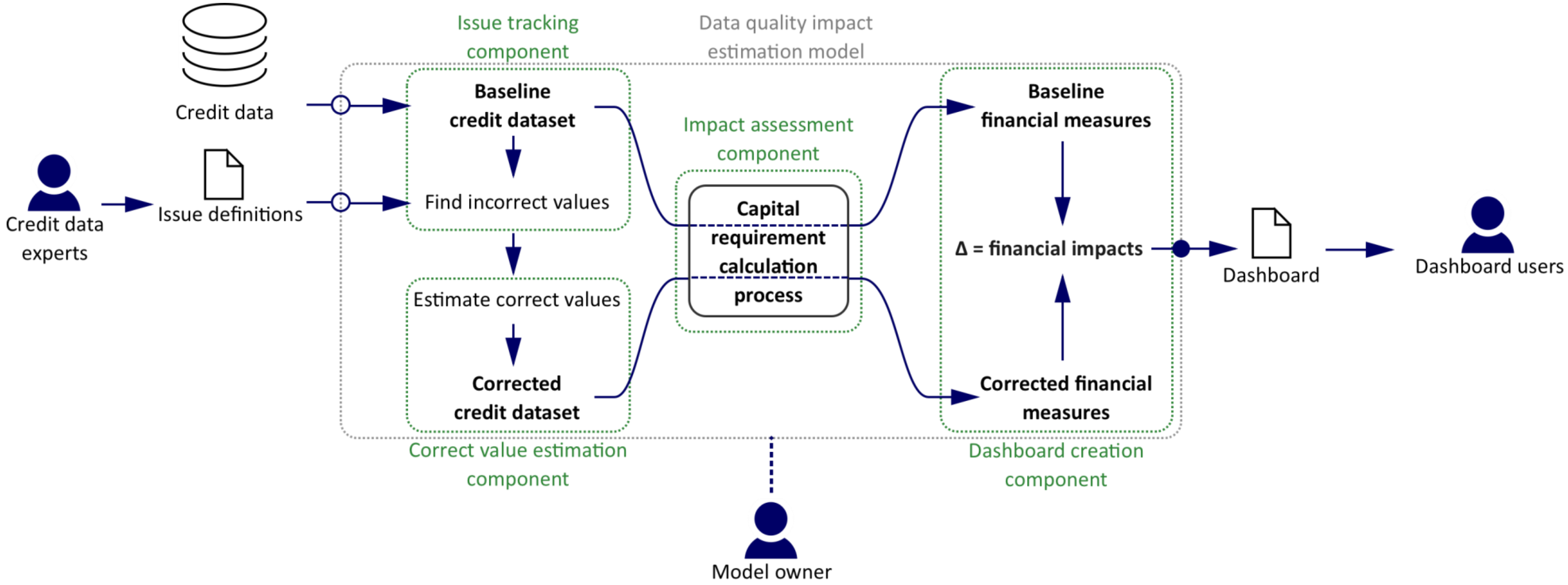
Analysis [Omitted in the public version of this report]

Conclusion It makes sense to keep using the simulation approach over the expected value approach in the model.

Appendix 8 Out-of-sample estimation performance on LTFV

[This appendix is omitted in the public version of this report]

Appendix 9 Model overview


Issue tracking component
Data quality impact estimation model
Credit data
Credit data experts
Issue definitions
Baseline credit dataset
Find incorrect values
Estimate correct values
Corrected credit dataset
Impact assessment component
Capital requirement calculation process
Baseline financial measures
Δ = financial impacts
Corrected financial measures
Dashboard
Dashboard users
Correct value estimation component
Dashboard creation component
Model owner

Appendix 10 Univariate analysis of system linearity

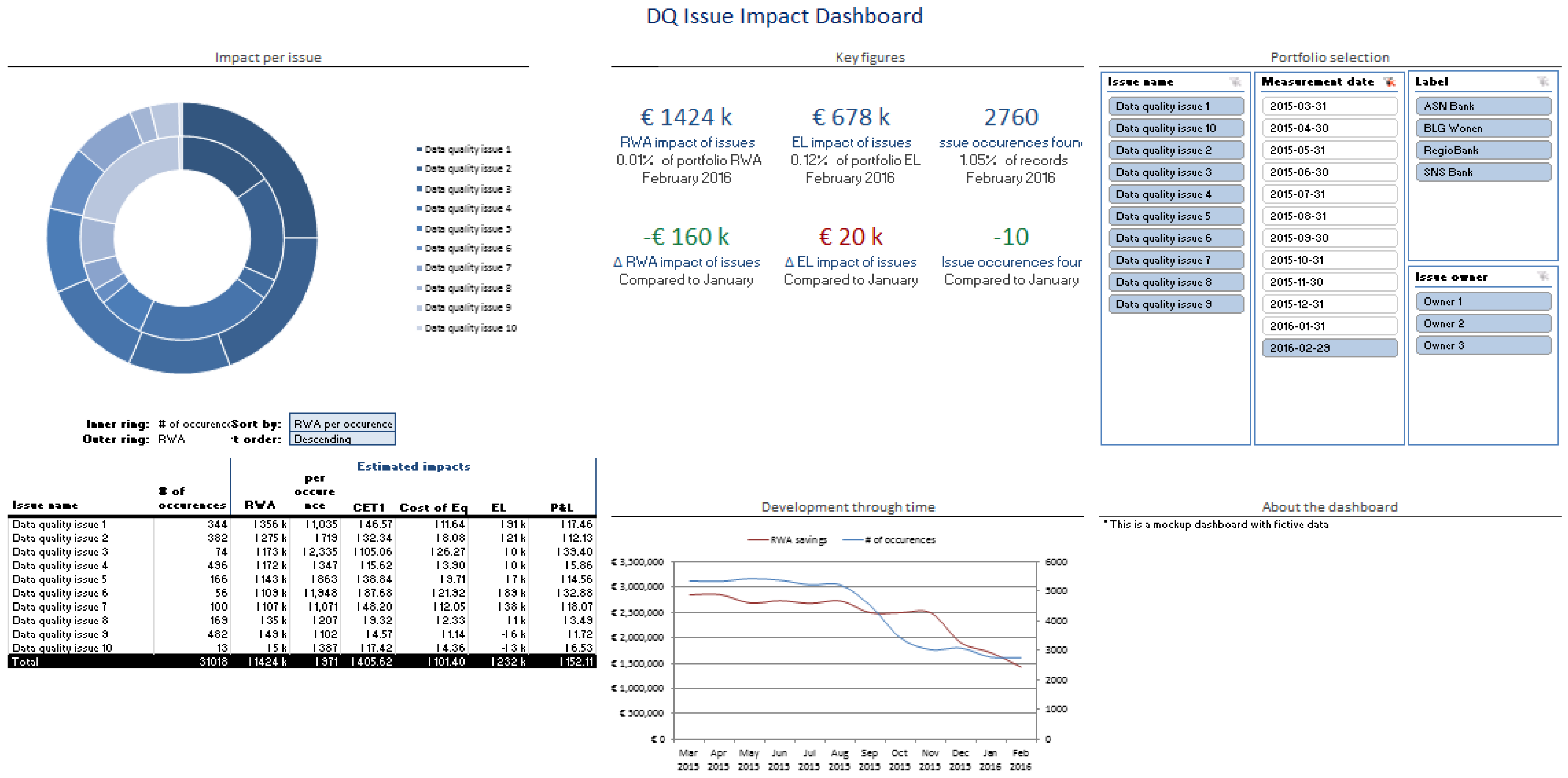
Introduction This analysis takes a random sample of 10000 records from the credit risk data and scales the values on every individual input variable of the credit risk models (the risk drivers) by a certain factor, holding the remaining input variables constant. The credit risk models are evaluated for every combination of input variable and scaling factor. The mean of the resulting output variables are then plotted in a line chart. If the system of credit risk models were linear for every individual input variable, then this analysis would show a straight line or ‘no effect’ (i.e. a flat line) for every input variable/output variable combination.

Only variables at the interval and ratio measurement level have been included. Discrete variables are rounded to the next nearest integer after scaling.

[The remainder of this appendix is omitted in the public version of this report]

Appendix 11 Example of the dashboard Excel file

[Contains fictive numbers]



DQ Issue Impact Dashboard

Impact per issue

Inner ring: # of occurenc Sort by: RWA per occurence
Outer ring: RWA t order: Descending

		Estimated impacts					
Issue name	# of occurences	RWA	per occurence	CET1	Cost of Eq	EL	P&L
Data quality issue 1	344	I 356 k	I 1,035	I 46.57	I 11.64	I 91 k	I 17.46
Data quality issue 2	382	I 275 k	I 719	I 32.34	I 8.08	I 21 k	I 12.13
Data quality issue 3	74	I 173 k	I 2,335	I 105.06	I 26.27	I 0 k	I 39.40
Data quality issue 4	496	I 172 k	I 347	I 15.62	I 3.90	I 0 k	I 5.86
Data quality issue 5	166	I 143 k	I 863	I 38.84	I 9.71	I 7 k	I 14.56
Data quality issue 6	56	I 109 k	I 1,948	I 87.68	I 21.92	I 89 k	I 32.88
Data quality issue 7	100	I 107 k	I 1,071	I 48.20	I 12.05	I 38 k	I 18.07
Data quality issue 8	169	I 35 k	I 207	I 9.32	I 2.33	I 1 k	I 3.49
Data quality issue 9	482	I 49 k	I 102	I 4.57	I 1.14	-I 6 k	I 1.72
Data quality issue 10	13	I 5 k	I 387	I 17.42	I 4.36	-I 3 k	I 6.53
Total	31018	I 1424 k	I 371	I 405.62	I 101.40	I 232 k	I 152.11

Key figures

€ 1424 k
RWA impact of issues
0.01% of portfolio RWA
February 2016

€ 678 k
EL impact of issues
0.12% of portfolio EL
February 2016

2760
ssue occurences foun
1.05% of records
February 2016

-€ 160 k
Δ RWA impact of issues
Compared to January

€ 20 k
Δ EL impact of issues
Compared to January

-10
Issue occurences four
Compared to January

Development through time

Portfolio selection

Issue name
Data quality issue 1
Data quality issue 10
Data quality issue 2
Data quality issue 3
Data quality issue 4
Data quality issue 5
Data quality issue 6
Data quality issue 7
Data quality issue 8
Data quality issue 9

Measurement date
2015-03-31
2015-04-30
2015-05-31
2015-06-30
2015-07-31
2015-08-31
2015-09-30
2015-10-31
2015-11-30
2015-12-31
2016-01-31
2016-02-29

Label
ASN Bank
BLG Wonen
RegioBank
SNS Bank

Issue owner
Owner 1
Owner 2
Owner 3

About the dashboard

* This is a mockup dashboard with fictive data