README - Do Banks Overestimate Green Credit Risk? Evidence from SME Lending

Authors: ***Patrick Zeitinger, Sandeep Singh, Robert Schulz, Florian Egli***

Affiliations:Professorship of Public Policy for the Green Transition (PPGT), TU Munich; International Finance Corporation, World Bank Group, Washington D.C.; ProCredit Holding AG, Frankfurt am Main

Contact: patrick.zeitinger@tum.de

Items marked [CONFIRM]: require author input before final submission.

Overview

This replication package reproduces all tables and figures in the working paper "Do Banks Overestimate Green Credit Risk? Evidence from SME Lending." The analysis uses a proprietary ten-year panel of SME loan accounts from ProCredit Group (11 countries, 2015–2024, 43,000+ rated firms), shared with the World Bank reproducibility team under a Non-Disclosure Agreement (NDA).

To reproduce all results:place the NDA-shared data file in ` Data/` , update the single path on line 23 of ` main.do` , and execute ` main.do` . The package runs two do-files in sequence and writes all outputs to ` Outputs/` . Estimated runtime: **approximately 20 minutes on a standard desktop.

Data Availability and Provenance Statements

☑ No data can be made publicly available.**

All data are proprietary to ProCredit Group and are subject to an NDA between the authors and ProCredit Holding AG, Frankfurt am Main, Germany. The World Bank reproducibility team has been granted temporary access to the starting-point dataset listed below solely for verification purposes, under a separate NDA with ProCredit.

Confidential data used in this paper and not provided as part of the public replication package will be preserved for 5 years after publication, in accordance with journal policies.

Statement about Rights

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data. *(Does not apply — data is proprietary and may not be redistributed. The World Bank reproducibility team has separate NDA access for verification only.)*

Summary of Availability

Data file	Source	Notes	Provided to WB reviewer
`Loans_merged_2015.dta`	ProCredit Group internal database	Confidential; NDA required	Yes (under NDA)

Details on Each Data Source

Dataset — Contract-level panel (sole input file)

The package uses a single NDA-shared contract-level dataset covering all SME loan accounts at ProCredit Group from 2015 to 2024. `Code/01_cleaning.do` transforms this file into the client-year panel used for regression analysis, and `Code/02_analysis.do` Section 11 uses it directly for the loan-level portfolio figures.

- **Filename**: `Loans_merged_2015.dta`

- **Format**: Stata binary (.dta), approx. **1.8 GB**

- **Coverage**: 783,188 contract-year observations; 155,357 unique SME clients; 12 countries (11 after Germany drop); 2015–2024

- **Source**: ProCredit Group, proprietary internal credit database

- **Access**: Researchers interested in access should contact ProCredit Holding AG: Robert Schulz (robert.schulz@procredit-group.com) or the legal department (PCH.Legal@procredit-group.com). Access requires execution of a Data Use Agreement / NDA. It may take several weeks to arrange access.

- **Access year**: Data accessed continuously by the authors 2023–2026 under a data use agreement between Patrick Zeitinger and ProCredit Group.

- Key variables: Contract identifier, client identifier, calendar year, country (Bank variable), sector (ISIC), internal risk classification (RC_EOY, RC_initial), binary default indicators (default_nextyear, default_thisyear), green loan type indicators (EEtype, REtype, GRtype), outstanding principal (EUR), eco-classifier codes (ProCredit green taxonomy sub-subcategories), financial statement ratios (CurrentRatio, RoAA, DebtEBITDA).

- **Data dictionary**: Variable labels are embedded in the Stata file. A codebook is available from the authors upon request.

- **License**: Proprietary. Redistribution is not permitted.

- **Provenance note**: Steps 1–4 of the data pipeline (how raw ProCredit database extracts were cleaned, merged, and aggregated to produce `Loans_merged_2015.dta`) are documented in `Code/00_data_documentation.do`. That file cannot be executed without direct access to the ProCredit database. `Loans_merged_2015.dta` serves as the reproducible starting point in the sense of the WB FAQ on starting points for reproducibility packages.

- Data Citation

*ProCredit Group. (2024). *SME Loan Portfolio Panel 2015–2024* [Proprietary dataset]. ProCredit Holding AG, Frankfurt am Main, Germany. Accessed by the authors under a data use agreement between Patrick Zeitinger and ProCredit Group, 2023–2026.*

Computational Requirements

Software Requirements

- ☑ The replication package contains a program (`main.do`) that installs all Stata dependencies and sets up the necessary directory structure.

- StataNow 19.5 BE-Basic Edition**, running on Windows 11 Education

  - `estout` version 3.31 (26 Apr 2022, Ben Jann) — table export to RTF/Excel

  - `cem` — Coarsened Exact Matching (Iacus, King, Porro; version not tagged in SSC package)

  - `psmatch2` version 4.0.12 (30 Jan 2016, Leuven & Sianesi) — Propensity Score Matching

- `reghdfe` — high-dimensional fixed effects regression
- `ftools` — dependency for reghdfe
- `winsor` — winsorisation of continuous variables
- All packages are installed automatically by `main.do` on first run via SSC.

Note on Stata edition: The analysis was run on **BE-Basic Edition** (max 2,047 variables; 1 processor; no multiprocessor). Replication on SE or MP editions should produce identical results.

Note on portability: The original analysis was conducted on Windows 11 Education. The code uses Stata globals for file paths (set via `main.do`) and should be portable to macOS/Linux, but has not been tested outside of Windows.

Controlled Randomness

- ☑ Random seed is set at the top of the PSM matching block in `Code/02_analysis.do` (`set seed 12345`), immediately before the `psmatch2` step.

- The same seed (`set seed 12345`) is also set in `main.do` Step 6 so the package is self-contained.

Memory, Runtime, and Storage Requirements

Summary

Approximate time to reproduce all analyses on a standard 2024 desktop machine:

- ☐ < 10 minutes

- ☑ 10–60 minutes (approximately 20 minutes on a standard desktop)

- ☐ 2–8 hours

- ☐ > 8 hours

Approximate storage required (including input data):

- ☐ < 250 MB

- ☐ 250 MB – 2 GB

- ☑ 2 – 10 GB (input file ~1.8 GB + intermediate panel files ~1.5 GB + outputs ~4 MB ≈ **3.5 GB total**)

- ☐ > 10 GB

Hardware Used by Authors

Code was last run on: Windows 11 Education, 16 GB RAM, AMD Ryzen 5 PRO 8540U w/ Radeon 740M Graphics, 3.20 GHz

Instructions to Replicators

1. Ensure you have received the NDA-shared data file from the World Bank reproducibility team (or contact the authors at patrick.zeitinger@tum.de).

2. Place the file in the `Data/` subfolder:

Data/

Loans_merged_2015.dta

3. Install **Stata 19 or later** (BE-Basic Edition or higher).

4. Open `main.do` in Stata.

5. On **line 23**, update the `global project` path to your local package root:

```stata
global project "C:\YOUR\PATH\HERE"
```

6. Run `main.do`. All packages install automatically on first run; all outputs are written to `Outputs/`.

Note on `Code/00_data_documentation.do`: This file documents Steps 1–4 of the data pipeline (raw ProCredit database extracts → `Loans_merged_2015.dta`). It is included for transparency but is **not** called by `main.do` and cannot be executed without raw ProCredit database access.

List of Tables and Programs

The provided code reproduces:

- ☑ All tables and figures in the paper

- ☐ All in-text numbers **[CONFIRM — verify whether statistics cited in text body are explicitly output by code]**

Figure/Table	Description	Output file	Script	Note
Table 1	Summary statistics	`Table1_SummaryStats.xlsx`	`02_analysis.do`	Section 4
Table 2	Default rates by risk rating	`Table2_DefaultByRating.xlsx`	`02_analysis.do`	Section 4
Table 3	Main Cox + FE-LPM results	`Table_Main_Cox.rtf`, `Table_Main_FE.rtf`	`02_analysis.do`	Section 5
Figure 1	Portfolio split: green vs non-green exposure	`Figure1_final.png`	`02_analysis.do`	Section 11
Figure 2	Green composition by type	`Figure2_final.png`	`02_analysis.do`	Section 11
Annex — Matching balance	CEM/PSM covariate balance	`TableAnnex_Balance.xlsx`, `Love-Plot.gph`	`02_analysis.do`	Section 3
Annex Table A2	Sector-level green firm share	`TableA2_sector.xlsx`	`02_analysis.do`	Section 4
Annex Table A4	Sector-level green loan share	`TableA4_sector_green_share.xlsx`	`02_analysis.do`	Section 11
Annex — Intensity	Dose-response (green share)	`TableAnnex_Intensity.rtf`	`02_analysis.do`	Section 6
Annex — Non-sticky	Non-sticky treatment robustness	`TableAnnex_Nonsticky.rtf`	`02_analysis.do`	Section 7
Annex — Financials	Financial ratios as alternative controls	`TableAnnex_Financials.rtf`	`02_analysis.do`	Section 8
Annex — Time split	Pre/post-2020 temporal stability	`TableAnnex_TimeSplit.rtf`	`02_analysis.do`	Section 9
Annex — Country subsample	Country-level robustness (11 countries)	`TableAnnex_CountrySubsample.rtf`	`02_analysis.do`	Section 10

| Annex Figure A1 | EE sub-subcategory composition | `FigureA1_vol.png` | `02_analysis.do` | Section 11 |

| Annex Figure A2 | RE sub-subcategory composition | `FigureA2_vol.png` | `02_analysis.do` | Section 11 |

| Annex Figure A3 | Green Measures sub-subcategory | `FigureA3_vol.png` | `02_analysis.do` | Section 11 |

Description of Programs / Code

- **`main.do`** (root): Sets the single top-level directory path; derives all sub-paths; creates output folder structure; installs required Stata packages; sets random seed; calls `01_cleaning.do` and `02_analysis.do` in sequence.

- **`Code/00_data_documentation.do`** — *Included for documentation only; not called by `main.do`.*

  Documents Steps 1–4: how ProCredit raw database extracts (loan contracts, risk classifications, default events) were filtered, reformatted, and merged to produce `Loans_merged_2015.dta`. Cannot be executed without raw ProCredit database access.

- **`Code/01_cleaning.do`**: Data cleaning and panel construction. Opens `Loans_merged_2015.dta` and performs: (1) recoding of string financial ratio variables; (2) deduplication from contract-year to client-year level; (3) variable construction (green indicators, risk rating fill, loan controls); (4) two-pass right-censoring at firm-level default event; (5) sticky green variable construction. Saves `MasterData_censored_nextyear.dta` as input to `02_analysis.do`.

- **`Code/02_analysis.do`**: Main analysis script. Opens `MasterData_censored_nextyear.dta`. Performs: (1) CEM matching on country, sector, and panel exposure; (2) PSM on firm-level covariates; (3) matching balance diagnostics; (4) summary statistics; (5) Cox Proportional Hazard and Two-Way FE LPM regressions (main tables); (6) all robustness checks (Sections 6–10); (7) loan-level portfolio figures and annex tables (Section 11, using `Loans_merged_2015.dta` directly).

License for Code

Folder Structure

```
Reproducibility Package/      ← package root
│
├── main.do                ← Run this (change path on line 23 only)
├── README.md              ← This file
├── LICENSE.txt
│
├── Code/
│   ├── 00_data_documentation.do ← Documents Steps 1–4 (not run)
│   ├── 01_cleaning.do       ← Panel construction + right-censoring
│   ├── 02_analysis.do       ← All regressions + figures
│   └── ado/                ← Stata packages (auto-populated by main.do)
│
├── Data/
│   ├── Loans_merged_2015.dta                     [NDA — provided]
│   ├── MasterData_Clean_panel2015-2024_without adjustment.dta   [generated by 01_cleaning.do]
│   ├── MasterData_censored_nextyear.dta                [generated by 01_cleaning.do]
│   └── MasterData2015-2024_final.dta                   [generated by 02_analysis.do]
│
└── Outputs/                ← Created automatically by main.do
    ├── Tables/
    │   ├── Main/
    │   │   ├── Table1_SummaryStats.xlsx
    │   │   ├── Table2_DefaultByRating.xlsx
    │   │   ├── Table_Main_Cox.rtf
```

```
│  │  └── Table_Main_FE.rtf
│  └── Annex/
│      ├── TableAnnex_Balance.xlsx
│      ├── TableAnnex_Intensity.rtf
│      ├── TableAnnex_Nonsticky.rtf
│      ├── TableAnnex_Financials.rtf
│      ├── TableAnnex_TimeSplit.rtf
│      ├── TableAnnex_CountrySubsample.rtf
│      ├── TableA2_sector.xlsx
│      └── TableA4_sector_green_share.xlsx
├── Figures/
│  ├── Main/
│  │  ├── Figure1_final.png
│  │  └── Figure2_final.png
│  └── Annex/
│      ├── FigureA1_vol.png
│      ├── FigureA2_vol.png
│      └── FigureA3_vol.png
├── Love-Plot.gph
└── Matching/
    ├── collapsed_client_averages_cem.dta
    └── collapsed_client_averages_cem_psm.dta
```

References

*ProCredit Group. (2024). *SME Loan Portfolio Panel 2015–2024* [Proprietary dataset]. ProCredit Holding AG, Frankfurt am Main, Germany. Accessed by the authors under a data use agreement between Patrick Zeitinger and ProCredit Group, 2023–2026.*

Acknowledgements

This paper was edited in part with the assistance of Claude (Anthropic) for writing support and Grammarly for grammar checking. All analytical decisions, code, and conclusions are the authors' own.