Contributions and Applications in Loan Loss Provisioning, Stress Testing, and Visual Analytics

Von der Wirtschaftswissenschaftlichen Fakultät der
Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften
– Doktor rerum politicarum –
– Dr. rer. pol. –

genehmigte Dissertation
von

M.Sc. Nikolas Friedrich Siegfried Stege
geboren am 04. Oktober 1985 in Groß-Gerau

2023

Referent/Referentin:	Prof. Dr. Michael H. Breitner
Koreferent/Koreferentin:	Prof. Dr. Stefan Wielenberg
Tag der Promotion:	07.12.2023

Für Janine

Abstract

This cumulative dissertation summarizes and discusses six research articles that are either published in academic journals and conference proceedings or submitted for review. The topics described are cross-disciplinary and can be allocated to Accounting, Finance, and Information Systems Research. In Accounting, we analyze the methodological differences between ratings and lifetime default risk to develop a proof for the use of rating changes for the determination of significant increases in credit risk in accordance to the impairment requirements of the International Financial Reporting Standards. Our results and findings contribute to more transparency with regard to decision-relevant information for stakeholders of financial statements. In Finance, we combine machine learning techniques with cointegration analysis to produce adequate projections of macroeconomic variables for stress testing exercises. Our results and findings have practical relevance for risk managers in the financial services industry and help to validate the execution of stress tests and to ensure compliance. In Information Systems Research, we develop a general process model and visualization framework to identify and highlight unusual data in subsets for further investigation. Our process model and visualization framework empower domain experts and data analysts to jointly gain and discuss insights from underlying data. Our results and findings show that both our process model and visualization framework contribute to interactive visual analytics, storytelling, and well-founded decision support.

Keywords. Loan Loss Provisioning, Stress Testing, Visual Analytics.

Zusammenfassung

Diese kumulative Dissertation diskutiert und fasst sechs Forschungsartikel zusammen, die entweder in wissenschaftlichen Zeitschriften und Tagungsbänden veröffentlicht oder eingereicht wurden. Die beschriebenen Themen sind fächerübergreifend und lassen sich den Bereichen Rechnungslegung, Finanzierung und Wirtschaftsinformatik zuordnen. Im Bereich Rechnungslegung analysieren wir die methodischen Unterschiede zwischen Ratings und Restlaufzeit-Ausfallrisiko, um einen Nachweis für die Verwendung von Ratingänderungen zur Bestimmung signifikanter Erhöhungen des Kreditrisikos gemäß den Impairment-Anforderungen der International Financial Reporting Standards zu erarbeiten. Unsere Ergebnisse und Erkenntnisse tragen zu mehr Transparenz hinsichtlich entscheidungsrelevanter Informationen für Stakeholder von Jahresabschlüssen bei. Im Bereich Finanzierung kombinieren wir Techniken des maschinellen Lernens mit Kointegrationsanalysen, um angemessene Projektionen makroökonomischer Variablen für Stresstests zu erstellen. Unsere Ergebnisse und Erkenntnisse haben praktische Relevanz für Risikomanager in der Finanzdienstleistungsbranche und helfen, die Durchführung von Stresstests zu validieren und deren Compliance sicherzustellen. Im Bereich Wirtschaftsinformatik entwickeln wir ein allgemeines Prozessmodell sowie ein Visualisierungsverfahren, um ungewöhnliche Daten in Teilmengen für weitere Untersuchungen zu identifizieren und hervorzuheben. Unser Prozessmodell und unser Visualisierungsverfahren versetzen Fachexperten und Datenanalysten in die Lage, gemeinsam Erkenntnisse aus den zugrundeliegenden Daten zu gewinnen und zu diskutieren. Unsere Ergebnisse und Erkenntnisse zeigen, dass sowohl unser Prozess als auch unser Visualisierungsverfahren zu interaktiver visueller Analyse, Storytelling und fundierter Entscheidungsunterstützung beitragen.

Schlagworte. Risikovorsorge, Stress Tests, Visuelle Analyse.

Management Summary

In this cumulative dissertation, research questions arising from real-world problems in different domains are presented and discussed. Information Systems Research (ISR) "embraces researchers and research from a wide array of fields under the umbrella of analytics [...] without compromising on the quality of research" (Gupta, 2017). Accordingly, this dissertation contributes to the three domains Accounting, Finance, and ISR. The topics described are all based on articles either published in academic journals and conference proceedings or currently under review. The overarching research motivation is to find new and innovative data-driven applications that are practical, relevant, and meaningful. Chronologically, the conducted research and related publications can be allocated to the following three topics:

I Determining Significant Increases in Credit Risk to Ensure Adequate Recognition of Loan Loss Provisions

II Applying Artificial Neural Networks in Stress Testing to Ensure Compatibility with Provided Scenarios by Regulators

III Visualizing Unusual Data in Subsets to Highlight Potential Areas for Further Investigation

For consistency and comparability, the same structure is used throughout the dissertation to summarize each topic. The chapters all start with a theoretical background of the problem, the derivation of research questions, and a summary of our main contributions. Then, the underlying methodology is described in a process-oriented manner. Following the approach of Rosemann and Vessey (2008), applicability checks are conducted to demonstrate the practical relevance of our research. Based on the results and findings, the main implications, recommendations, and limitations are discussed. Each chapter concludes with an overview of further research opportunities.

The following is a brief summary of our addressed research questions, proposed solutions, main contributions, and related publications.

I. Determining significant increases in credit risk

»Too little, too late« is the popular term used to summarize the weakness of existing accounting standards with regard to the recognition of credit losses of loans and other financial instruments in the time of the 2008 global financial crisis, see Brixner, Schaber, and Bosse (2013). To counteract the weakness, a new forward-looking expected credit loss (ECL) model was developed and published in the International Financial Reporting Standards (IFRS) 9, set effective in 2018, see EY Global CRS (2018). Following, the amount of ECL recognized as loan loss provision for a financial instrument depends on whether there is a significant increase in credit risk (SICR) since its initial recognition. If there is SICR, an amount equal to lifetime ECL must be recognized as loan loss provision. As long as there is no SICR or default, an amount equal to 12-month ECLs, i.e. only a portion of lifetime ECLs, must be recognized. The determination of SICR is one of the key elements of the IFRS 9 ECL model. To determine SICR, entities, i.e. banks, shall compare the lifetime default risk at the reporting date with the lifetime default risk at the date of initial recognition. Because ratings or credit scores are much more common measures in practice than lifetime default risk, the use of rating or credit score changes rather than lifetime default risk changes for the determination of SICR is easier to communicate and more transparent for stakeholders of financial statements. We analyze the methodological differences between changes of ratings and lifetime default risk for determining SICR (Bosse, Stege, & Hita Hochgesand, 2017a) and develop a proof for the use of rating changes for SICR in accordance with the impairment regulations in IFRS 9 (Bosse, Stege, & Hita Hochgesand, 2017b). We address the following research questions:

RQ1 Under what conditions can ratings be used for the determination of SICR in the IFRS 9 ECL model?

RQ2 How to demonstrate that rating changes are a reasonable approximation of lifetime default risk changes for the determination of SICR in the IFRS 9 ECL model?

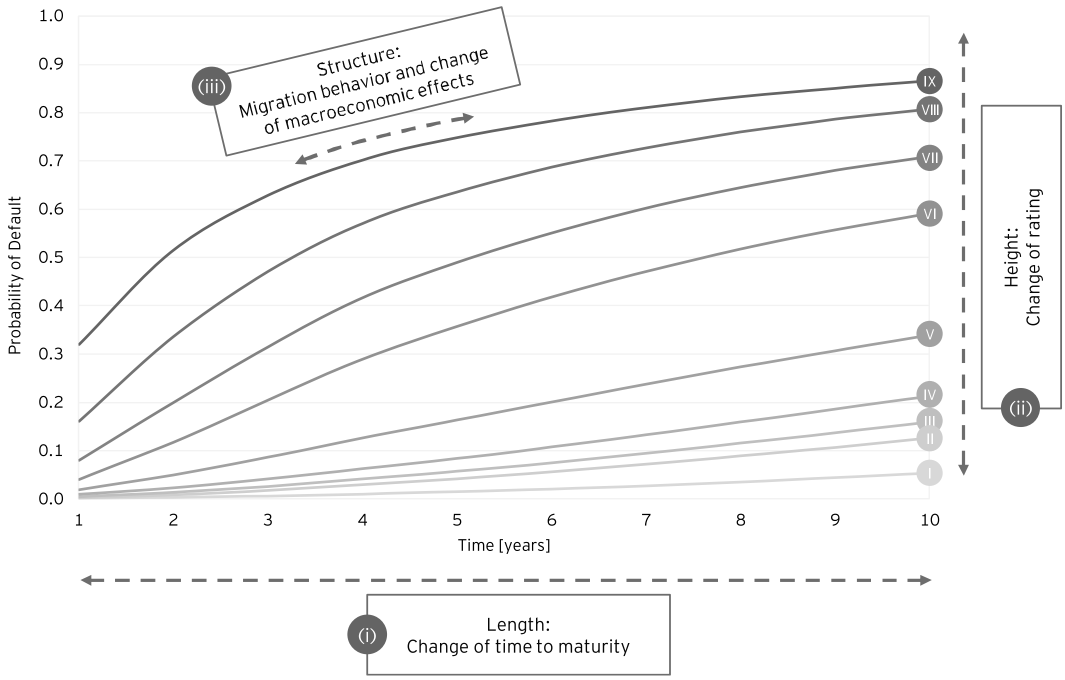

Figure 1: Factors of Lifetime Default Risk Changes (adaption of Bosse et al., 2017b, p. 439).

To address RQ1, we identify three factors that affect lifetime default risk changes. Figure 1 illustrates the three factors and their effect on lifetime default risk changes. The horizontal axis describes time in years, the vertical axis describes default risk, i.e. the probability of default (PD). If it can be shown that rating changes are the key factor for lifetime default risk changes, then rating changes can be used to determine SICR in accordance with the impairment requirements in IFRS 9.

To address RQ2, we develop a test that can be used to continuously monitor the adequacy of rating changes for SICR determination. The test is based on the interdependency between ratings and corresponding PD values that is usually described by »master scales«, a typical tool in credit risk modeling. We use the PD ranges from master scales to derive PD corridors that describe the PD ranges over time. This allows to evaluate the impact of factor changes to lifetime default risk changes. We find that leaving a PD corridor is equivalent to a rating change. Following, rating changes are the key factor for lifetime default risk changes, as long as changes caused by other factors stay within the corresponding corridor.

To demonstrate how our test can be applied in practice, we perform an applicability check and describe all necessary steps and data for replication. Our results and findings contribute to more transparency with regard to decision-relevant information for stakeholders of financial statements. An important limitation is that the application of our test is only meaningful in combination with a general validity of ratings, i.e. subject to the quality of information captured by ratings.

II. Applying artificial neural networks in stress testing

In risk management, stress testing is one of the most important tools. It describes a special form of scenario analysis to estimate losses in certain environments and has evolved considerably since the global financial crisis, see Aragonés, Blanco, and Dowd (2001) and Kohn and Liang (2019). In 2016, the European Banking Authority and European Central Bank required all major banks in the European Union to participate in a stress test specifically designed with a baseline and an adverse scenario, see European Systemic Risk Board (2016). The regulators provided official scenario estimates for a number of relevant macroeconomic factors. Estimates for other relevant factors not provided had to be generated in a way that ensured compatibility with the official scenario estimates. We demonstrate how innovative techniques from machine learning and time series analysis can be employed to map macroeconomic variables provided by regulators to relevant variables that were not provided, see Stege, Wegener, Basse, and Kunze (2021). We address the following research question:

RQ3 How can Artificial Neural Networks (ANNs) be used to produce adequate projections of macroeconomic factors consistent with regulatory guidelines on stress testing?

To address RQ3 we develop a step-by-step process with the objective to model the relationship between provided scenario variables and relevant variables that were not provided by regulators. For this, we use a basic ANN

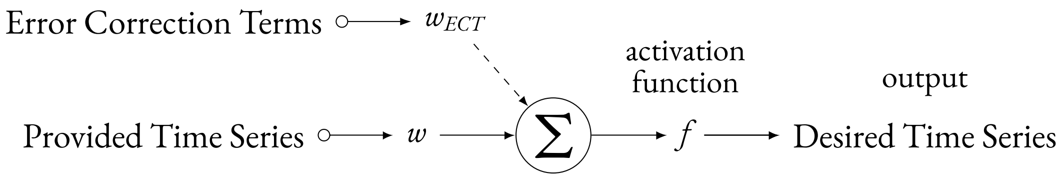

Figure 2: ANN Design of Model Types A and B (adaption of Stege et al., 2021, p.316).

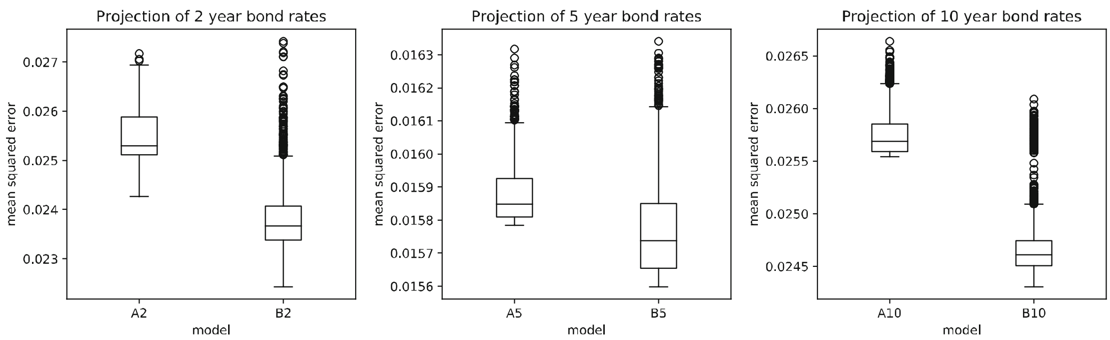

Figure 3: Comparison of Model Accuracy (Stege et al., 2021, p. 318).

»Model A« with the provided scenario variable as input and the desired variable as output. We design a second ANN »Model B« that uses cointegration information, i.e. error correction terms (ECTs), to account for the long-run relationship between both time series, see Figure 2. We then show how to generally configure the ANNs and describe the normalization of inputs, the number of neurons, and our fitting process. In an applicability check, we fit 12000 models in total. We use the mean squared error to evaluate model results. Figure 3 shows our results of both model types A and B for different maturities of the desired time series. Our results show that model B outperforms model A for all maturities. For empirical evidence, we apply non-parametric median tests to the model pairs to show that the medians of model errors differ systematically.

Our results and findings are relevant for risk managers in the financial services industry. From a different perspective, our approach can also be applied by auditors and regulators of financial institutions to validate the execution of stress tests and to ensure compliance. There is great potential for further research, because in its current state the complex configuration requirements most likely limit general applicability and acceptance.

III. Visualizing unusual data in subsets

Visualization methods in an analytics context, i.e. visual analytics, has been an active research field with applications in many sectors, see Sun, Wu, Liang, and Liu (2013). A powerful tool in visual analytics is storytelling, which describes the investigative process of connecting dots between seemingly disconnected information, see Hossain, Andrews, Ramakrishnan, and North (2011). In auditing, for example, accountants and auditors are increasingly using visual analytics and storytelling to analyze and communicate results, see, e.g., Sekar (2022). For complex analytical tasks, auditors are often assisted by specialists who have the required technical qualifications for these tasks. Visual analytics and the competency in storytelling can improve collaboration of auditors and specialists to jointly increase the overall quality of an audit. We develop a visualization framework to identify and assess unusual items in financial data for further investigation, see Stege and Breitner (2020). Because our framework is not limited to auditing, we generalize it for application in many domains where structured data is available, see Stege and Breitner (2023). We address the following research questions:

RQ4 How can visual analytics be embedded into a decision-making process to improve collaboration between data analysts and domain experts?

RQ5 How can the visualization framework be efficiently applied in practice to support anomaly explanation, well-founded decisions, and storytelling?

To address RQ4, we deduce a general process model and embed a visualization framework for the visual analysis of subset-dataset relationships. The process model consists of four main phases that are run through sequentially and cyclically until a well-founded decision is reached. The process is initiated with a problem statement and an underlying dataset that are run through the four phases. Accordingly, in a nutshell, subsets are identified from the dataset which are then visualized and jointly discussed by domain experts and data analysts to utilize insights and draw conclusions for decision support, see Figure 4.

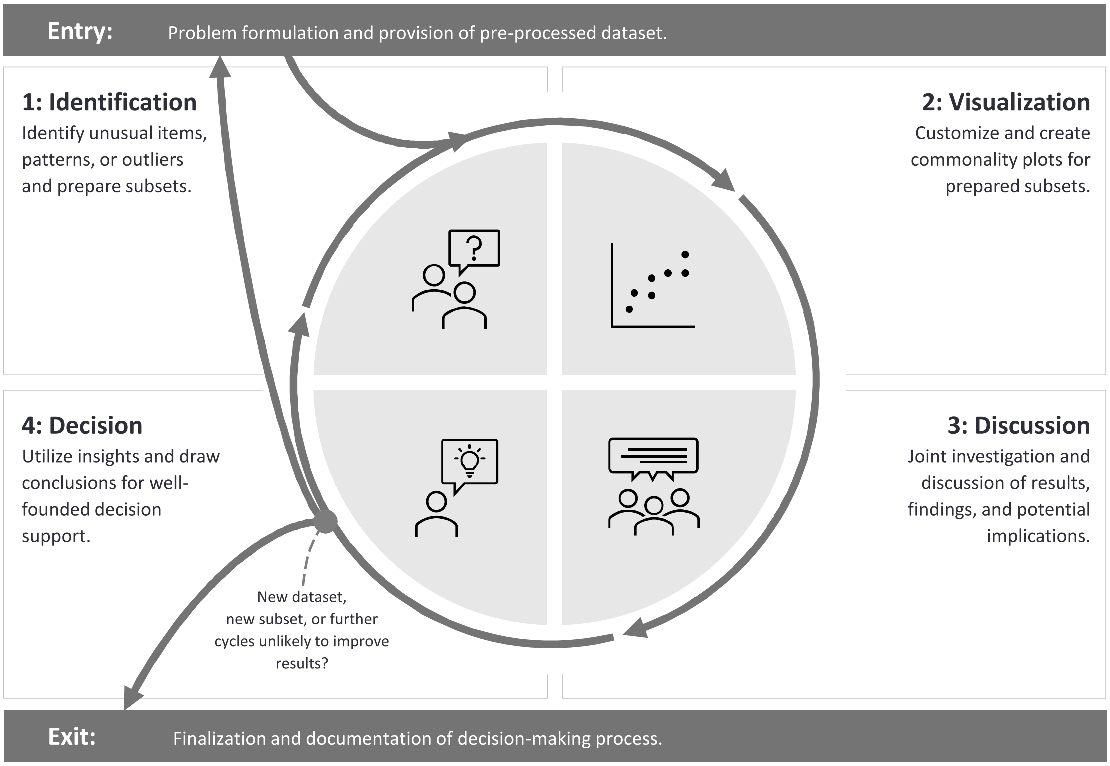

Figure 4: Interactive Process Model for Decision Support (Stege & Breitner, 2023, p. 6).

For the visualization phase, we introduce »commonality plots«, a novel visualization framework specifically designed to visualize commonalities in subsets and highlight items that have a low likelihood of occurrence. We create commonality plots for each feature in a dataset. Our visualization framework is inspired by Zoomable Circle Packing by Bostock (2018), which is closely related to treemaps, but better reveals the underlying hierarchy of data. This is very useful for the visualization of subset-dataset relationships. Figure 5 shows the commonality plots for three features of an arbitrary subset of the well-known Titanic dataset, which is publicly available and contains information about all passengers aboard the famous ship, see Kaggle (2012). The light gray circles show all values of the respective feature. The size of the circles indicates the number of occurrence in the main dataset. The dark gray circles show the values of the subset. We say there is commonality, if a feature of the subset only consists of equal values. Accordingly, the subset for feature »Age« has commonality, because there are only 23-year-old passen-

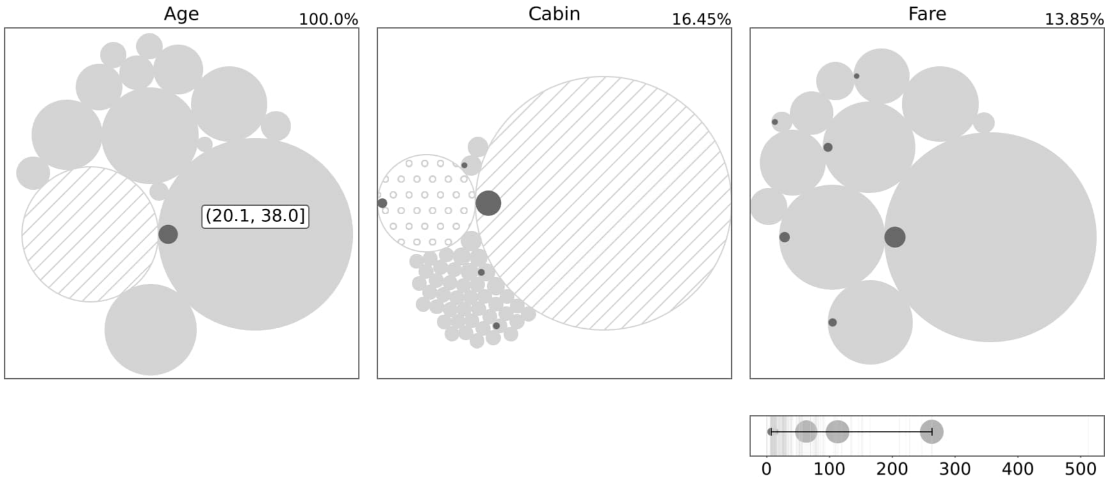

Figure 5: Commonality Plots for Selected Data Fields of the Titanic Dataset.

gers in the subset. The commonality plot for feature »Fare« shows the ticket prices passengers paid to get aboard. Because the feature is numeric, we show the distribution as a one-dimensional rug plot. The line plotted on top of the distribution shows the value range of ticket prices in the subset, i.e. the distribution of ticket prices paid by 23-year-old passengers. The feature »Cabin« describes the cabin labels of the passengers. A striped hatching indicates missing values, a dotted hatching indicates individual values that only appear once in the dataset. Apparently, the cabin label information is missing for the majority of passengers. Also lots of passengers have booked single cabins. Finally, we measure the unlikeliness of occurrence of the observed items in a subset. The results are presented on the top right corner of each visualization.

To address RQ5, we perform an applicability check to demonstrate how our process model and embedded visualization framework are applied in practice for well-founded decision support and storytelling. We show how data analysts and auditors jointly apply our process to analyze risk data from a European bank. Our process is most efficient for the analysis of structured datasets, because all measures necessary for the visualizations need to be calculated only once, regardless of the number of subsets. Our process is least efficient for the analysis of time series data that is constantly updated and where chronological order is of importance.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ADF	Augmented Dickey-Fuller test
AEG	Augmented Engle-Granger test
ANN	Artificial Neural Network
AOR	Annals of Operations Research
BI&A	Business Intelligence & Analytics
CBY	covered bond yields
DKE	Data & Knowledge Engineering
EBA	European Banking Authority
ECB	European Central Bank
ECL	expected credit loss
ECT	error correction term
ESRB	European Systemic Risk Board
ETL	extract, transform, and load
EU	European Union
EUR	Euro
IASB	International Accounting Standards Board
IFRS	International Financial Reporting Standards
IML	International Conference on Internet of Things and Machine Learning
IS	Information Systems
ISR	Information Systems Research
IT	Information Technology
ITG	IFRS Transition Resource Group for Impairment of Financial Instruments
PD	probability of default
PP	Phillips-Perron test

SICR significant increase in credit risk
SR swap rates
STAC Statistical Tests for Algorithms Comparison
VECM Vector Error Correction Model
VHB Verband der Hochschullehrer für Betriebswirtschaft
WI International Conference on Wirtschaftsinformatik
WPg Die Wirtschaftsprüfung (Fachzeitschrift)

Overview of Publications and Task Allocation

This dissertation is based on six papers written in collaboration with different co-authors. Five papers were peer-reviewed, accepted, and published in different journals or conference proceedings. One paper is submitted and still under review at the time of writing this dissertation. An overview of all papers in chronological order is presented in Table 1 also showing the JOURQUAL3 2015 ranking from the Verband der Hochschullehrer für Betriebswirtschaft (VHB). The following paragraphs briefly describe the contents of each paper as well as the allocation of tasks between authors.

In "Beurteilung der signifikanten Verschlechterung der Kreditqualität nach IFRS 9: Voraussetzungen für die Verwendung von Ratings und Lifetime-PD" (Bosse, Stege, & Hita Hochgesand, 2017a) we analyze methodological differences between ratings and the PD over the lifetime of a financial instrument for detecting a significant increase in credit risk (SICR) in the context of the impairment requirements set out in the International Financial Reporting Standards (IFRS). Dr. Michael Bosse had the idea to publish parts of the theoretical research we performed as quantitative specialists of a large accounting firm during the implementation of IFRS 9 Impairment requirements at a German bank. All parts of the paper were developed and written jointly by Dr. Michael Bosse and myself. Our colleague Dr. Manuel Hita Hochgesand supported us with the general conception. The paper was accepted and published in Die Wirtschaftsprüfung (WPg), Volume 70, Issue 1. WPg is a journal primarily aimed at auditors, tax consultants, as well as specialists and executives in companies from the areas of corporate governance, finance, and controlling.[1]

In "Stufenzuordnung nach IFRS 9: Nachweis zur Verwendung von Ratings als geeignetes Kriterium zur Beurteilung der signifikanten Verschlechterung der Kreditqualität" (Bosse, Stege, & Hita Hochgesand, 2017b) we develop a proof for the use of rating changes as an adequate SICR criterion in the context of IFRS 9 Impairment requirements. The paper is directly re-

[1] See https://www.idw.de/idw-verlag/wpg/, accessed April 2023.

lated to the previously described paper, as it builds upon our designed framework. Again, Dr. Michael Bosse had the idea to publish parts of our theoretical research on the matter, and we wrote all parts of the paper together, while Dr. Manuel Hita Hochgesand supported us in the conception phase. The paper was accepted and published in WPg, Volume 70, Issue 8.

In ”Mapping interest rate projections using neural networks under cointegration: An application from stress testing approaches” (Stege, Wegener, Basse, & Kunze, 2017) we discuss the application of techniques of business analytics in the banking industry, namely the use of neural networks in the context of financial risk management. The paper is a collaboration of four authors. Prof. Dr. Christoph Wegener had the idea to use cointegration analysis information as an input to neural networks to apply it for stress testing exercises. I was responsible for the design and application of the neural networks and also performed the statistical evaluation of results together with Prof. Dr. Christoph Wegener. Prof. Dr. Tobias Basse was responsible for literature research and theoretical background, while Dr. Frederik Kunze provided data and insights from practice. In October 2017, I presented our results at the International Conference on Internet of Things and Machine Learning (IML) in Liverpool, UK.

The paper ”Mapping swap rate projections on bond yields considering cointegration: An example for the use of neural networks in stress testing exercises” (Stege et al., 2021) is directly related to the previously described paper and further discusses the application of techniques of business analytics in the banking industry. The paper is a collaboration of the same four authors and builds upon our findings from previous research on the matter and conference feedback. Again, I was responsible for the design and application of the neural networks and also performed the statistical evaluation of results together with Prof. Dr. Christoph Wegener, who also refined the underlying methodology of our approach. Prof. Dr. Tobias Basse was responsible for literature research and theoretical background. We used the same data as in our previous paper which was provided by Dr. Frederik Kunze along with insights from practice. The paper was published in the special issue ”Neural Networks, Nonlinear Dynamics, and Risk Management in Banking and Finance" in the Annals of Operations Research (AOR), Volume 297, Issue 1. The paper was accepted after a double-blind peer-review process with two reviewers and two revisions. The AOR publishes full-length research articles dealing with key aspects of operations research, reports on computational studies, and case studies that present new and innovative practical applications.[2]

In "Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists" (Stege & Breitner, 2020) we introduce a visualization method that enables domain experts and data scientists to jointly discuss results from pattern recognition analyses. The paper is motivated by a typical problem from auditing I encountered while working as a quantitative specialist for a large accounting firm. I developed the visualization method, performed the application example, and was responsible for the entire paper. Prof. Dr. Michael H. Breitner supervised the writing process and thoroughly reviewed and edited all sections. The paper was accepted for the International Conference on Wirtschaftsinformatik (WI) 2020 and was published in the conference proceedings after one double-blind peer review round. In March 2020, I presented our results at the WI in Potsdam, Germany.

In "Identify, visualize, discuss, and decide: A collaborative framework to explore unusual subset-dataset relationships" (Stege & Breitner, 2023) we derive and evaluate an efficient interactive process model and visualization framework to enable value creating collaboration in interdisciplinary teams. The paper is partly based on the previously described conference paper and is advanced by diverse and extensive improvements regarding contribution (analytic process and visualization framework), generalizability, and applicability. I developed and enhanced the visualization framework, designed and performed the application example, and was responsible for the entire paper. Prof. Dr. Michael H. Breitner supported me with the conception of the generalized process model that embeds our visualization framework. Additionally, he supervised the writing process and thoroughly reviewed and edited all sections. The paper was submitted to the Data & Knowledge Engineer-

[2] See https://www.springer.com/journal/10479, accessed April 2023.

ing (DKE) journal in April 2023 and is still under review at the time of writing this dissertation. DKE publishes full-length research articles dealing with technical advances concerning data engineering, knowledge engineering, and interdependencies between these fields.[3]

[3] See https://www.sciencedirect.com/journal/data-and-knowledge-engineering/, accessed April 2023.

Year	Title	Authors	Conference/Journal	VHB/JQ3[a]	Chapter	Appendix
2017	Beurteilung der signifikanten Verschlechterung der Kreditqualität nach IFRS 9: Voraussetzungen für die Verwendung von Ratings und Lifetime-PD	Bosse, M. **Stege, N.** Hita Hochgesand, M.	Die Wirtschaftsprüfung (WPg)	C	2	A.1
2017	Stufenzuordnung nach IFRS 9: Nachweis zur Verwendung von Ratings als geeignetes Beurteilungskriterium der signifikanten Verschlechterung der Kreditqualität	Bosse, M. **Stege, N.** Hita Hochgesand, M.	Die Wirtschaftsprüfung (WPg)	C	2	A.2
2017	Mapping interest rate projections using neural networks under cointegration: An application from stress testing approaches	**Stege, N.** Wegener, C. Basse, T. Kunze, F.	Proceedings of the International Conference on Internet of Things and Machine Learning (IML)	-	3	A.3
2020	Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists	**Stege, N.** Breitner, M.H.	Proceedings of the Internationale Tagung für Wirtschaftsinformatik (WI)	C	4	A.4
2021	Mapping swap rate projections on bond yields considering cointegration: An example for the use of neural networks in stress testing exercises	**Stege, N.** Wegener, C. Basse, T. Kunze, F.	Annals of Operations Research (AOR)	B	3	A.5
2023[b]	Identify, visualize, discuss, and decide: A collaborative framework to explore unusual subset-dataset relationships	**Stege, N.** Breitner, M.H.	Data & Knowledge Engineering (DKE)	B	4	A.6

Table 1: Overview of Publications

[a] JOURQUAL3 Verband der Hochschullehrer für Betriebswirtschaft (VHB 2015)
[b] In review process. Submitted in April 2023.

Our research has been alternately praised and criticized for being too cross-disciplinary, but we believe this is strength and not a weakness in today's data rich environment.

Ritu Agarwal & Vasant Dhar

1 Introduction

1.1 Research motivation and relevance

The introductory quote of Agarwal and Dhar (2014) on Information Systems Research (ISR) also applies to this cumulative dissertation. The papers presented and discussed are all motivated by real-world problems identified in different domains. The overarching research motivation is to find new and innovative data-driven applications that are practical, relevant, and meaningful. The versatile focus areas of the earlier research phase significantly sparked motivation and ideas for the final phase of research. Chronologically, allocating the papers to their respective domains, the course of research starts in Accounting, continues in Finance, and concludes in ISR.

1.1.1 Research in Accounting

Our research in Accounting focuses on regulatory requirements for the recognition of expected credit losses under the International Financial Reporting Standards (IFRS) 9 »Impairment of Financial Instruments«. During the financial crisis of 2008 the delayed recognition of credit losses on financial instruments was identified as a weakness in existing accounting standards ("too little, too late"), see EY Global CRS (2018) and Brixner et al. (2013), so the International Accounting Standards Board (IASB) issued IFRS 9 Impairment that came into effect in 2018. Following, a new impairment model that requires more timely recognition of expected credit losses has been introduced as part of IFRS 9. Because adopters are allowed to apply self-developed credit risk models as a basis to determine expected credit losses, it becomes necessary to analyze the existing models to evaluate IFRS 9 compliance and identify potential needs for adjustment. A central challenge for adopters is the definition and setting of criteria that timely indicate a significant increase in credit risk (SICR) that requires the recognition of lifetime expected credit losses as loan loss provisions. We develop a proof for the use of rating changes as an adequate SICR determination criterion in the context of IFRS 9 Impairment requirements. For this, we define a framework under which conditions ratings can be generally used as a SICR determination criterion (Bosse, Stege, & Hita Hochgesand, 2017a). Based on the framework, we develop a proof for the use of ratings as a SICR determination criterion that also includes a validation test for adequacy (Bosse, Stege, & Hita Hochgesand, 2017b).

1.1.2 Research in Finance

Our research in Finance focuses on the application of techniques of data analytics in the banking industry, i.e. modeling neural networks in the context of stress testing. As one of the most important tools in applied financial risk management, stress testing is a special form of scenario analysis to estimate losses in certain macroeconomic environments, see, e.g., Hirtle, Kovner, Vickery, and Bhanot (2016). Following, stress tests have been used by supervisors and central banks to assess the overall capital adequacy of the banking system. The European Banking Authority (EBA) and the European Central Bank (ECB), for example, together designed and carried out their 2016 stress testing exercises. All major banks in the European Union (EU) were required to participate in this stress test. For the 2016 EU-wide stress test, the European Systemic Risk Board (ESRB) defined a baseline scenario and an adverse scenario. For both scenarios, estimates of the development of some relevant macroeconomic variables were provided (European Systemic Risk Board, 2016). Estimates for other relevant variables not provided by the ESRB had to be generated in a way that ensured compatibility with the forecasts of the official scenarios of the ESRB. Many banks used techniques of time series analysis and other quantitative modeling strategies to produce the required additional forecasts. We adapt a multi-methodology approach by Wegener, von Spreckelsen, Basse, and von Mettenheim (2016) who combine neural networks with techniques of cointegration analysis. The objective of our approach is to enable practitioners to map the macroeconomic variables provided by the regulators to covered bond yields that are an important source of funding for many European banks (Stege, Wegener, Basse, & Kunze, 2021).

1.1.3 Information Systems Research

In our final research phase, we derive and evaluate an efficient interactive process model driven by data analytics and domain knowledge to enable value creating collaboration in interdisciplinary teams. This is relevant, because the integration of data analytics in business processes and workflows for data-driven decision support has been an important challenge for companies and organizations (Chen, Chiang, & Storey, 2012). A powerful set of algorithms, methods, and tools is available for this, but insights do not emerge easily and not automatically from application (Sharma, Mithas, & Kankanhalli, 2014). In short, there is a great need for humans to interact efficiently with data (Saghafi, Wand, & Parsons, 2022). Following Alpar and Schulz (2016) and Eilers, Köpp, Gleue, and Breitner (2017), the increasing complexity of analytical business tasks requires large amounts of data that can only be managed by highly specialized teams of data analysts. While data analysts have the essential technical skills, in the working world the most valuable domain knowledge resides with managers and decision-makers with years of experience. It is the domain experts who face business needs, and the data analysts who apply their technical skills, but often lack domain knowledge to explain certain phenomena in the underlying data. This is a gap that can result in sub-optimal decision-making (Eilers et al., 2017; Hagen, 2021). Thus, improving communication and collaboration between domain experts and data analysts is of great importance for well-founded decision support, making it a highly relevant field for ISR. Motivated by a typical problem from auditing, we introduce a visualization framework that helps to detect unusual data in a subset and highlights potential areas for investigation (Stege & Breitner, 2020). Finally, we enhance our visualization framework and embed it in a generalized process model applicable in many domains where structured data is available (Stege & Breitner, 2023).

1.2 Structure of the dissertation

The dissertation is divided by research domain. Chapter 2 presents and discusses our publications in the Accounting practice, Chapter 3 covers our research in Finance, and Chapter 4 presents and discusses our publications in ISR. For consistency and comparability, the same structure is used to present our research. Each chapter starts with a summary of the theoretical background, the derivation of our research questions, and a summary of our main contributions. Then, the underlying methodology is described in a process-oriented manner. Following the approach of Rosemann and Vessey (2008), applicability checks are conducted to demonstrate the practical relevance of our research. Based on our results and findings, the main implications, recommendations, and limitations are discussed, followed by an overview of further research opportunities. Chapter 5 concludes with a reflective discussion of the overall course of research.

The main parts of the dissertation are preceded by sections containing abstract, management summary, and an overview of publications and task allocation.

Like forest fires, unanticipated crises can also be regenerative — revealing gaps in our thinking, they can shake loose deeply held assumptions.

Stephen C. Nelson & Peter J. Katzenstein

2 Determining Significant Increases in Credit Risk

2.1 Theoretical background and contribution

In the time of the 2008 global financial crisis, the impairment requirements of the existing accounting standards did not recognize credit losses until a credit loss event occurred. This was identified as a weakness, because credit losses of loans and other financial instruments were recognized too little, too late, see EY Global CRS (2018) and Brixner et al. (2013). The IASB addressed these concerns and developed a new forward-looking expected credit loss (ECL) model as set out in IFRS 9. Following, the amount of ECLs recognized as loss allowance for a financial instrument depends on the extent of deterioration of the underlying credit quality, i.e. increase in credit risk, since initial recognition.

The IFRS 9 ECL model for the recognition of loss allowances is divided

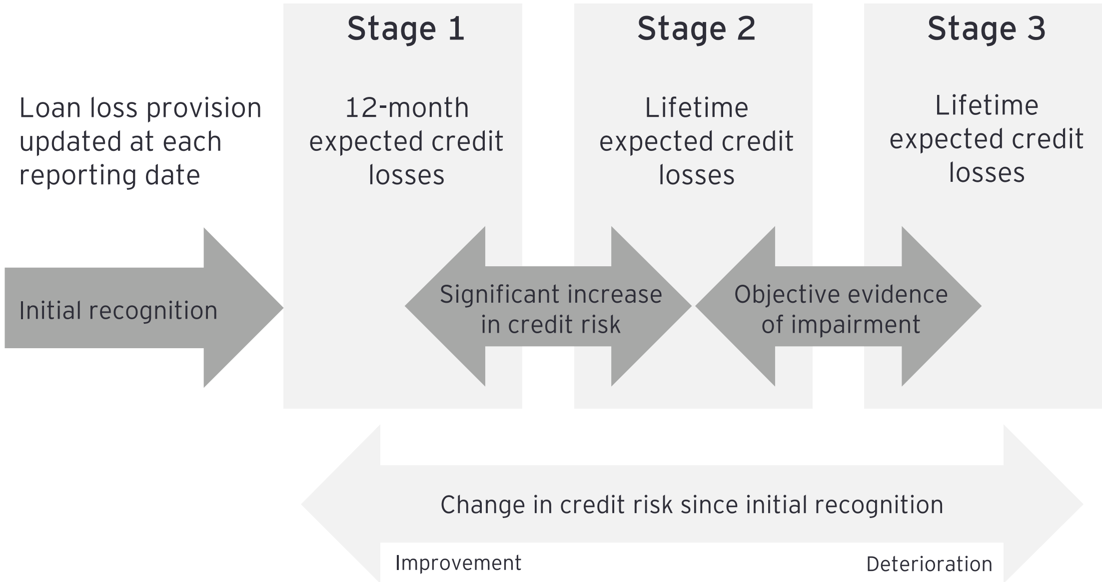

Figure 2.1: General Approach of the IFRS 9 ECL Model (adaption of EY Global CRS, 2018, p. 9).

into three stages. According to IFRS 9.5.5.3 and 5.5.5, as long as there is no SICR or default, an amount equal to 12-month ECLs must be recognized that represent credit losses from default events that are possible within 12 months after the reporting date (stage 1 of the IFRS 9 ECL model). Otherwise, the amount must be equal to lifetime ECLs, which are losses that result from all possible default events over the remaining lifetime of a financial instrument (stages 2 and 3). The assumption is that there will be SICR before a default occurs, so prior to an allocation to stage 3, financial instruments should generally have been allocated to stage 2 (IFRS 9.B5.5.7). This corresponds to the overall objective to ensure a more timely recognition of credit losses. Figure 2.1 summarizes the general approach of the ECL model for the recognition of either 12-month or lifetime ECLs.

Against this background, the determination of SICR for correct stage allocation of financial instruments is one of the key elements of the IFRS 9 ECL model. In determining whether there has been a SICR, "an entity shall use the change in the risk of a default occurring over the expected life of the financial instrument" (IFRS 9.5.5.9). In other words, entities shall compare the lifetime default risk at the reporting date with the lifetime default risk at the date of initial recognition of the financial instrument. Despite this, in certain circumstances, changes of the 12-month default risk may be a reasonable approximation of lifetime risk changes. ”In such cases, an entity may use changes in the risk of a default occurring over the next 12 months to determine whether credit risk has increased significantly since initial recognition” (IFRS 9.B5.5.13). The risk of a default occurring over the next 12 months can be defined as the 12-month probability of default (PD) often presented by ratings or credit scores, which are well-known measures of the credit quality of transactions or debtors. On the other hand, lifetime default risk, i.e. lifetime PD, is a new and uncommon measure, at least in an accounting context. Thus, a resulting stage allocation from 12-month PD and corresponding rating changes is easier to communicate and more transparent to users of financial statements than stage allocation results based on changes of lifetime PDs.

We analyze the methodological differences between changes of ratings and lifetime PDs for determining SICR (Bosse et al., 2017a) and develop a proof for the use of rating changes as an adequate stage allocation criterion in the IFRS 9 ECL model (Bosse et al., 2017b). We address the following research questions:

RQ1 Under what conditions can ratings be used as a determination criterion for SICR in the IFRS 9 ECL model?

RQ2 How to demonstrate that rating changes are a reasonable approximation of lifetime PD changes for determining SICR in the IFRS 9 ECL model?

The implementation of complex IFRS 9 requirements such as the ECL model for recognizing credit losses is challenging. Even more challenging is an implementation that is transparent and conveys decision-relevant information for stakeholders of financial statements. Our findings and implications contribute to transparency with regard to the recognition and disclosure of loan loss provisions in financial statements.

2.2 Methodology

2.2.1 Conditions for determining SICR with ratings

In principle, IFRS 9 regards 12-month PD changes as reasonable approximation for lifetime PD changes when determining SICR, "unless circumstances indicate that a lifetime assessment is necessary" (IFRS 9.B5.5.13). This is the case when there are significant payment obligations beyond 12 months or changes in relevant factors occur that are not adequately reflected in the 12-month PD or only have an impact beyond 12 months (IFRS 9.B5.5.14). A possible approach to ensure that these circumstances do not exist can be designed as suggested by the IFRS Transition Resource Group for Impairment of Financial Instruments (ITG), see ITG (2015b):

1. identification of key factors that affect the appropriateness of 12-month PD changes as reasonable approximation of lifetime PD changes;

2. monitoring of identified factors and consideration whether potentially observed changes of identified factors affect the appropriateness of 12-month PD changes as reasonable approximation of lifetime PD changes.

Consequently, it needs to be demonstrated that changes of 12-month PDs and corresponding rating changes account for the key factors that affect lifetime PD changes.

2.2.2 Factors determining lifetime PD changes

A common approach to calculate lifetime PDs is the multiplication of rating migration matrices that contain the probabilities for all possible combinations of rating changes within a given time interval. Typically, migration matrices describe a time interval of 12 months, so the migration probability of a specific rating into default corresponds to its 12-month PD.

As Figure 2.2 shows, the multiplication of migration matrices results in specific PD curves that describe the lifetime PD over time for each credit rating. The vertical axis of the plot shows probabilities, the horizontal axis shows time in years, and there are nine ratings »I« to »IX«. While the height of a PD

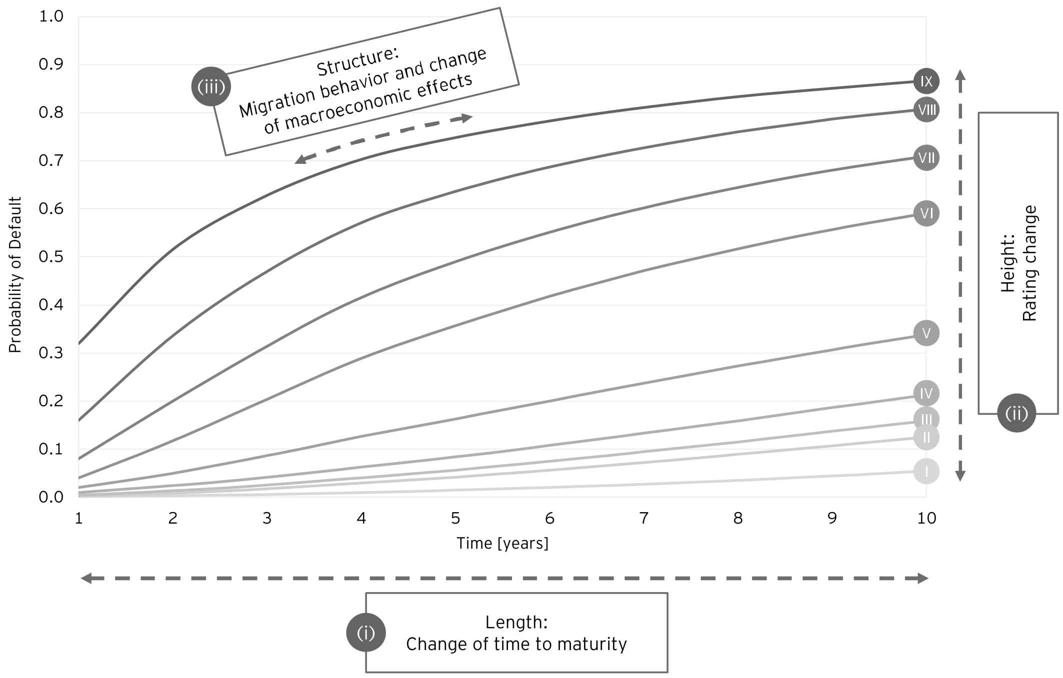

Figure 2.2: Factors Determining Lifetime PD Changes (adaption of Bosse et al., 2017b, p. 439).

curve depends on the corresponding rating, its length is determined by time. The general structure of a PD curve depends on the migration probabilities of the corresponding rating and relevant factors that affect the probabilities over time. With respect to the determination of changes in lifetime PD, the following factors can be derived:

(i) Change of time to maturity

(ii) Rating changes

(iii) Migration behavior and change of macroeconomic effects

If we can show that rating changes are the main factor for determining lifetime PD changes and the other factors have no significant impact, then rating changes represent a reasonable approximation for lifetime PD changes in the context of determining SICR.

2.2.3 Rating changes as main factor for lifetime PD changes

The longer the time to maturity of a financial instrument, the higher the probability that the debtor will not be able to meet the payment obligations arising from the financial instrument. Vice versa, "the risk of a default occurring over the expected life usually decreases as time passes if the credit risk is unchanged and the financial instrument is closer to maturity" (IFRS 9.B5.5.11). The change in lifetime PD that is triggered solely by a change of time to maturity, i.e. factor (i) in Figure 2.2, must be neutralized when determining SICR. The impact of factor (i) can be derived by moving along a PD curve and measuring the difference between two points in time. In this context, neutralization simply means that the comparison of original and current lifetime PD must be based on the same term length.

Credit ratings represent the judgment of financial analysts who use quantitative modeling and qualitative adjustments to the results of their quantitative models to produce ratings (Bozanic, Kraft, & Tillet, 2022). Although ratings typically correspond to 12-month PDs, risks beyond 12 months are taken into account when generating a rating, e.g. in the form of estimates of future financial conditions of industries and companies. Thus, rating changes implicitly address changes in the risk of a default occurring over the expected life of a financial instrument, see factor (ii) in Figure 2.2. The impact of factor (ii) equals the difference between PD curves at a specific point in time.

While it is straightforward to derive the impacts of factors (i) and (ii), i.e. movement along and between curves, the analysis of the third factor is more complex, because changes in macroeconomic effects and migration behavior change the structure of PD curves, see factor (iii) in Figure 2.2. To measure the impact of factor (iii), it is necessary to analyze how changes of macroeconomic effects alter a PD curve. If information is available that is not accounted for by ratings used for the lifetime PD calculation, it is possible to iteratively adjust the resulting lifetime PD curves as follows:

$$\begin{aligned}
PD_t^{conditional} &= \frac{PD_t^{cumulative} - PD_{t-1}^{cumulative}}{1 - PD_{t-1}^{cumulative}}, \\
PD_t^{\Delta conditional} &= PD_t^{conditional} \cdot (1 + \Delta_t), \\
PD_t^{\Delta cumulative} &= PD_t^{\Delta conditional} \cdot (1 - PD_{t-1}^{\Delta cumulative}) + PD_{t-1}^{\Delta cumulative},
\end{aligned} \tag{2.1}$$

where $PD_t^{conditional}$ is the conditional PD at time t, Δ_t is the scaling factor of macroeconomic change, and $PD_t^{cumulative}$ are the original lifetime PDs with $PD_0^{cumulative} = 0$. Then, the adjusted lifetime PD curves $PD_t^{\Delta cumulative}$ are iteratively calculated based on the adjusted conditional PDs $PD_t^{\Delta conditional}$. For a detailed description of conditional and cumulative PDs and their interdependencies, see Bosse et al. (2017b).

Structural changes to lifetime PD curves due to macroeconomic effects, as calculated in (2.1), could mean that the change in lifetime PD is no longer properly approximated by a rating change, because the adjusted lifetime PD no longer corresponds to the original rating. The extent to which the change in macroeconomic impact can be considered insignificant must be determined to test whether rating changes are still a reasonable approximation for lifetime PD changes.

2.2.4 Test for adequacy of rating changes as SICR criterion

The interdependency between ratings and PDs is the basis for deriving a test for adequacy of rating changes as reasonable approximation for lifetime PD changes. Following Bluhm, Overbeck, and Wagner (2016), in credit risk modeling, the process of assigning PDs to ratings is called calibration. The end product of a calibration is a master scale that maps ratings to specific PD ranges or single values of the ranges, typically the mid-points. In the context of lifetime PD calculation on the basis of rating migration matrix multiplication, a rating not only describes lifetime PD curves, but also a PD corridor that corresponds to the PD range of the master scale over time. Thus, leaving a PD corridor is equivalent to a rating change. If the structural change of a PD profile implies leaving the PD corridor of a rating, the lifetime PD is no

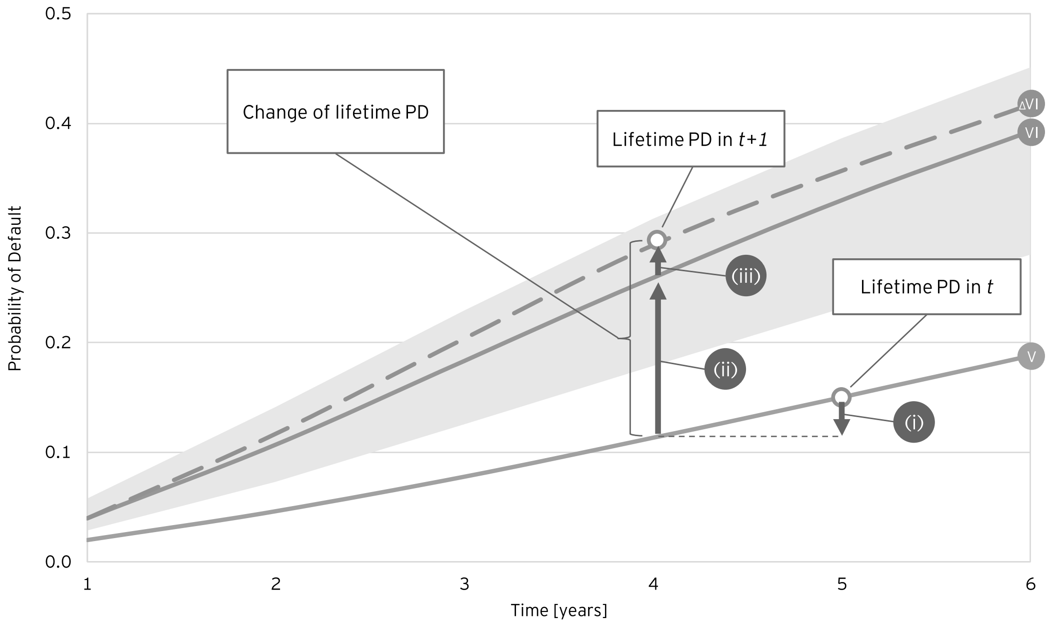

Figure 2.3: Impact of Factors of Lifetime PD Changes (adaption of Bosse et al., 2017b, p. 446).

longer represented by that rating. So PD corridors can be interpreted as an adequate range of tolerance within which a change of macroeconomic effects is irrelevant for the relationship between lifetime PD and corresponding rating.

The PD corridor can be determined by rolling forward the ratio of PD range boundaries and midpoint PDs of two ratings i and $i+1$ over time t:

$$\begin{aligned} v(i) &= \frac{PD^{lb}(i+1) - PD(i)}{PD(i+1) - PD(i)}, \\ PD^{ub}_t(i) = PD^{lb}_t(i+1) &= v(i) \cdot (PD_t(i+1) - PD_t(i)) + PD_t(i), \end{aligned} \tag{2.2}$$

where $v(i)$ is the boundary ratio for rating i used to derive the upper bound $PD^{ub}_t(i)$ of the PD corridor of rating i, which equals the lower bound $PD^{lb}_t(i+1)$ of the PD corridor of rating $i+1$. For a detailed description of PD corridor derivation, see Bosse et al. (2017b). Based on (2.2), the range of tolerance for rating i at time t equals:

$$[PD^{lb}_t(i).PD^{ub}_t(i)]. \tag{2.3}$$

In summary, rating changes are the main factor for lifetime PD changes and

Rating in t	Rating in t+1									
	I	II	III	IV	V	VI	VII	VIII	IX	Def.
I	92	4	2	1	0.5	0.25	0.1	0.02	0.00	0.13
II	3	85	5.85	2.5	1	1.25	0.5	0.35	0.30	0.25
III	1	6	82	5	2.5	1.75	0.75	0.4	0.1	0.5
IV	0.5	1.25	7	80	6	3	1	0.2	0.05	1
V	0.25	2	4	6	75	5	3	1.75	1	2
VI	0	0.25	0.75	3.5	6.5	60	10	9.25	5.75	4
VII	0	0.13	0.25	1.12	3.5	7	55	15	10	8
VIII	0	0	0.1	0.25	1.65	4	8	50	20	16
IX	0	0	0	0.25	1.25	3.5	6	12	45	32
Def.	0	0	0	0	0	0	0	0	0	100

Table 2.1: Exemplary 12-month Rating Migration Matrix (adaption of Bosse et al., 2017a, p. 9). All Values in Percent [%].

represent a reasonable approximation of lifetime PD changes as long as adjusted lifetime PDs stay within the range of tolerance described by (2.3). The derived factors of lifetime PD changes and the range of tolerance between lifetime PD curves are illustrated in Figure 2.3. In addition to a magnification of lifetime PD curves of ratings V and VI from Figure 2.2, rating VI was adjusted for macroeconomic effects resulting in lifetime PD curve described by ΔVI. The light gray area encompassing lifetime PD curve VI illustrates its range of tolerance. Figure 2.3 shows that the rating change from V to VI is the main factor for the overall lifetime PD change, and the other factors have no significant impact. If this can be shown for the other ratings as well, it is proven that rating changes are the main factor for lifetime PD changes and thus represent a reasonable approximation in the context of determining SICR.

2.3 Applicability check

Three inputs are required to perform the adequacy test of rating changes for SICR determination from the previous section: a rating migration matrix to derive lifetime PDs, the corresponding master scale to derive tolerance ranges, and estimates of macroeconomic effects that are not accounted for by the original ratings to retrospectively adjust lifetime PDs.

Rating	lower bound PD	midpoint PD	upper bound PD
I	0%	0.13%	0.2%
II	0.2%	0.25%	0.4%
III	0.4%	0.5%	0.7%
IV	0.7%	1%	1.4%
V	1.4%	2%	2.9%
VI	2.9%	4%	5.8%
VII	5.8%	8%	10.6%
VIII	10.6%	16%	19.1%
IX	19.1%	32%	100%
Def.	100%	100%	100%

Table 2.2: Exemplary Master Scale of a Rating Model (adaption of Bosse et al., 2017a, p. 8).

Table 2.1 shows an exemplary rating migration matrix in which rows describe original ratings of debtors at time t and columns describe the ratings of debtors observed at time $t+1$. The matrix' entries describe the probabilities of rating migrations between t and $t+1$. In our case, t is measuring years, so the matrix represents rating migrations within one year. There are nine ratings »I« to »IX« and the default state »Def.«. The higher the rating, the lower the associated credit quality, thus the higher the probability to migrate into default. Per definition, default is an absorbing state, so once a debtor defaults, the probability of leaving default is zero (Andersson & Vanini, 2010).

Table 2.2 shows an exemplary master scale that describes the lower and upper bounds of PD ranges for ratings I to IX. The midpoint PDs of the master scale, i.e. 12-month PDs, are equal to default values of the migration matrix in Table 2.1.

First, lifetime PDs are derived for ten years by multiplying the migration matrix in Table 2.1 ten times with itself. Figure 2.2 shows the resulting lifetime PD curves, Tables 2.3 and 2.4 show the lifetime PD values for ratings V to VII for six years. For the lifetime PD results of all ratings for ten years, see Bosse et al. (2017a). Next, we assume a recession for years 2 to 4 and estimate a 15% increase of the conditional PD values. We adjust the conditional PDs

	Time [years]					
	1	2	3	4	5	6
$PD_t^{cumulative}(VI)$	0.04	0.107	0.183	0.26	0.33	0.393
$PD_t^{conditional}(VI)$	0.04	0.07	0.086	0.093	0.095	0.094
Δ_t	0	0.15	0.15	0.15	0	0
$PD_t^{\Delta conditional}(VI)$	0.04	0.08	0.097	0.107	0.095	0.094
$PD_t^{\Delta cumulative}(VI)$	0.04	0.117	0.204	0.289	0.357	0.418

Table 2.3: Calculation Results for Lifetime PD Adjustments (adaption of Bosse et al., 2017b, p. 445).

	Time [years]					
	1	2	3	4	5	6
$PD_t^{cumulative}(V)$	0.02	0.046	0.078	0.113	0.15	0.188
$PD_t^{lb}(VI)$	0.029	0.074	0.125	0.179	0.231	0.28
$PD_t^{cumulative}(VI)$	0.04	0.107	0.183	0.26	0.33	0.393
$PD_t^{lb}(VII)$	0.058	0.141	0.23	0.313	0.387	0.451
$PD_t^{cumulative}(VII)$	0.08	0.184	0.286	0.378	0.456	0.522

Table 2.4: Calculation Results for Tolerance Range Derivation (adaption of Bosse et al., 2017b, p. 445).

and then the cumulative PDs according to expression (2.1). From top to bottom, Table 2.3 shows the step by step results for the lifetime PD adjustment of rating VI. Figure 2.3 shows the original and adjusted lifetime PD curves, labeled VI and ΔVI. Finally, we derive the tolerance ranges by projecting the boundary PDs of the master scale according to expression (2.3). The resulting tolerance range for rating VI is shown in Table 2.4 and Figure 2.3.

The results show that the adjustment of macroeconomic factors for years 2 to 4 does not shift the lifetime PD out of tolerance range. The adequacy test can be applied to the other ratings and corresponding lifetime PDs in the same way to prove that rating changes are the main factor for lifetime PD changes and can be used as reasonable criterion for determining SICR in accordance with paragraphs 5.5.9, B5.5.13, and B5.5.14 of IFRS 9.

2.4 Implications, recommendations, and limitations

With the identification of overall factors that drive lifetime PD changes and highlighting rating changes as necessary key factor, we answer RQ1. The applicability check demonstrates how the interdependencies between ratings and lifetime PDs can be combined to test whether rating changes are a reasonable approximation of lifetime PD changes, see RQ2. While ratings generally address information relevant for assessing changes in credit risk as listed in IFRS 9.B5.5.17, rating models were not designed with the requirements of IFRS 9 in mind (ITG, 2015b). Following, it should not be assumed that ratings will always be appropriate for determining SICR. The appropriateness of using ratings as SICR criterion "depends on whether the ratings are reviewed with sufficient frequency, include all reasonable and supportable information and reflect the risk of default over the expected life of the financial instrument" (EY Global CRS, 2018). Passing our adequacy test demonstrates that ratings reasonably reflect lifetime PDs. However, if ratings are not frequently reviewed and do not reasonably resemble credit quality in the first place, there is no point in quantitatively demonstrating that lifetime PD changes are reasonably approximated by rating changes. The application of our adequacy test is only meaningful in combination with a general validity and applicability of ratings in an IFRS 9 context, i.e. subject to the quality of information captured by ratings as listed in IFRS 9.B5.5.17.

As stated by the ITG (2015b), the adequacy of SICR criteria needs to be continuously monitored. We recommend to use the tolerance range results from expression (2.3) to derive limits for Δ_t in (2.1) for every rating. As long as changes of macroeconomic factors do not exceed the derived limits, adjusted lifetime PDs will stay within tolerance ranges and pass the adequacy test.

If lifetime PD adjustments lead to a failed test, we recommend to consider adjusting the significance thresholds rather than immediately dropping ratings as SICR criterion. Let a rating change from V to VII define SICR. Then, a financial instrument with an initial rating of V and a current rating of VI is allocated to stage 1 of the IFRS 9 ECL model, because its credit risk has not increased significantly since initial recognition. If macroeconomic downturn adjustments lead to a failed adequacy test for rating VI, the lifetime PD does not correspond to rating VI but to rating VII. Then, determining SICR for the same financial instrument based on lifetime PD changes would result in allocation to stage 2 of the IFRS 9 ECL model. The same result can be achieved by lowering the previously defined significance threshold from rating VII to VI, so SICR is triggered by a rating change from V to VI. Consequently, exceeding the tolerance range can be counteracted by lowering the significance threshold that triggers SICR. Vice versa, raising the threshold counteracts falling below the tolerance range.

Besides distinctive domain expertise, the application of our proof and adequacy test requires a certain level of technical qualification. The test can be used to monitor the adequacy of ratings as SICR criterion and as such requires regular customization and configuration. A matured version of our proof and adequacy test that is integrated into existing internal control systems requires monitoring and close collaboration between different departments of a bank. Thus, in summary, general applicability is limited by technical and communication barriers, technical skills, and the need of regular customization and configuration. Finally, as stated, the overall application of our adequacy test is limited by a general validity and applicability of ratings in an IFRS 9 context.

2.5 Further research agenda

The implementation of IFRS 9 requirements is challenging. While IFRS 9 provides opportunities to reduce complexity, the use of simplifications often suggests a devaluation of implementation quality. The opposite is true regarding IFRS 9.B5.5.13. The use of ratings as determination criterion for SICR is a simplification that increases implementation quality because it supports comprehensibility and interpretability of results of the IFRS 9 ECL model. Providing a proof for the use of ratings that meets the requirements set out in IFRS 9.B5.5.13 involves additional effort compared to the use of lifetime PDs as SICR criterion. With regard to more transparency and more meaningful and decision-relevant information for stakeholders of financial statements, the additional effort can be worthwhile. In this context, further research can concentrate on an evaluation of financial statements with regard to different SICR criteria and their effect on comprehensibility and interpretability for stakeholders. Equally important is an evaluation of effectiveness of our adequacy test with regard to the tolerance ranges. In the current setting, we use master scale midpoint PDs for the projection of tolerance ranges. An evaluation of ECL backtesting results can reveal whether a further restriction of tolerance ranges leads to adjusted significance thresholds that in turn result in more accurate ECL results.

Finally, another research opportunity is to assess whether the application of our adequacy test is beneficial for the evaluation of significance thresholds with regard to the incorporation of multiple macroeconomic scenarios when calculating ECLs, see ITG (2015a).

»Making neural nets uncool again«. The world needs everyone involved with Artificial Intelligence. Being cool is about being exclusive, and that's the opposite of what we want.

Jeremy Howard, Founder of fast.ai

3 Applying Artificial Neural Networks in Stress Testing

3.1 Theoretical background and contribution

Business Intelligence & Analytics (BI&A) describes the extensive use of data in combination with statistical techniques and other tools of quantitative analysis to improve decision making processes in companies and organizations, see, e.g., Davenport and Harris (2007); Rikhardsson and Yigitbasioglu (2018). Risk management in the context of BI&A is of central importance for the financial services industry and gained even more prominence after the 2008 global financial crisis with a constant focus on risk detection, measurement, and reporting, see, e.g., Hirtle et al. (2016); Leo, Sharma, and Maddulety (2019). In risk management, one of the most important tools is stress testing that describes a special form of scenario analysis to estimate losses in certain environments, see, e.g., Aragonés et al. (2001). Growing out of the

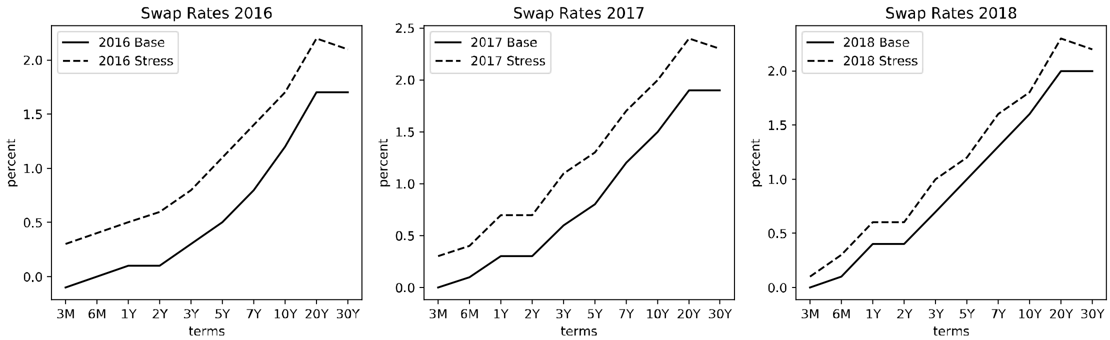

Figure 3.1: EUR Swap Rates Provided by the ESRB for the 2016 EU-wide Stress Test for Baseline and Adverse Scenarios of 2016, 2017, and 2018 (Stege et al., 2021, p. 312).

financial crisis, stress testing has evolved considerably, "but the underlying rationale remains to assure that banks can continue to supply credit to households and businesses in circumstances of deep economic and financial distress" (Kohn & Liang, 2019). Regulators provide detailed guidance to be complied with when carrying out stress tests. The guidelines "are designed to identify the relevant building blocks required for an effective stress testing programme, from simple sensitivity analysis on single risk factors or portfolios to complex macroeconomic stress testing on an institution-wide basis" (European Banking Authority, 2018). Following, the EBA intends to assist institutions to understand supervisory expectations for the use of stress testing as a risk management tool.

In 2016, the EBA and ECB together designed and conducted stress testing exercises that required all major banks in the EU to participate. The 2016 EU-wide stress test was designed with a baseline scenario and an adverse scenario. The adverse scenario reflected four systemic risks identified by the ESRB, the most significant being an abrupt reversal of compressed risk premia in the United States that would eventually trigger financial stress in other countries as well (European Systemic Risk Board, 2016). In the adverse scenario, the economic difficulties were assumed to materialize in 2016, continue in 2017, and end in 2018. The ESRB provided scenario estimates for some relevant factors, most importantly, projections for Euro (EUR) swap rates, see Figure 3.1. Estimates for other relevant variables not provided by the ESRB had to be generated in a way that ensured compatibility with the forecasts of the official scenarios of the ESRB. Many banks used techniques of time series analysis and other quantitative modeling strategies to produce the required additional forecasts, or relied on expert judgment.

We demonstrate how innovative techniques from machine learning and time series analysis can be employed to map macroeconomic variables provided by regulators to covered bond yields that are an important source of funding for many European banks (Stege et al., 2021). For this, we adapt a multi-methodology approach by Wegener et al. (2016) that combines Artificial Neural Networks (ANNs) with techniques of cointegration analysis. We address the following research question:

RQ3 How can ANNs be used to produce adequate projections of macroeconomic factors consistent with regulatory guidelines on stress testing?

We discuss the application of techniques from BI&A in the banking industry to address stress testing, a challenge in applied financial risk management. On the one hand, our approach and findings are particularly relevant for risk managers that have to map interest rate projections for one specific financial instrument on bond yields of other fixed income securities. On the other hand, our approach contributes to accounting in general, as it can be applied by auditors and regulators of financial institutions to validate the execution of stress tests and compliance with the guidelines provided.

3.2 Methodology

3.2.1 General approach

We use ANNs to map the provided scenarios for EUR swap rates to other factors relevant for the financial services industry, so that their projections can be used for stress testing. Following Qi and Zhang (2008), who analyze how to use ANNs to best model and forecast time series, it can be helpful to examine differenced data rather than time series in levels. Under certain circumstances, namely cointegration among analyzed time series, the examination of differenced time series without consideration of cointegration will lead to a loss of information (Wegener et al., 2016). This implies that neglecting cointegration when combining time series with ANNs can lead to an error like in the linear framework. We adapt the multi-methodology approach by Wegener et al. (2016) to enhance ANNs when mapping the provided scenarios for EUR swap rates to other time series. Our approach is summarized by the following process:

1. Prepare time series data,
2. test time series for cointegration,
3. specify ANNs with and without cointegration information, and
4. evaluate accuracy of ANNs.

While step 1 refers to common pre-processing tasks when working with time series data, steps 2 to 4 are described in more detail in the following sections.

3.2.2 Cointegration analysis

As proposed by Engle and Granger (1987), cointegration exists when two or more time series are individually non-stationary, but share a common stochastic trend, so that a linear combination of the time series is stationary. To test whether a time series is non-stationary, a unit root test can be applied. Typical unit root tests are the Augmented Dickey-Fuller test (ADF) and the Phillips-Perron test (PP), both testing the null hypothesis that a unit root exists in time series data. If the null hypothesis cannot be rejected, the time series are non-stationary.

The next step is to test whether non-stationary time series are cointegrated. Typical cointegration tests are the Augmented Engle-Granger test (AEG) and the Phillips-Ouliaris test, both testing the null hypothesis that there is no cointegration. If the null hypothesis is rejected, the non-stationary time series are cointegrated.

Finally, in traditional time series analysis, the Vector Error Correction Model (VECM) is applied when cointegration exists. VECMs include ECTs to account for the long-run cointegration relationship between time series.

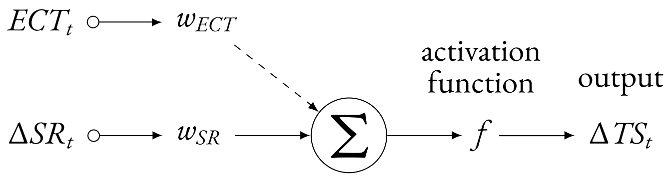

Figure 3.2: ANN Design of Model Types A and B (adaption of Stege et al., 2021, p.316).

Thus, following Wegener et al. (2016), we consider an »Error Correction Neural Network« with ECTs as additional input similar to the linear VECM to account for the long-run relationship.

3.2.3 Artificial neural network specification

In the context of time series modeling, the results of Qi and Zhang (2008) indicate that ANNs perform better when the underlying variables are stationary. As stated, if stationary variables result from non-stationary variables that are cointegrated, neglecting the cointegration relation would lead to a loss of information. We use two ANNs, one that accounts for cointegration by using ECTs as an additional input, and the other without additional information. The objective is to test and demonstrate that the ANN with ECT input outperforms the other without ECTs.

We strive for a simple and transparent ANN setup to increase trust and acceptance for practitioners applying the approach, so we keep the minimalistic design of a three-layered feedforward ANN as proposed by Wegener et al. (2016). The first layer of an ANN is the input layer that describes the independent variables, i.e. features. As stated, we design two ANNs, model »A« with swap rates (SR) as single input and model »B« with both SR and ECTs as inputs. The third layer is the output layer that describes the dependent variables, i.e. responses. In our approach, the responses are the time series to be used in the stress test. Figure 3.2 shows the ANN design where model A is illustrated by solid lines and the additional input for model B is illustrated by the dashed line. The output of the models is labeled »ΔTS« and describes the time series in differences used for the stress test. The second layer, i.e. hidden layer, transforms the weighted sum of inputs by a non-linear activation function, e.g. the hyperbolic tangent, see Gleue, Eilers, von Mettenheim, and Breitner (2019):

$$f(X) = tanh(X) = 1 - \frac{2}{e^{2X} + 1}. \tag{3.1}$$

For better convergence when using (3.1), we normalize the input and output data to the interval [-1, 1] as follows:

$$\text{norm}(\Delta X) = 2 \cdot \frac{\Delta X - min(\Delta X)}{max(\Delta X) - min(\Delta X)} - 1. \tag{3.2}$$

In contrast to the approach of Wegener et al. (2016), the purpose of our approach is not to forecast, but instead provide a qualified mapping of time series. Consequently, we follow a common deep learning rule of thumb to select the number of neurons in the hidden layer:

$$n_{\text{hidden}} = \frac{n_{\text{observations}}}{\omega \cdot (n_{\text{input}} + n_{\text{output}})}, \tag{3.3}$$

where ω is a scaling factor that describes the generalization degree of the model. The higher ω, the more general the ANNs. Vice versa, the smaller ω, the more the ANNs memorize the training sample to perfection, resulting in poor performance for any observations not part of the training data. For better comparison, we keep both models as identical as possible and set up the hidden layers with the same number of neurons by using a higher ω for model type A. We compensate the difference in ω with an early stopping algorithm that sets a maximum of T epochs for fitting the ANNs on the training set and stops when evaluation on the validation set has not improved for the last p epochs. For details, see Stege et al. (2021).

3.2.4 Evaluation of accuracy

We use the mean squared error to evaluate the models' accuracy and compare their performance. Then, to provide statistical evidence, we apply normality tests to ascertain whether parametric or non-parametric tests can be used for the statistical evaluation of the models' performance. Typical normality tests are Shapiro-Wilk (Shapiro & Wilk, 1965), D'Agostino-Pearson (D'Agostino

& Pearson, 1973), and Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1939). The null hypothesis of each test is that the data follows a normal distribution. If rejected, non-parametric tests can be used for statistical evaluation. Typical non-parametric tests are Wilcoxon (Wilcoxon, 1945), Binomial Sign and Mann-Whitney-U (Mann & Whitney, 1947), each with null hypothesis that the medians of the differences between two groups of data are equal. If rejected, it can be inferred that the medians of the mean squared errors of the models differ systematically, so that it is statistically evident that one model is superior to the other.

3.3 Applicability check

We examine German covered bonds, i.e. »Pfandbriefe«, because they are an important liability for German banks, but were not provided for the scenarios of the stress test. We use monthly data from January 1999 to December 2016 for three maturities, 2, 5, and 10 years. Both SR and covered bond yields (CBY) are extracted from Bloomberg. For the 10 year CBY, we consider the more liquid »Jumbo Pfandbrief« debt securities.

The results of the unit root tests do not clearly indicate that the time series are non-stationary, see Table 3.1. Despite our ambiguous test results, we follow the argumentation and assumptions of Wegener, Basse, Sibbertsen, and Nguyen (2019) whose examination of CBY indicate that the time series are integrated of order 1, i.e. non-stationary. We then test for cointegration, see Tables 3.2 and 3.3 for results. Subject to critical values, the results indicate that the time series are cointegrated, so it is possible to describe the relationship between the time series with ECTs, see, e.g., Asteriou and Hall (2021). We employ the methodology suggested by Engle and Granger (1987) to generate the ECTs.

We end up with 216 observations per time series as we examine SR and CBY in differences and use the ECTs for model B with a lag of one. We use a split ratio of 0.7, leading to training sets with 151 observations, and select a low ω in expression (3.3) to place 20 neurons in the hidden layers of both model types. For an overview of the final ANN specifications, see Table 3.4.

	Test stat. (levels)	Critical value 5%	Test stat. (first diff.)	Critical value 5%	Test stat. (linear)	Critical value 5%
ADF test						
SR 2Y	-0.76	-2.87	-10.86	-2.87	-2.51	-3.43
SR 5Y	-0.46	-2.87	-11.90	-2.87	-2.89	-3.43
SR 10Y	0.06	-2.87	-12.79	-2.87	-3.27	-3.43
CBY 2Y	-0.52	-2.87	-11.91	-2.87	-2.38	-3.43
CBY 5Y	0.05	-2.87	-12.41	-2.87	-2.80	-3.43
CBY 10Y	0.07	-2.87	-13.48	-2.87	-2.55	-3.43
PP test						
SR 2Y	-0.79	-2.87	-11.13	-2.87	-2.70	-3.43
SR 5Y	-0.44	-2.87	-12.01	-2.87	-3.12	-3.43
SR 10Y	-0.20	-2.87	-12.81	-2.87	-3.46	-3.43
CBY 2Y	-0.60	-2.87	-12.11	-2.87	-2.58	-3.43
CBY 5Y	-0.26	-2.87	-12.49	-2.87	-2.99	-3.43
CBY 10Y	-0.06	-2.87	-13.48	-2.87	-2.64	-3.43

Table 3.1: Unit Root Test Results for 2, 5, and 10 Year EUR SR and German CBY (adaption of Stege et al., 2021, p. 315).

	Test stat.	Critical value 10%	Critical value 5%
AEG test			
2Y	-4.44	-2.91	-3.17
5Y	-3.61	-2.91	-3.17
10Y	-3.13	-2.91	-3.17

Table 3.2: AEG Cointegration Test Results for 2, 5, and 10 Year EUR SR and German CBY (adaption of Stege et al., 2021, p. 316).

	Test stat. (tau)	p-value	Test stat. (z)	p-value
Phillips-Ouliaris test				
2Y	-3.47	0.039	-23.71	0.022
5Y	-3.32	0.056	-21.70	0.035
10Y	-3.77	0.058	-25.52	0.059

Table 3.3: Phillips-Ouliaris Cointegration Test Results for 2, 5, and 10 Year EUR SR and German CBY (adaption of Stege et al., 2021, p. 316).

Model	Input	Hidden	Output
A2	1 ΔSR(2 year)	20	1 ΔCBY(2 year)
A5	1 ΔSR(5 year)	20	1 ΔCBY(5 year)
A10	1 ΔSR(10 year)	20	1 ΔCBY(10 year)
B2	1 ΔSR(2 year) 2 ECT(2 year, lag 1)	20	1 ΔCBY(2 year)
B5	1 ΔSR(5 year) 2 ECT(5 year, lag 1)	20	1 ΔCBY(5 year)
B10	1 ΔSR(10 year) 2 ECT(10 year, lag 1)	20	1 ΔCBY(10 year)

Table 3.4: Overview of ANN Specifications (adaption of Stege et al., 2021, p. 318).

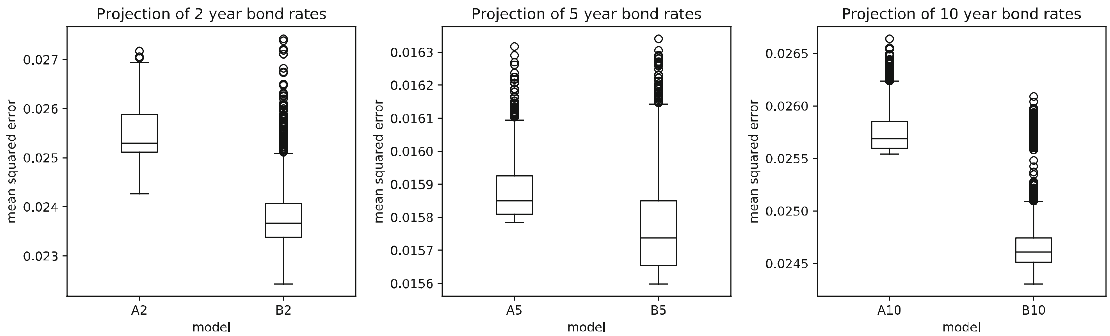

Figure 3.3: Comparison of Model Accuracy (Stege et al., 2021, p. 318).

We fit 12,000 models in total with different random seeds for 20 different subsets of data to eliminate potential dependencies between initially randomized weights and results. To compare model results, we use boxplots, see Figure 3.3. The results show that model B outperforms model A for all evaluated maturities. This clearly indicates that the use of cointegration information as additional ANN feature can improve model accuracy.

For statistical evaluation, we use Statistical Tests for Algorithms Comparison (STAC) by Rodríguez-Fdez, Canosa, Mucientes, and Bugarín (2015) and VassarStats by Lowry (2017). We apply normality tests to the mean squared errors of each model. The results of the normality tests show that the mean squared errors are not normally distributed, i.e. the null hypothesis is rejected

Model	Sh.-Wilk	D'Ag.-Pears.	Kol.-Smir.	Wilcox.	Bin. Sign	M.-Wh.-U
A2	0.93817 (0.000)	131.496 (0.000)	0.50968 (0.000)	1,994,728 (0.000)	1,963 (0.000)	125,851 (0.000)
B2	0.89019 (0.000)	677.138 (0.000)	0.50895 (0.000)			
A5	0.89376 (0.000)	427.818 (0.000)	0.50622 (0.000)	836,704 (0.000)	1,408 (0.000)	1,348,896 (0.000)
B5	0.82279 (0.000)	627.624 (0.000)	0.50627 (0.000)			
A10	0.75189 (0.000)	962.064 (0.000)	0.50969 (0.000)	2,001,000 (0.000)	2,001 (0.000)	120,298 (0.000)
B10	0.85299 (0.000)	436.670 (0.000)	0.51019 (0.000)			

The table reports the test statistics and p-values (in parentheses) for each test.

Table 3.5: Test Results of the Statistical Evaluation of Model Accuracy (Stege et al., 2021, p. 319).

for each test. We then apply non-parametric median tests to model pairs. The tests show that the medians of model errors differ systematically when comparing model types A and B. Thus, it is statistically evident that model type B is superior to model type A. The test results are reported in table 3.5.

3.4 Implications, recommendations, and limitations

The empirical evidence from the applicability check shows that using cointegration information as additional feature improves model accuracy when mapping SR to CBY with ANNs. While the overall improvement of the projections under consideration of cointegration is rather small, we show how ANNs can be used in general to produce projections of scenarios that are consistent with regulatory guidelines, see RQ3. On the one hand, our approach follows a step-by-step process that facilitates applicability for practitioners. On the other hand, the application is subject to numerous assumptions that affect and limit the efficiency and accuracy of the approach. To name a few, the number of hidden layers, the number of neurons in each layer, the choice of activation functions, and the specification of hyperparameters, such as the learning rate, all affect results and performance of the ANNs. There are multiple methods that can be used to close in on the optimal ANN specifications. This is a complex task in itself that often comes at cost of clarity and simplicity of the final model. With respect to RQ3, the objective is to produce adequate, not optimal projections. Thus, we recommend to use a setup that is as simple and transparent as possible, even if this may negatively affect accuracy of the model.

The overall applicability of our approach is limited by the different ways to apply and interpret the results of statistical tests. From a theoretical point of view, we successfully obtain statistical evidence that corroborates the superiority of using ECTs as an additional feature when producing scenario projections with ANNs. From a practical point of view, it remains questionable whether our approach is superior to classical approaches already used by practitioners in the financial services industry.

3.5 Further research agenda

We lay out a new technique of empirical economics that combines traditional time series analysis with machine learning in the form of neural networks. Our applicability check shows how the approach can generally be employed by risk managers working in the financial services industry. In its current state, the high number of existing assumptions is likely to limit acceptance and user friendliness of the approach. This results in great potential for further research with implications for both theory and practice. Following, additional empirical evidence is necessary to evaluate whether our approach can be a useful tool with regard to overall applicability and quality of results when compared to established approaches from practice.

Finally, from a regulatory perspective, our approach may be helpful for auditors and regulators to test for compatibility and adequacy of the projections used by stress test participants. In fact, there is great potential for future research regarding the adaption of the approach as a useful stress test validation tool for the audit.

Data visualization can help us both to understand complex issues a bit better, but also to provide images to debate about, to refer back to, and sometimes just to meditate over.

Moritz Stefaner

4 Visualizing the Unusual

4.1 Theoretical background and contribution

The integration of data analytics in business processes and workflows for data-driven decision support has been an important challenge for companies and organizations for decades (Chen et al., 2012). Increasingly complex data-driven tasks need to be managed by highly specialized teams of data analysts, but while data analysts have the essential technical skills, in the working world the most valuable domain knowledge resides with experienced managers and decision-makers (Alpar & Schulz, 2016; Eilers et al., 2017). Following Bresciani and Eppler (2009), collaboration and knowledge sharing in organizations improve significantly through the use of interactive visualization. Visualization methods in an analytics context, i.e. visual analytics, has been an active research field with applications in many sectors (Sun et al., 2013). A powerful set of algorithms, methods, and tools is available for visual analytics, but insights do not emerge automatically from application (Sharma et al.,

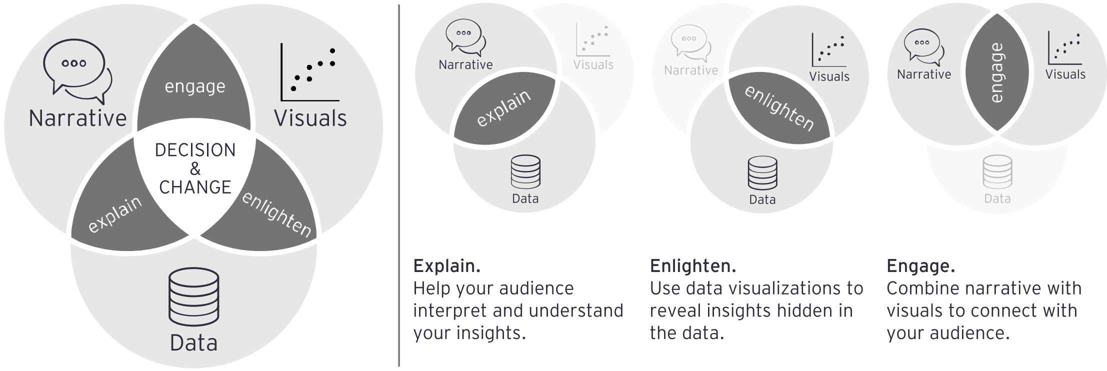

Figure 4.1: The Key Elements of Storytelling (adaption of Dykes, 2020, pp. 31–32).

2014). This is emphasized by Endert et al. (2014), who state that interaction is the critical glue that integrates analytics, visualization, and humans, who often face results without understanding their relevance and meaning. They study a typical analysis task defined as storytelling that describes the investigative process of connecting the dots between seemingly disconnected information (Hossain et al., 2011; Thomas, 2005). More precisely, following Dykes (2020), storytelling is a combination of three key elements, data, narrative, and visuals, that intersect to drive well-founded decisions and change, see Figure 4.1. Sophisticated analytic support for storytelling remains a significant area for research and has found its way into many sectors. In auditing, for example, accountants and auditors are increasingly using BI&A and visual analytics, i.e. storytelling, to communicate results (Schneider, Dai, Janvrin, Ajayi, & Raschke, 2015; Sekar, 2022). Audit efficiency highly depends on the auditors' competency "in recognizing patterns in financial data and in hypothesizing likely causes of those patterns to serve as a guide for further testing" and investigation (Bedard & Biggs, 1991). For this, auditors are often assisted by specialists from, e.g., Information Technology (IT) and forensic accounting, but collaboration of specialists and audit team members is a challenge for both groups (Boritz, Kochetova, Robinson, & Wong, 2020). The competency in storytelling can improve collaboration of auditors and specialists to jointly increase the overall quality of an audit. This is also emphasized by Bolt and Tregidga (2023), who explore the implications of storytelling on how people understand and make sense of materiality, which is a fundamental concept in financial and non-financial reporting and auditing. In addition to their findings, they conclude that the potential of storytelling in general relates to "how individuals who hold different pieces of information are able to collectively construct new meaning" (Maitlis & Christianson, 2014). Finally, auditing is particularly suited for advanced data analytics and visualization techniques due to the combination of numerous repetitive tasks and the challenge to gain insight from vast volumes of data, with the overall objective to perform analytical audit procedures not only based on samples, but on entire company transactions, see Kokina and Davenport (2017).

Against this background, we develop a visualization framework to identify and assess unusual items in financial data, which is a typical analytical tasks for auditors (Stege & Breitner, 2020). The framework improves the interactive process of engagement between data analysts and auditors. As our method is not limited to auditing, we generalize it for application in many domains where structured data is available (Stege & Breitner, 2023). We address the following research questions:

RQ4 How can visual analytics be embedded into a decision-making process to improve collaboration between data analysts and domain experts?

RQ5 How can the visualization framework be efficiently applied in practice to support anomaly explanation, well-founded decisions, and storytelling?

The need to present more complex data in more intuitively understandable and informative ways, and the increasing sophistication of data visualization in general, are key challenges for efficient IT driven processes (Lycett, 2013). As stated, this is of particular relevance with regard to audits and the integration of specialists, because the quality of audits is critically dependent on judgment and the derivation of conclusions regarding the financial statements of companies and organizations (Knechel, 2016). Our generalized process model and visualization framework contribute to interactive visual analytics and profound storytelling in auditing and many other domains where structured data is available.

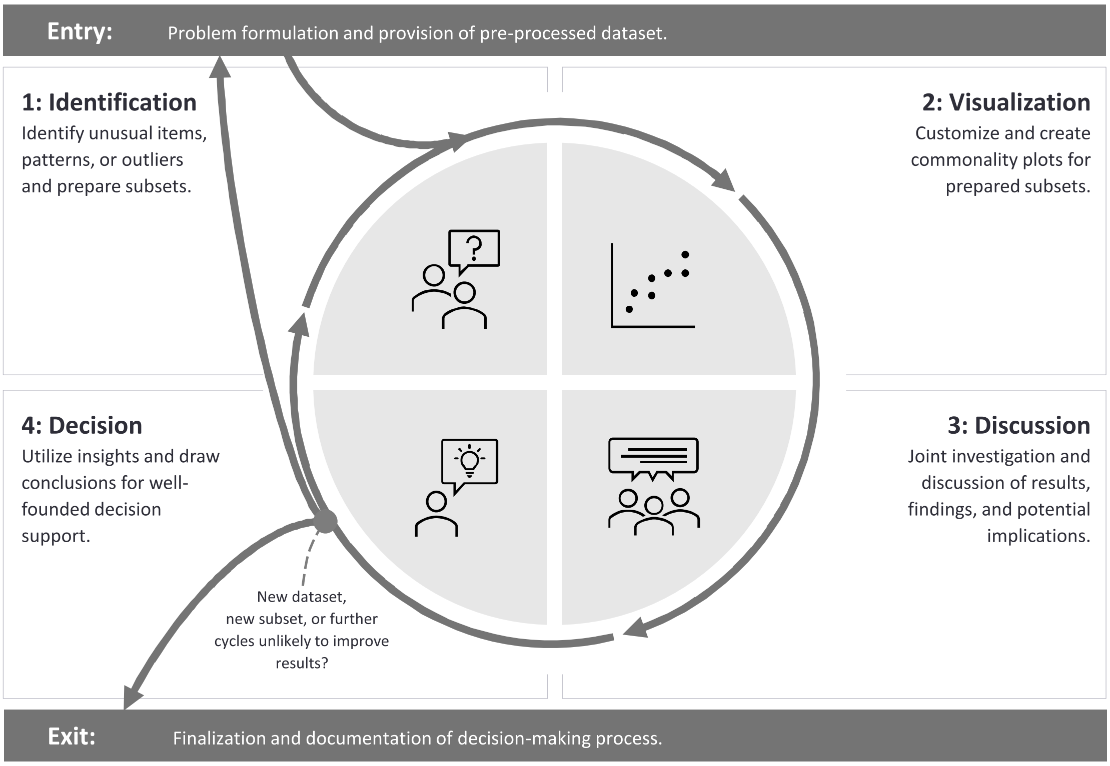

Figure 4.2: Interactive Process Model for Decision Support (Stege & Breitner, 2023, p. 6).

4.2 Methodology

4.2.1 Interactive process model

Our process model contains four phases that are run through sequentially and cyclically until a well-founded decision is reached, see Figure 4.2. The process is initiated with a pre-processed dataset that relates to a specific problem statement. The objective of the first phase is to identify subsets from the dataset that provide insights for the problem statement at hand. In general, the subsets can be identified manually through domain knowledge and experience or through the application of statistical and analytical techniques. In the second phase, subset-dataset relationships are visualized using a technique we call »commonality plots«. The plots are specifically designed to visualize commonalities in the subsets and highlight items that have a low likelihood of occurrence, see the following sections for a description of the underlying concept and general setup. The visualizations are used as a basis for the third phase of our process, where domain experts and data analysts jointly investigate and discuss results and findings to decide whether these provide helpful insights with respect to the problem statement. The objective of the fourth phase is to draw conclusions from the results of the discussion and utilize them for decision support. If there are follow-up questions from discussion or a need to investigate further subsets, the identification phase can be re-entered. If results indicate that the main dataset needs adjustments, e.g. additional features or a different approach to impute missing values, the process can be re-started with a new or adjusted main dataset. Finally, if the analysis of further subsets is unlikely to improve results, the process is exited and the visual analysis is finished.

In summary, our process describes how domain experts and data analysts can collaboratively use visual analytics for well-founded decision support and storytelling. The resulting commonality plots can also be used for the overall documentation of the decision-making process.

4.2.2 The concept of commonality

Our whole concept of commonality is built on »cardinality«, which describes the number of different elements in a list of values. For example, the list $\{0, 0, 0, 1\}$ has two different elements 0 and 1, so its cardinality is 2. We introduce a basic formal notation to describe subset-dataset relationships in the context of cardinality. Let D describe the main dataset with a certain number of columns (features) n_c and rows n_r. From D, we select a sample of rows that form subset d. Then d has the same number of features as D, but fewer rows, so $d \subset D$ with $n_c^d = n_c^D$ and $n_r^d < n_r^D$. Let $|d[i]|$ describe the cardinality of a single feature i from subset d. Then, by definition, the minimum of $|d[i]|$ is 1 if all values of $d[i]$ are identical, and the maximum of $|d[i]|$ is equal to the number of rows n_r^d if all values in $d[i]$ are different. Based on this setting, we define that there is commonality if the cardinality of a feature of a subset equals 1:

$$Commonality := |d[i]| \stackrel{!}{=} 1. \tag{4.1}$$

Our definition of commonality works for every data type, but is not always meaningful. Following Boslaugh (2012), data can be broadly classified into categorical data and continuous data. Then, commonality as in (4.1) is only applicable for either categorical data or for continuous data that is categorized prior to application. When categorization of continuous data is not possible, another approach for measuring commonality can be defined based on the value distribution of a continuous feature. Let f describe the probability density function of the distribution of a feature in the main dataset $D[i]$. Then, following Silverman (2018); Wand and Jones (1994), kernel density estimation can be used to derive an estimated density function as follows:

$$\hat{f}(D[i]) = \frac{1}{nh} \cdot \sum_{j=1}^{n} K\left(\frac{D[i] - X_j}{h}\right), \tag{4.2}$$

where h describes the width of kernels K that are added up to form the estimated density curve. Now, we say there is commonality for continuous data, i.e. continuous commonality, if all values in $d[i]$ lie close to each other when compared to the entire distribution $\widehat{f}(D[i])$.

4.2.3 The measurement of commonality

Our concept of commonality indicates whether items in a subset have anything in common. Next, we need a measure that indicates how unusual commonality is, i.e. a commonality likelihood measure between 0 and 1. We can picture our problem with the famous saying »to find a needle in the haystack«, where subset d is the needle and the main dataset D is the haystack. Typically, finding the needle is highly unlikely, i.e. a probability close to 0. However, if the entire haystack is full of identical needles, then finding a needle is highly probable, i.e. a probability close to 1. For categorical commonality, we can calculate these probabilities from combinatorics using a simplified form of the classical approaches of Hoadley (1969) and Janardan (1973) to calculate the probability p to obtain a specific composition of items in a sample:

$$p = \frac{\binom{n_v^D}{n_r^d}}{\binom{n_r^D}{n_r^d}}, \tag{4.3}$$

where n_v^D is the total number of values v of a distinct element in data field $D[i]$. The probability measure in (4.3) is closer to zero the larger the difference between n_r^D and n_v^D (i.e., there are very few needles in the haystack), and is equal to 1 in the extreme case $n_r^D = n_v^D$ (i.e., the entire haystack is full of needles). If there is no commonality as defined in (4.1), we use the expected value of the binomial distribution as indicator for unusualness, because the probability measure in (4.3) is too sensitive for large datasets and easily converges to zero.

For continuous commonality likelihood, we calculate the probabilities directly from the density function in (4.2):

$$p = \int_{min(d[i])}^{max(d[i])} \hat{f}(D[i])dD[i], \tag{4.4}$$

where $[min(d[i]),\ max(d[i])]$ is the interval of the value range of the continuous data in the subset. The function returns 0 if all values in the subset are identical, and it returns 1 if the value ranges of the subset and the main dataset are equal.

The expressions in (4.3) and (4.4) return probabilities of occurrence. For our final measure of unusualness of occurrence, we use the complementary probabilities $1 - p$.

4.2.4 The visualization of commonality

Our approach to visualize commonality is inspired by Zoomable Circle Packing by Bostock (2018), which is closely related to treemaps. In comparison to treemaps, circle packing does not use space as efficiently, but better reveals the underlying hierarchy of data (Heer, Bostock, & Ogievetsky, 2010). This is very useful for the visualization of subset-dataset relationships. We create the commonality plots for each feature i of D as follows:

1. We use the letter-value box technique by Hofmann, Wickham, and

Kafadar (2017) to categorize continuous features.

2. We adapt the circle packing algorithm of Bostock (2018) to create a circle for each distinct element in a feature $D[i]$ and to position the resulting circles on a plot. The circle sizes are subject to the number of values n_v^D of each element. The number of circles is equal to the cardinality $|D[i]|$. If the number of resulting circles is too high, we group the circles by size, starting with the smallest, until an adequate number of circles is reached.

3. We treat missing values as a category that is visualized with different hatching. For grouped values, we use different hatching as well.

4. We repeat the second step for the subset $d[i]$ and position the resulting circles on top of the existing circles of $D[i]$.

5. For continuous features, we create an additional visualization of the original distribution of the data to compensate for the loss of information through categorization. The visualization is a combination of rug plots to display the original distribution and one-dimensional scatter plots to project the subset values onto the rug plots, see Stege and Breitner (2020).

In the machine learning community, one of the most popular datasets is the Titanic dataset that contains information about the passengers aboard the famous ship and whether they survived the sinking, see Kaggle (2012). We use the dataset to illustrate the features of commonality plots. Table 4.1 summarizes the dataset.

First, we draw an arbitrary sample from the dataset. For this, we select all passengers who are 23 years old, so there is complete commonality for the feature »Age« in the subset, i.e. $|d[Age]| = 1$, because all values are equal to 23. We then visualize the subset-dataset relationships with commonality plots. Figure 4.3 shows the resulting commonality plots for three features of the dataset, »Age«, »Cabin«, and »Fare«. The dark gray circles show the values of the subset in comparison to the original value distribution of the main

i	Feature	Cardinality	Data type	Description
1	Sex	2	text	F = female, M = male
2	Survived	2	int	0 = No, 1 = Yes
3	Pclass	3	int	Ticket class
4	Embarked	4	text	Port of embarkation
5	Parch	7	int	Number of parents/children
6	SibSp	7	text	Number of siblings/spouses
7	Age	89	int	Age in years
8	Cabin	148	text	Cabin number
9	Fare	248	float	Ticket price
10	Ticket	681	text	Ticket number
11	Name	891	text	Passenger name
12	PassengerId	891	int	Identification number

Table 4.1: Titanic Dataset Summary (891 Rows and 12 Columns).

dataset. The features »Age« and »Fare« are numerical, so their values were grouped prior to plotting. As the subset for »Age« only contains value 23, there is a single dark gray circle visible in the larger circle of the main dataset that includes all ages within range $[20.1, 38.0]$. For illustration purposes, we choose not to replace missing values and to visualize them with striped hatching. The feature »Cabin« describes the cabin labels of the passengers. Apparently, the information is not available for the majority of passengers. Also, there are values that appear only once in the dataset, i.e. individual values. We group individuals and visualize them with a dotted circle. The numerical feature »Fare« describes the ticket price that passengers paid to get aboard. The one-dimensional rug plot of the distribution indicates that it is right-skewed, so low ticket prices are much more common than expensive tickets. The line plotted on top of the distribution shows the value range of ticket prices in the subset, i.e. the distribution of ticket prices paid by 23-year-old passengers. Although feature »Age« is numeric as well, there is no plot of the distribution, because kernel density estimation does not work in case there are missing values. The likelihood measures are presented on the top right corner of each visualization. Labels can be added for all categories or selected categories, e.g. those that caused a certain level of unusualness. In Figure 4.3, the common-

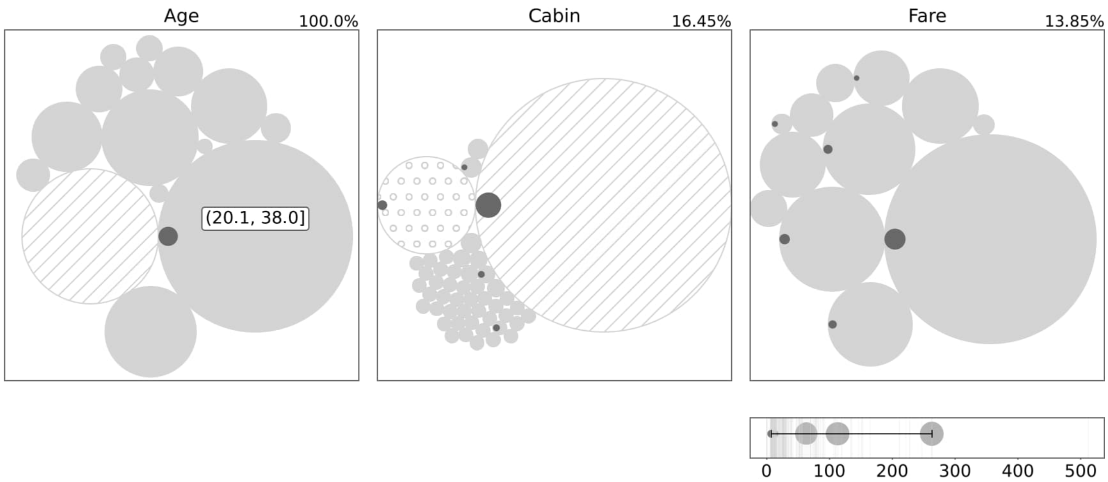

Figure 4.3: Commonality Plots for Selected Data Fields of the Titanic Dataset.

ality for feature »Age« is highly unusual ($1 - p = 100.0\%$), so the label for the corresponding category is shown. We omit the commonality plots for the remaining features of the dataset because there is no additional explanatory gain from illustration.

4.3 Applicability check

This applicability check is based on a problem statement directly linked to our research in Accounting and Finance summarized in Chapters 2 and 3. As stated, financial institutions must establish provisions to prepare for potential credit losses, see Sections 1.1.1 and 1.1.2. In this context, the IFRS 9 ECL model from Chapter 2 and stress testing from Chapter 3 are both important tools for the estimation of adequate loan loss provisions and equity capital. Auditors are required to assure that financial institutions fulfill the capital requirements and the regulations on loan loss provisioning. For their analysis, they require detailed financial information from audit clients. We use an anonymized sample of such a dataset. The sample contains information about 1000 credit agreements of a European bank, see Table 4.2.

We follow our process model as set out in 4.2.1. In the identification phase, we select two subsets from the dataset, one in the role of an auditor based on domain knowledge and one as data analyst applying pattern recognition al-

i	Feature	Cardinality	Data type	Description
1	rep_date	1	date	Reporting date
2	flag_watchlist	2	float	Watchlist flag
3	portfolio	2	text	Name of portfolio
4	stage	3	int	Risk class of ECL model
5	cntr_id	9	text	Country of origin
6	product_type	11	text	Product type
7	init_rating	25	int	Initial rating at recognition
8	rating	26	int	Current rating
9	contract_age	28	text	Age of contract in years
10	loss_provision	296	float	Loan loss provision
11	start_date	344	date	Date of initial recognition
12	prob_dflt	757	float	PD
13	contract_id	1000	text	Identification number

Table 4.2: European Bank Dataset Summary (1000 Rows and 13 Columns), (adaption of Stege & Breitner, 2023, p. 16).

gorithms. In the visualization phase, we create commonality plots for further investigation and discussion. We use results and findings for well-founded decision support and storytelling.

For the first subset, we select all items from the dataset with an initial rating larger than 20 for further investigation, because according to the bank's lending policy, loans will not be granted to debtors with a credit rating worse than 20. Figure 4.4 shows the resulting commonality plots. The plots reveal that all items have the same start date, age, product type, country, and reporting date. For reasons of anonymity, we have omitted the country name. The discussion revolves around the visualizations:

- The plot for »init_rating« confirms the subset selection, i.e. $D[init\_rating] > 20$.
- Although there is commonality for the feature »rep_date«, this is not unusual at all ($1-p = 0.0\%$). This is due to the fact that the whole dataset is a snapshot at a single reporting date, so all items in the dataset have the same reporting date, i.e. $|D[i]| = 1$ implies $|d[i]| = 1$.
- The commonality for feature »cntr_id« is not unusual ($1-p = 26.63\%$),

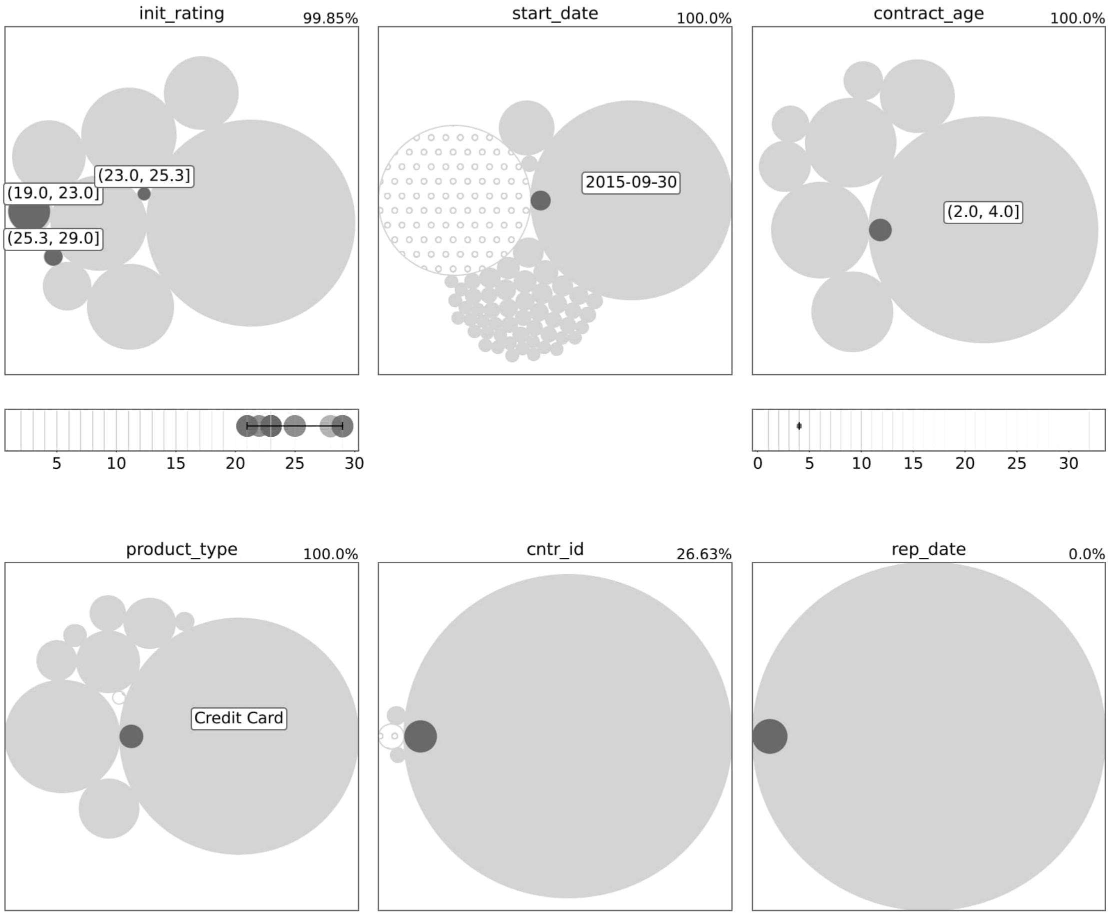

Figure 4.4: Commonality Plots for Selected Data Fields of the Bank Dataset (Subset 1).

because the majority of the dataset originates from the same country.

- The features »start_date« and »contract_age« are closely related, because the age of the contract equals the difference between reporting date and date of initial recognition of the credit agreement. Nevertheless, the commonality plots indicate that it is highly unusual to end up with the same start date and age of the contracts.

The insights from the commonality plots are used as a basis to question the bank. It turns out that the bank merged with a retail bank in September 2015. When the existing contracts of the retail bank were added to the portfolio of the bank, the start date was set to the date of the merger and not to the original date of initial recognition. This directly affects the determination of SICR

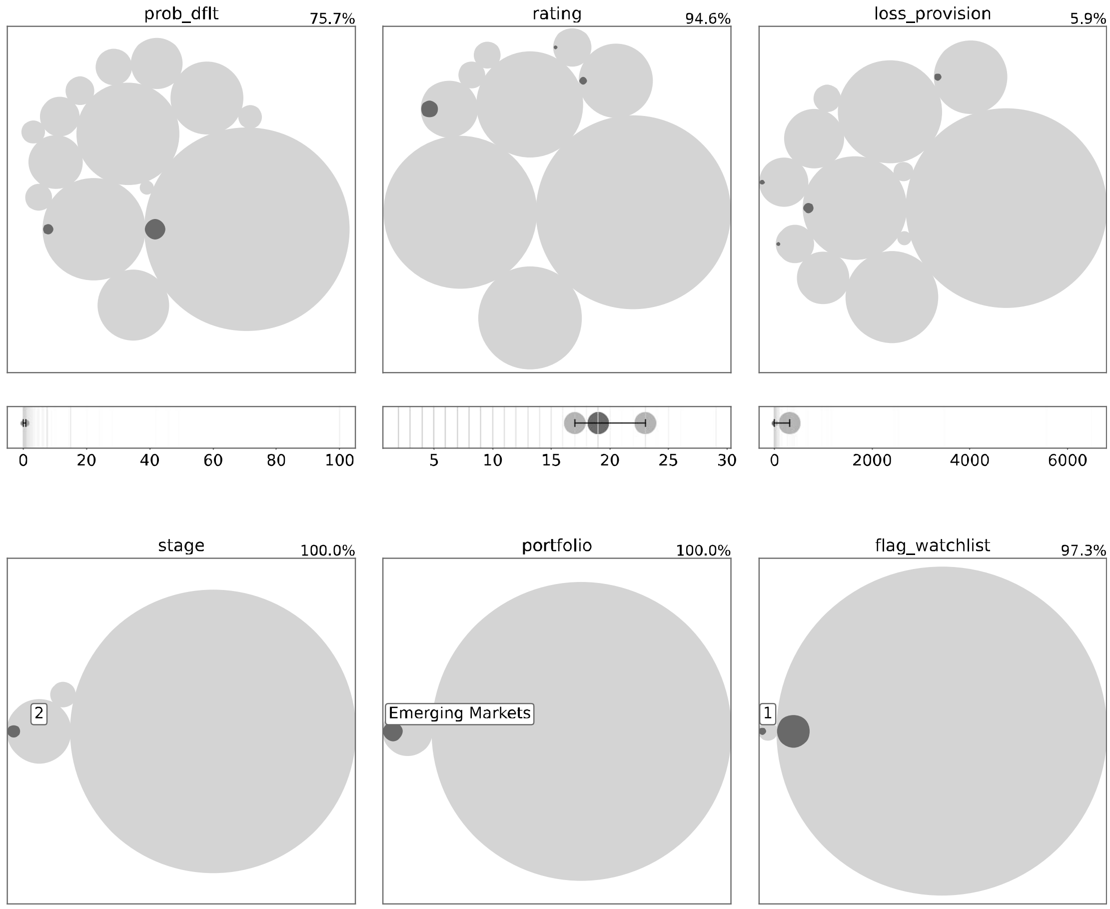

Figure 4.5: Commonality Plots for Selected Data Fields of the Bank Dataset (Subset 2), (Stege & Breitner, 2023, p. 17).

(see Chapter 2), because the credit quality at origination is directly linked to the start date. In turn, this affects the results of the IFRS 9 ECL model, which, if undetected, may lead to misstatements in the bank's financial statement.

For the second subset, we apply pattern recognition algorithms to all numerical features and identify an unusual value cluster regarding the PD. We select all items from that cluster for further investigation. Figure 4.5 shows the resulting commonality plots that are the basis for discussion:

- The plot for »prob_dflt« reflects the low PD value cluster resulting from pattern recognition algorithm.
- The low PDs do not relate to the credit ratings of the items that show

poor credit quality, see the distribution of »rating«.

- All items are allocated to stage 2 of the IFRS 9 ECL model, see »stage«. Following, the credit quality of the items has deteriorated significantly since initial recognition, which implies the calculation of lifetime ECLs for loan loss provisioning.
- Despite SICR and poor credit quality with regard to the rating, the loan loss provision for the items is fairly low, see the distribution of »loss_provision«.
- The plot for »portfolio« shows that all items are from the same portfolio, which is highly unusual.
- Some of the items have a watchlist flag that indicates loan repayment issues, see »flag_watchlist«, which also contradicts low PDs.

The insights are again used as a basis to question the bank. It turns out that a subsidiary of the bank reported percentages in a different value range. The bank uses $[0, 100]$ for disclosure, whereas the subsidiary uses $[0, 1]$. These values were also used for the calculation of loan loss provisions which lead to a significant undervaluation of the credit risk. As with the first subset, if undetected, the results would have lead to misstatements in the bank's financial statement.

4.4 Implications, recommendations, and limitations

With our general process model for the visual analysis of subset-dataset relationships, we answer RQ4. The applicability check demonstrates how our process can be applied by domain experts and data analysts in practice, see RQ5. The results show how the visualizations and insights can be used to support storytelling and well-founded decisions. In our case, the visualized findings caused a correction of numbers that are ultimately disclosed in the bank's financial statements. Although our intention is not to measure exact probabilities, our likelihood measures can still have a significant impact on how results are presented to users, which in turn influences decision-making.

This also applies to the categorization of values, handling missing values, and general visual customization such as coloring and hatching.

Prior to the identification of subsets, we recommend a dry run of our process, meaning to visualize only the main dataset without subsets. This helps users to become familiar with the underlying dataset and to discuss categorization and how to handle missing values. When imputing missing values of numerical features, we found it very helpful to use a different color in the distribution to mark the value used for imputation, as it makes changes to the original dataset visible for the later phases of the process. Besides the identification of subsets based on domain knowledge and experience, for structured datasets we found isolation forests by Liu, Ting, and Zhou (2008) to be an efficient technique. It is important that all phases of our process are applicable without unnecessary technical barriers that slow down performance and negatively affect discussion and the overall investigative flow. We recommend to integrate our process into tool-supported extract, transform, and load (ETL) workflows and BI&A tools to make commonality plots seamlessly available at the push of a button. Equally important are version control for already used samples and extraction functionalities for the visualizations to facilitate documentation of results and findings.

Our process is most efficient for the analysis of structured datasets that remain unaltered throughout the process, because all measures, categorizations, and distribution estimates need to be calculated only once for the main dataset, regardless of the number of subsets. On the other hand, performance of our process is very limited when applied to constantly changing datasets such as real-time data. Time series data is a sequence of data points that are indexed chronologically by time. Chronological order is not accounted for in our visualization technique, making the process very inefficient, if not meaningless, when applied to time series data. Finally, the overall efficiency of commonality plots is limited by the calibration of the underlying measures and by the customization of the visualizations.

4.5 Further research agenda

Our results and findings show how our process model and visualization framework can be used efficiently for visual analytics, storytelling, and decision-making. It is of utmost importance that the whole process and its techniques are transparent, replicable, user-friendly, and without unnecessary technical barriers. These are the essential drivers for interactive visual analytics, profound storytelling, and decision support. For greater adoption, further research is needed to remove technical barriers to enable seamless application of our process.

Our application example comes from practice, namely auditing. But our process and visualization framework are not restricted to auditing and can be applied to many domains where structured data is available. For example, during user acceptance testing of our visualization framework, we collaborated with a physician to analyze an open source dataset available for building disease prediction models and Healthcare systems (Patil, 2020). There are numerous use cases for theoretical research as well. For example, datasets for meta-analysis of ISR journals by Palvia, Daneshvar Kakhki, Ghoshal, Uppala, and Wang (2015) who study ISR trends, a literature database for COVID-19 publications by Butler-Henderson, Crawford, Rudolph, Lalani, and Sabu (2020) who ask "Can we curate the first six months of published literature to support future researchers?", and a general overview of representative sampling in empirical Software Engineering research by Baltes and Ralph (2022) who propose "extensive guidelines for improving the conduct, presentation, and evaluation of sampling". The first and second examples are built around structural datasets that allow instant application of our visualization framework. For the third example on representative sampling, our visualization framework can be used to evaluate sample representativeness, i.e., a representative sample is not highlighted as unusual by commonality plots.

We encourage researchers and practitioners to integrate our visualization method into existing analytics workflows and apply and develop our process model in their specific domains. Consequently, further research can focus on general calibration rules for specific use cases and specific data to derive best practices. Further research can then in turn evaluate best practices in terms of robustness and efficiency. This includes an investigation of color themes and their impact on interaction, interpretability, and overall efficiency.

It has to start somewhere, it has to start sometime. What better place than here, what better time than now?

Zack de la Rocha

5 Closing Thoughts

5.1 Reflective discussion and outlook

For decades, the Information Systems (IS) discipline "has been thinking and researching questions at the intersection of technology, data, business, and society" (Agarwal & Dhar, 2014). Accordingly, the research areas covered in this dissertation are diverse and include Accounting, Finance, and ISR. From a chronological point of view, the course of research starts with very domain-specific problems and becomes more process-oriented towards the more recent phase. From a role-oriented point of view, in retrospect, not only the research areas are diverse. For our research in Accounting, my contributions come in the role of a domain expert with regard to IFRS 9 requirements for impairment and the implementation of the ECL model. For our research in Finance, my contributions come in the role of a data analyst with regard to the application of machine learning techniques. In more recent research projects, I resemble both roles as we derive and evaluate an efficient interactive process model and visualization framework driven by data analysts and domain experts to enable value creating collaboration in interdisciplinary teams. The underlying research design is process-oriented and aims for a general applicability of methods to create value and meaning. Ultimately, this best reflects the overall motivation for this dissertation to find new and innovative data-driven applications that are practical, relevant, and meaningful.

Against this background, I use a homogeneous structure to summarize the different research projects in order to highlight applicability and practical relevance. I consistently start with theoretical background, our research questions, and the overall contribution and relevance. Then, I summarize the underlying methodology and list all necessary building blocks for general application that is demonstrated in an applicability check. Finally, implications, recommendations, and limitations are discussed from a theoretical and practical perspective. Because our recent research already is process-oriented and includes an applicability check, its summary required less restructuring than that of previous research projects. Instead of restructuring, the focus is placed on an extension of the applicability check and the description of the visual setup. The homogeneous structure enables comparability and allows for the identification of overall limitations. It turns out that the general applicability of our approaches is mainly limited by numerous assumptions and technical barriers. This is especially true for the application of ANNs, where results are affected by the specification of hyperparameters and the general architecture of ANNs, see Chapter 3. Although our visualization framework described in Chapter 4 is more matured in terms of general applicability, it requires heavy customization. With regard to our step-by-step process that can be applied to prove that rating changes are adequate criteria for SICR determination, the effort required to perform the described steps is the technical barrier that limits applicability, see Chapter 2. Regardless the domain, the technical qualifications required to overcome the different barriers of our applications are rare. In summary, our data-driven processes and applications are not one-size-fits-all, so general applicability is limited by the requirement of certain technical skills necessary for configuration and customization.

With regard to our process model and the embedded visualization framework described in Chapter 4, we conclude that it is of utmost importance that its application is transparent, replicable, user-friendly, and without unnecessary technical barriers. This finding also applies to the applications resulting from our other research projects. The less complicated overall applicability, the higher the number of potential users. The more users, the higher the amount of valuable feedback. The more feedback, the more opportunities and ideas for relevant research and meaningful development. Thus, the reduction of unnecessary complexity and the simplification of accessibility imply increased relevance of our research and resulting applications. With regard to overall accessibility, the provision of web applications or low-code ETL workflows can be very helpful. With regard to the reduction of complexity it is necessary to make configuration and customization optional rather than mandatory. Both easy to say, but very difficult to implement, leaving many possibilities for further research. Accordingly, an important area for further research is the analysis of all underlying assumptions to understand their effect on results to evaluate their relevance for the overall application. Analyses can be performed both for each assumption individually and for their interdependencies. On this basis, best practices for various domains and areas for application can be derived. Equally important is the scientific evaluation of these best practices.

Finally, it is important to improve the dialogue between practitioners and the IS discipline in general to ensure that ISR is relevant and impactful, see Te'eni, Seidel, and Brocke (2017). This is also emphasized by Rosemann and Vessey (2008) who find that "without research outcomes relevant to practice, the very existence of a research discipline could be questioned because the discipline could well lack impact beyond its own (academic) community". In addition to that, it is important that research is conducted in a rigorous way to ensure high quality and credibility. The IS academic community often fears a gap between rigorous and relevant research, so Rosemann and Vessey (2008) propose to conduct applicability checks that address relevance, but leave intact a rigorous research process. While we provide applicability checks of our own research, there is always the chance for bias. Thus, it is important that reviewers and practitioners are enabled to apply and adapt our approaches in their specific domains. We lay out the necessary process steps and building blocks of our approaches, but that may not be enough. A next step can be to provide replicable examples in well-known blog or notebook formats for hands-on experience, see, e.g., Kluyver et al. (2016) who discuss »Jupyter Notebooks« as reproducible computational workflows for academic publishing. Naturally, this will not be possible for every domain and research project, especially when sensitive data is involved. But for data-driven research and processes, reproducible applicability checks can open the door for reviewers and practitioners. If the underlying data is restricted for publication, it may be worthwhile to provide examples based on publicly available datasets. Another advantage of open access data is better user familiarity, which increases transparency and credibility of the research underlying the application.

5.2 Conclusion

This dissertation contributes to both theory and practice with regard to loan loss provisioning, stress testing, and visual analytics. More precisely, our findings and implications from research in Accounting contribute to transparency with regard to the recognition and disclosure of loan loss provisions in financial statements. Our results related to stress testing contribute to the overall quality and compliance of stress test executions. Finally, our generalized process model and visualization framework contribute to interactive visual analytics, storytelling, and decision support in the audit and many other domains where structured data is available. All research projects covered in this dissertation are based on problems identified in practice and thus ensure practical relevance of our research, which is additionally emphasized by applicability checks. To ensure academic rigor, our research has undergone thorough peer-review processes prior to publication in academic journals and conference proceedings.

I conclude this dissertation as it started with a quote from Agarwal and Dhar (2014) that has motivated me throughout the course of research: "This is an exciting time to be an IS researcher and to think beyond IS to science in general."

References

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25*(3), 443–448.

Alpar, P., & Schulz, M. (2016). Self-service business intelligence. *Business & Information Systems Engineering*, *58*(2), 151–155.

Andersson, A., & Vanini, P. (2010). Credit migration risk modeling. *Journal of Credit Risk*, *6*(1), 3–30.

Aragonés, J. R., Blanco, C., & Dowd, K. (2001). Incorporating stress tests into market risk modeling. *Derivatives Quarterly*, *7*(3), 44–50.

Asteriou, D., & Hall, S. G. (2021). *Applied econometrics*. Bloomsbury Publishing.

Baltes, S., & Ralph, P. (2022). Sampling in software engineering research: A critical review and guidelines. *Empirical Software Engineering*, *27*(4), 1–31.

Bedard, J. C., & Biggs, S. F. (1991). Pattern recognition, hypotheses generation, and auditor performance in an analytical task. *Accounting Review*, 622–642.

Bluhm, C., Overbeck, L., & Wagner, C. (2016). *Introduction to credit risk modeling*. Chapman and Hall/CRC.

Bolt, R., & Tregidga, H. (2023). Methodological insights "Materiality is ...": Sensemaking and sensegiving through storytelling. *Accounting, Auditing & Accountability Journal*, *36*(1), 403–427.

Boritz, J. E., Kochetova, N. V., Robinson, L. A., & Wong, C. (2020). Auditors' and specialists' views about the use of specialists during an audit.

Behavioral Research in Accounting, *32*(2), 15–40.
Boslaugh, S. (2012). *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media.
Bosse, M., Stege, N., & Hita Hochgesand, M. (2017a). Beurteilung der signifikanten Verschlechterung der Kreditqualität nach IFRS 9: Voraussetzungen für die Verwendung von Ratings und Lifetime-PD. *Die Wirtschaftsprüfung*, *70*(1), 5–12.
Bosse, M., Stege, N., & Hita Hochgesand, M. (2017b). Stufenzuordnung nach IFRS 9: Nachweis zur Verwendung von Ratings als geeignetes Kriterium zur Beurteilung der signifikanten Verschlechterung der Kreditqualität. *Die Wirtschaftsprüfung*, *70*(8), 437–447.
Bostock, M. (2018). *Zoomable circle packing*. Retrieved from https://observablehq.com/@d3/zoomable-circle-packing
Bozanic, Z., Kraft, P., & Tillet, A. (2022). Qualitative disclosure and credit analysts' soft rating adjustments. *European Accounting Review*, 1–29. Retrieved from https://doi.org/10.1080/09638180.2022.2038227
Bresciani, S., & Eppler, M. J. (2009). The benefits of synchronous collaborative information visualization: Evidence from an experimental evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *15*(6), 1073–1080.
Brixner, J., Schaber, M., & Bosse, M. (2013). Der Exposure Draft ED/2013/3" Expected Credit Losses": Überblick über die neuen Wertminderungsvorschriften und deren Implikationen auf den Bilanzansatz und die Erfolgswirkung. *KoR*, *2013*(5), 221-235.
Butler-Henderson, K., Crawford, J., Rudolph, J., Lalani, K., & Sabu, K. (2020). COVID-19 in higher education literature database (CHELD V1): An open access systematic literature review database with coding rules. *Journal of Applied Learning & Teaching*, *3*(2), 1–6.
Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly*, *36*(4), 1165–1188.
D'Agostino, R., & Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of b2 and b1. *Biometrika*, *60*(3),

613–622.
Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press.
Dykes, B. (2020). *Effective data storytelling: How to drive change with data, narrative and visuals*. John Wiley & Sons.
Eilers, D., Köpp, C., Gleue, C., & Breitner, M. H. (2017). It's not a bug, it's a feature: How visual model evaluation can help to incorporate human domain knowledge in data science. *Proceedings International Conference on Information Systems*, 17 p.
Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., & Andrews, C. (2014). The human is the loop: New directions for visual analytics. *Journal of Intelligent Information Systems*, *43*(3), 411–435.
Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica: Journal of the Econometric Society*, 251–276.
European Banking Authority. (2018). Final report on guidelines on institutions' stress testing. *European Banking Authority EBA-GL-2018-04*, 130 p.
European Systemic Risk Board. (2016). Adverse macro-financial scenario for the EBA 2016 EU-wide bank stress testing exercise. *European Systemic Risk Board*, 15 p.
EY Global CRS. (2018). Applying IFRS - Impairment of financial instruments under IFRS 9. *Ernst & Young Global Limited*, 165 p.
Gleue, C., Eilers, D., von Mettenheim, H.-J., & Breitner, M. H. (2019). Decision support for the automotive industry: Forecasting residual values using artificial neural networks. *Business & Information Systems Engineering*, *61*, 385–397.
Gupta, A. (2017). Editorial thoughts: What and how ISR publishes. *Information Systems Research*, *28*(1), 1–4.
Hagen, J. A. (2021). Saving IS from another troubled marriage: Diagnosing and bridging the gap between data science and other business functions. *Proceedings Pacific Asia Conference on Information Systems*, 15 p.
Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM*, *53*(6), 59–67.

Hirtle, B., Kovner, A., Vickery, J., & Bhanot, M. (2016). Assessing financial stability: The capital and loss assessment under stress scenarios (CLASS) model. *Journal of Banking & Finance*, *69*(1), 35–55.
Hoadley, B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*, *64*(325), 216–229.
Hofmann, H., Wickham, H., & Kafadar, K. (2017). Letter-Value Plots: Boxplots for large data. *Journal of Computational and Graphical Statistics*, *26*(3), 469-477.
Hossain, M. S., Andrews, C., Ramakrishnan, N., & North, C. (2011). Helping intelligence analysts make connections. *Proceedings AAAI Conference on Artificial Intelligence*, 10 p.
ITG. (2015a). Meeting Summary – 11 December 2015. *IFRS Transition Resource Group for Impairment of Financial Instruments, Meeting Summary*. Retrieved from `https://www.ifrs.org/groups/transition-resource-group-for-impairment-of-financial-instruments/#meetings`
ITG. (2015b). Meeting Summary – 16 September 2015. *IFRS Transition Resource Group for Impairment of Financial Instruments, Meeting Summary*. Retrieved from `https://www.ifrs.org/groups/transition-resource-group-for-impairment-of-financial-instruments/#meetings`
Janardan, K. (1973). Chance mechanisms for multivariate hypergeometric models. *Sankhyā: The Indian Journal of Statistics, Series A*, *35*(4), 465–478.
Kaggle. (2012). *Titanic - Machine learning from disaster.* Retrieved from `https://www.kaggle.com/competitions/titanic`
Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... Corlay, S. (2016). Jupyter Notebooks - A publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS Press.
Knechel, W. R. (2016). Audit quality and regulation. *International Journal of Auditing*, *20*(3), 215–223.
Kohn, D., & Liang, N. (2019). Understanding the effects of the U.S. stress tests. *Federal Reserve System Conference on Stress Testing: A Discussion and*

Review, 30 p.
Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. *Journal of Emerging Technologies in Accounting*, *14*(1), 115–122.
Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83–91.
Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. *Risks*, *7*(1), 1–29. Retrieved from https://doi.org/10.3390/risks7010029
Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *Proceedings International Conference on Data Mining*, 10 p.
Lowry, R. (2017). *VassarStats: Website for statistical computation.* Retrieved from http://vassarstats.net/
Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world. *European Journal of Information Systems*, *22*(4), 381–386.
Maitlis, S., & Christianson, M. (2014). Sensemaking in organizations: Taking stock and moving forward. *Academy of Management Annals*, *8*(1), 57–125.
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.
Palvia, P., Daneshvar Kakhki, M., Ghoshal, T., Uppala, V., & Wang, W. (2015). Methodological and topic trends in Information Systems Research: A meta-analysis of IS journals. *Communications of the Association for Information Systems*, *37*(1), 30.
Patil, P. (2020). *Disease symptom prediction.* Retrieved from https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset
Qi, M., & Zhang, G. P. (2008). Trend time–series modeling and forecasting with neural networks. *IEEE Transactions on Neural Networks*, *19*(5), 808–816.
Rikhardsson, P., & Yigitbasioglu, O. (2018). Business intelligence & analytics in management accounting research: Status and future focus. *Interna-*

tional Journal of Accounting Information Systems, *29*, 37–58.
Rodríguez-Fdez, I., Canosa, A., Mucientes, M., & Bugarín, A. (2015). STAC: A web platform for the comparison of algorithms using statistical tests. *Proceedings International Conference on Fuzzy Systems*. Retrieved from http://tec.citius.usc.es/stac/
Rosemann, M., & Vessey, I. (2008). Toward improving the relevance of Information Systems Research to practice: The role of applicability checks. *Management Information Systems Quarterly*, *32*(1), 1–22.
Saghafi, A., Wand, Y., & Parsons, J. (2022). Skipping class: Improving human-driven data exploration and querying through instances. *European Journal of Information Systems*, *31*(4), 463–491.
Schneider, G. P., Dai, J., Janvrin, D. J., Ajayi, K., & Raschke, R. L. (2015). Infer, predict, and assure: Accounting opportunities in data analytics. *Accounting Horizons*, *29*(3), 719–742.
Sekar, M. (2022). Storytelling in Auditing. In *Machine Learning for Auditors: Automating Fraud Investigations Through Artificial Intelligence* (pp. 181–183). Springer.
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611.
Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, *23*(4), 433–441.
Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.
Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow University Mathematics Bulletin*, *2*(2), 3–14.
Stege, N., & Breitner, M. H. (2020). Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists. *Proceedings Internationale Wirtschaftsinformatik Tagung*, 16 p.
Stege, N., & Breitner, M. H. (2023). Identify, visualize, discuss, and decide: A collaborative framework to explore unusual subset-dataset relationships.

Unpublished.
Stege, N., Wegener, C., Basse, T., & Kunze, F. (2017). Mapping interest rate projections using neural networks under cointegration: An application from stress testing approaches. *Proceedings International Conference on Internet of Things and Machine Learning*, 5 p.
Stege, N., Wegener, C., Basse, T., & Kunze, F. (2021). Mapping swap rate projections on bond yields considering cointegration: An example for the use of neural networks in stress testing exercises. *Annals of Operations Research*, *297*(1), 309–321.
Sun, G.-D., Wu, Y.-C., Liang, R.-H., & Liu, S.-X. (2013). A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, *28*(5), 852–867.
Te'eni, D., Seidel, S., & Brocke, J. v. (2017). Stimulating dialog between Information Systems Research and practice. *European Journal of Information Systems*, *26*(6), 541–545.
Thomas, J. J. (2005). *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society.
Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. CRC Press.
Wegener, C., Basse, T., Sibbertsen, P., & Nguyen, D. K. (2019). Liquidity risk and the covered bond market in times of crisis: Empirical evidence from Germany. *Annals of Operations Research*, *282*, 407–426.
Wegener, C., von Spreckelsen, C., Basse, T., & von Mettenheim, H.-J. (2016). Forecasting government bond yields with neural networks considering cointegration. *Journal of Forecasting*, *35*(1), 86–92.
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

A
Appendix

A.1

Beurteilung der signifikanten Verschlechterung der Kreditqualität nach IFRS 9: Voraussetzungen für die Verwendung von Ratings und Lifetime-PD

Citation. Bosse, M., Stege, N., & Hita Hochgesand, M. (2017). Beurteilung der signifikanten Verschlechterung der Kreditqualität nach IFRS 9: Voraussetzungen für die Verwendung von Ratings und Lifetime-PD. Die Wirtschaftsprüfung, 70(1), 5–12.

Link. https://www.idw.de/IDW-Verlag/04-WPg/WPg/Jahresregister/Downloads/Down-Wpg-Jahresregister-2017.pdf

Abstract. Eine zentrale Herausforderung bei der Umsetzung einer IFRS-konformen Stufenzuordnung ist die Auswahl geeigneter Beurteilungskriterien. In den derzeitigen Umsetzungsprojekten zeichnet sich dafür vor allem die Verwendung von kumulierten Ausfallwahrscheinlichkeiten (Lifetime-PD) und Ratings ab. Während Ratings ein vertrautes Maß zur Beurteilung der Kreditqualität sind, handelt es sich bei Lifetime-PD – zumindest im bilanziellen Kontext – um eine Neuerung. Vor diesem Hintergrund wird in diesem Beitrag herausgearbeitet, welche methodischen Unterschiede zwischen Lifetime-PD und Ratings bestehen und unter welchen Voraussetzungen diese als Beurteilungskriterien für eine signifikante Verschlechterung der Kreditqualität verwendet werden können.

Keywords. IFRS 9, Expected-Loss-Modell, Stufenzuordnung, Lifetime-PD, Rating.

A.2

Stufenzuordnung nach IFRS 9: Nachweis zur Verwendung von Ratings als geeignetes Beurteilungskriterium der signifikanten Verschlechterung der Kreditqualität

Citation. Bosse, M., Stege, N., & Hita Hochgesand, M. (2017). Stufenzuordnung nach IFRS 9: Nachweis zur Verwendung von Ratings als geeignetes Kriterium zur Beurteilung der signifikanten Verschlechterung der Kreditqualität. Die Wirtschaftsprüfung, 70(8), 437–447.

Link. https://www.idw.de/IDW-Verlag/04-WPg/WPg/Jahresregister/Downloads/Down-Wpg-Jahresregister-2017.pdf

Abstract. Die Umsetzung der neuen Wertminderungsvorschriften nach IFRS 9 befindet sich in einem Spannungsfeld zwischen der Berücksichtigung der komplexen Anforderungen des Standards auf der einen Seite und der Transparenz und Kommunizierbarkeit im Sinne der Vermittlung entscheidungsnützlicher Informationen auf der anderen Seite. Eine Möglichkeit zur Reduzierung der Komplexität und gleichzeitigen Erhöhung von Transparenz und Kommunizierbarkeit kann in der Verwendung von Ratings als Beurteilungskriterium für die Stufenzuordnung bestehen. Dies setzt jedoch die Erbringung eines geeigneten Nachweises voraus, der sich in der Praxis bislang noch nicht etabliert hat und in diesem Beitrag herausgearbeitet wird.

Keywords. IFRS 9, Stufenzuordnung, Rating, Beurteilungskriterium, Makroökonomische Szenarien.

A.3

Mapping interest rate projections using neural networks under cointegration: An application from stress testing approaches

Citation. Stege, N., Wegener, C., Basse, T., & Kunze, F. (2017). Mapping interest rate projections using neural networks under cointegration: An application from stress testing approaches. Proceedings International Conference on Internet of Things and Machine Learning, 5 p.

Link. https://dl.acm.org/doi/10.1145/3109761.3109774

Abstract. This paper discusses the application of techniques of business analytics in the banking industry examining stress tests in the context of financial risk management. We focus on the use of neural networks in combination with techniques of cointegration analysis to map swap rate projections derived from given scenarios (e.g., a certain stress scenario from the EBA/ECB 2016 EU-wide stress test) on other relevant interest rates in order to ensure that contingent projections for these time series are produced and used in the process of stress testing.

Keywords. Risk Management, Net Interest Rate Income, Mapping Interest Rate Projections, Cointegration, Artificial Neural Networks.

A.4

Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists

Citation. Stege, N., & Breitner, M. H. (2020). Hybrid intelligence with commonality plots: A first aid kit for domain experts and a translation device for data scientists. Proceedings Internationale Wirtschaftsinformatik Tagung, 16 p.

Link. https://doi.org/10.30844/wi_2020_c7-stege

Abstract. There is a large gap between domain experts capable to identify business needs and data scientists who use insight producing algorithms, but often fail to connect these to the bigger picture. A major challenge for companies and organizations is to integrate practical data science into existing teams and workflows. We are driven by the assumption that efficient data science requires cross-disciplinary teams able to communicate. We present a methodology that enables domain experts and data scientists to analyze and discuss findings and implications together. Motivated by a typical problem from auditing we introduce a visualization method that helps to detect unusual data in a subset and highlights potential areas for investigation. The method is a first aid kit applicable regardless whether unusual samples were detected by manual selection of domain experts or by algorithms applied by data scientists. An applicability check shows how the visualizations facilitate collaboration of both parties.

Keywords. Commonality Plots, Domain Knowledge, Hybrid Intelligence, Visualization, Data Science.

A.5

Mapping swap rate projections on bond yields considering cointegration: An example for the use of neural networks in stress testing exercises

Citation. Stege, N., Wegener, C., Basse, T., & Kunze, F. (2021). Mapping swap rate projections on bond yields considering cointegration: An example for the use of neural networks in stress testing exercises. Annals of Operations Research, 297(1), 309–321.

Link. https://doi.org/10.1007/s10479-020-03762-x

Abstract. This paper discusses the application of techniques of business analytics in the banking industry examining stress tests in the context of financial risk management. We focus on the use of neural networks in combination with techniques of cointegration analysis to map swap rate projections derived from given scenarios (e.g., a certain stress scenario from the EBA/ECB 2016 EU-wide stress test) on other relevant interest rates in order to ensure that contingent projections for these time series are produced and used in the process of stress testing.

Keywords. Risk Management, Net Interest Rate Income, Modeling Interest Rates, Cointegration, Artificial Neural Networks.

A.6

Identify, visualize, discuss, and decide: A collaborative framework to explore unusual subset-dataset relationships

Note. The paper was submitted to the Data & Knowledge Engineering (DKE) journal in April 2023 and is still under review at the time of writing this dissertation.

Abstract. Domain experts are driven by business needs, while data analysts develop and use various algorithms, methods, and tools, but often without domain knowledge. A major challenge for companies and organizations is to integrate data analytics in business processes and workflows. We deduce an efficient interactive process and visualization framework to enable value creating collaboration in inter- and cross-disciplinary teams. Domain experts and data analysts are both empowered to jointly analyze and discuss results and come to well-founded insights and implications. Inspired by a typical auditing problem, we develop and apply a visualization framework to single out unusual data in general subsets for potential further investigation. Our framework is applicable to both unusual data detected manually by domain experts or by algorithms applied by data analysts. An application example shows typical interaction, collaboration, visualization, and decision support.

Keywords. Data Visualization, Visual Analytics, Commonality Plots, Subset-Dataset Relationships, Anomaly Explanation, Decision Support.