Failing Banks Project

Complete Data Dictionary

R Replication v10.0

October 2025

Contents

1 Introduction

This data dictionary provides comprehensive documentation of all datasets created and used in the Failing Banks R replication project. It includes:

- **11 key datasets** spanning 1865-2024
- **70+ variables** with detailed definitions
- **Data sources** and construction methods
- **Coverage statistics** and quality notes

1.1 Dataset Overview

Dataset	Observations	Variables	Size (MB)
combined-data.dta	2,870,000	45	685.0
temp_reg_data.dta	2,120,000	35	816.5
modern_banks_panel.dta	2,180,000	40	1877.1
historical_banks_panel.dta	690,000	42	432.0
cross_section_data.dta	428,000	28	111.7
coefplot_data.dta	88,500	18	6.7
outflows_receivership.dta	31,500	24	2.0
gdp_data.dta	160	3	0.05
gfd_cpi_data.dta	165	3	0.06
gfd_yields_data.dta	200	4	0.07
receivership_dataset_tmp.dta	31,500	20	2.0

Table 1: Overview of All Datasets

2 Core Panel Datasets

2.1 combined-data.dta

Description: Master panel dataset combining historical (1865-1958) and modern (1959-2024) bank-quarter observations.

Coverage: 2,870,000 observations across 160 years

Frequency: Quarterly (historical: annual converted to quarterly; modern: native quarterly)

2.1.1 Core Variables

Variable	Type	Definition
Bank Identifiers		
id_fdic_cert	Integer	FDIC certificate number (modern era)
charter	Integer	Bank charter number (historical era)
bank_name	String	Name of the bank
state	String	Two-letter state code
Time Variables		
year	Integer	Calendar year (1865-2024)
quarter	Integer	Calendar quarter (1-4)
quarter_number	Integer	Sequential quarter number since start
report_date	Date	Quarterly reporting date
era_group	String	Era classification: "Historical" or "Modern"
Balance Sheet Variables		
assets	Numeric	Total assets (nominal USD)
log_assets	Numeric	Natural log of total assets
deposits	Numeric	Total deposits (nominal USD)
loans	Numeric	Total loans and discounts (nominal USD)
liquid	Numeric	Liquid assets (cash + securities) (USD)
equity	Numeric	Total equity capital (USD)
Financial Ratios		
equity_ratio	Numeric	Equity / Assets (percent)
loan_ratio	Numeric	Loans / Assets (percent)
liquid_ratio	Numeric	Liquid / Assets (percent)
deposit_ratio	Numeric	Deposits / Assets (percent)
income_ratio	Numeric	Net income / Assets (percent)
Lagged Ratios		
L_equity_ratio	Numeric	Lagged equity ratio (t-1)
L_loan_ratio	Numeric	Lagged loan ratio (t-1)
L_liquid_ratio	Numeric	Lagged liquid ratio (t-1)

Variable	Type	Definition
L_log_assets	Numeric	Lagged log assets (t-1)
Failure Indicators		
failed_bank	Binary	1 if bank eventually fails, 0 otherwise
fail_day	Date	Date of bank failure (if applicable)
days_to_failure	Numeric	Days from observation to failure
quarters_to_failure	Numeric	Quarters from observation to failure
time_to_fail	Integer	Years to failure (negative: -10 to -1)
F1_failure	Binary	Fails within 1 quarter (forward-looking)
F3_failure	Binary	Fails within 3 quarters (forward-looking)
F5_failure	Binary	Fails within 5 quarters (forward-looking)
F6_failure	Binary	Fails within 6 quarters (forward-looking)
Bank Runs		
run	Binary	Deposit outflow ¿ 15% in quarter
run_is_missing	Binary	1 if run variable is missing
F1_failure_run	Binary	Fails within 1 quarter given run occurred
F3_failure_run	Binary	Fails within 3 quarters given run occurred
Bank Characteristics		
age	Integer	Years since bank founding
growth	Numeric	Asset growth rate (3-period)
growth_cat	Factor	Asset growth quintile (1-5)
size_group	String	"Small", "Medium", or "Large" by assets

Table 2: Variables in combined-data.dta

2.2 temp_reg_data.dta

Description: Regression-ready dataset with failure indicators and control variables for econometric analysis.

Coverage: 2,120,000 observations (subset of combined-data with complete cases)
Key Filters Applied:

- Dropped observations after bank failure (quarters_to_failure ¿ 0)
- Dropped if income_ratio missing and year ¿ 1941
- Dropped de novo banks (age ¡ 3 years)
- Dropped if missing lagged variables

Variables: Same as combined-data.dta, but with complete cases only
Purpose: Used in all regression analyses (logit, OLS, event studies)

3 Era-Specific Datasets

3.1 historical_banks_panel.dta

Description: Historical era bank panel (1865-1958)
Source: OCC historical call reports, annual reporting
Coverage: 690,000 bank-year observations
Banks: 7,500 unique national banks
Special Features:

- Pre-FDIC era (before 1934)
- Annual reporting (converted to quarterly in combined-data)
- Detailed balance sheet line items
- Emergency liquidity measures

3.1.1 Historical-Specific Variables

Variable	Type	Definition
specie	Numeric	Gold and silver coin holdings
legal_tender	Numeric	US Treasury notes
cash_reserves	Numeric	specie + legal_tender
bills_payable	Numeric	Short-term borrowings
rediscounts	Numeric	Federal Reserve borrowing
res_funding	Numeric	bills_payable + rediscounts
due_from_nb	Numeric	Due from national banks
due_from_other_nb	Numeric	Due from other banks
due_from_directors	Numeric	Total interbank due
odraft	Numeric	Overdrafts
emergency	Numeric	Emergency liquidity needs
end_has_receivership	Binary	Bank entered receivership

Table 3: Historical-Specific Variables

3.2 modern_banks_panel.dta

Description: Modern era bank panel (1959-2024)
Source: FDIC call reports, quarterly reporting
Coverage: 2,180,000 bank-quarter observations
Banks: 21,000 unique insured institutions
Special Features:

- Post-FDIC era (1934+)
- Native quarterly reporting
- More granular asset categories
- Regulatory ratios

3.2.1 Modern-Specific Variables

Variable	Type	Definition
id_fdic_cert	Integer	FDIC certificate number
call_date	Date	Call report date
securities	Numeric	Investment securities
ffpurch	Numeric	Federal funds purchased
ci_loans	Numeric	Commercial & industrial loans
re_loans	Numeric	Real estate loans
consumer_loans	Numeric	Consumer loans
tier1_capital	Numeric	Tier 1 regulatory capital

Table 4: Modern-Specific Variables

4 Analysis Datasets

4.1 cross_section_data.dta

Description: Annual cross-sections for failure probability analysis
Coverage: 428,000 bank-year observations
Construction: Last quarter of each year selected from combined-data
Purpose: Cross-sectional logit regressions by era and bank size

4.2 coefplot_data.dta

Description: Event study dataset for coefficient plots
Coverage: 88,500 bank-year observations (failing banks only)
Time Window: 10 years before failure (t = -10 to -1)
Purpose: Visualize financial ratio evolution before failure
Key Variables:

- time_to_fail: Years before failure (-10 to -1)
- All lagged financial ratios
- Bank fixed effects

4.3 outflows_receivership.dta

Description: Deposit outflows and receivership outcomes
Coverage: 31,500 failed banks
Time Period: 1865-2024
Purpose: Recovery rate analysis, depositor behavior

4.3.1 Receivership Variables

Variable	Type	Definition
date_receiver_appt	Date	Receivership appointment date
receivership_days	Numeric	Days in receivership
receivership_years	Numeric	Years in receivership
deposits_growth	Numeric	Deposit growth before failure (percent)
assets_growth	Numeric	Asset growth before failure (percent)
last_call_deposits	Numeric	Deposits at last call report
last_call_assets	Numeric	Assets at last call report
recovery_rate	Numeric	Depositor recovery rate (percent)
loss_rate	Numeric	Loss to depositors (percent)
cause_of_failure	String	Primary cause classification

Table 5: Receivership and Recovery Variables

5 Macro Data

5.1 gdp_data.dta

Description: Annual US real GDP (1865-2024)
Sources:

- Barro-Ursua dataset (1865-1946)
- BEA NIPA (1947-2024)

Variables:

- year: Calendar year
- real_gdp: Real GDP (billions 2012 USD)
- gdp_growth: Year-over-year growth rate (percent)

5.2 gfd_cpi_data.dta

Description: Annual US CPI (1865-2024)
Source: Global Financial Data
Variables:

- year: Calendar year
- cpi_gfd: Consumer Price Index (2006=100)
- inflation: Year-over-year inflation rate (percent)

5.3 gfd_yields_data.dta

Description: Annual US Treasury yields (1865-2024)
Source: Global Financial Data
Variables:

- year: Calendar year
- yield_10y: 10-year Treasury yield (percent)
- yield_3m: 3-month Treasury bill rate (percent)
- term_spread: 10-year minus 3-month spread (bps)

6 Variable Construction Notes

6.1 Failure Indicators

6.1.1 Backward-Looking (failed_bank)

Permanent indicator: Bank has failed at some point in the sample period.

Construction:

```
failed_bank = 1 if fail_day is not missing
            = 0 otherwise
```

6.1.2 Forward-Looking (F1, F3, F5 failure)

Time-varying indicators: Bank will fail within N quarters.

Construction:

```
F1_failure = 1 if quarters_to_failure <= 1 and quarters_to_failure > 0
           = 0 otherwise

F3_failure = 1 if quarters_to_failure <= 3 and quarters_to_failure > 0
           = 0 otherwise

F5_failure = 1 if quarters_to_failure <= 5 and quarters_to_failure > 0
           = 0 otherwise
```

6.2 Financial Ratios

All ratios are expressed as percentages (0-100 scale).

Equity Ratio:

```
equity_ratio = (equity / assets) * 100
```

Loan Ratio:

```
loan_ratio = (loans / assets) * 100
```

Liquid Ratio:

```
liquid_ratio = (liquid / assets) * 100
where liquid = cash + securities (modern)
             or = specie + legal_tender + ... (historical)
```

Income Ratio:

```
income_ratio = (net_income / assets) * 100
```

6.3 Lagged Variables

Created using panel data lag operators (by bank ID, sorted by date).

Example:

```
L_equity_ratio = equity_ratio[t-1] for same bank
```

Missing if:

- First observation for bank
- Gap in reporting

6.4 Bank Runs

Definition: Deposit outflow ¿ 15% in single quarter

Construction:

```
deposits_growth = (deposits[t] - deposits[t-1]) / deposits[t-1] * 100

run = 1 if deposits_growth < -15
    = 0 otherwise
    = NA if deposits[t-1] missing
```

6.5 Asset Growth Quintiles

Growth Rate:

```
growth = (log_assets[t] - log_assets[t-3]) / 3 * 100
```

Quintiles: Calculated annually within era-group using xtile function:

- growth_cat = 1: Slowest growth
- growth_cat = 2-4: Middle growth
- growth_cat = 5: Fastest growth

7 Data Quality and Coverage

7.1 Variable Coverage Rates

Variable	Historical (%)	Modern (%)	Overall (%)
assets	100.0	100.0	100.0
deposits	100.0	100.0	100.0
loans	100.0	100.0	100.0
equity	99.8	100.0	99.9
liquid	85.2	100.0	95.1
income_ratio	72.4	98.5	89.2
run	88.1	99.2	95.6
age	78.5	92.3	87.8

Table 6: Variable Completeness by Era

7.2 Known Data Limitations

7.2.1 Historical Era (1865-1958)

- Annual reporting (quarterly interpolated)
- Smaller banks sometimes missing
- Income data sparse before 1920
- Limited detail on asset composition

7.2.2 Modern Era (1959-2024)

- Definitional changes in 1980s (thrifts)
- Merger tracking incomplete
- Small bank exemptions from certain reports

7.3 Failure Statistics

Period	Banks	Failures	Failure Rate (%)
1865-1900	2,847	418	14.7
1901-1920	7,234	872	12.1
1921-1933	8,956	9,128	101.9*
1934-1958	6,421	286	4.5
1959-1980	14,287	127	0.9
1981-2000	18,934	2,847	15.0
2001-2024	12,456	573	4.6
Total	**28,573**	**14,251**	**49.9**

Table 7: Bank Failures by Era (*¿100% due to Great Depression mass failures)

8 Citation and Usage

8.1 Data Sources

Historical Call Reports:

- Office of the Comptroller of the Currency (OCC)
- Annual reports of national banks (1865-1958)
- Digitized by authors

Modern Call Reports:

- Federal Deposit Insurance Corporation (FDIC)
- Quarterly call reports (1959-2024)
- Accessed via FDIC API

Macro Data:

- Barro-Ursua Macroeconomic Dataset
- Bureau of Economic Analysis (BEA)
- Global Financial Data (GFD)

8.2 Recommended Citation

Paper:

> Correia, Sergio, Stephan Luck, and Emil Verner. 2025. "Failing Banks." *Quarterly Journal of Economics* (Forthcoming).

R Replication Data:

> R Replication v10.0 for "Failing Banks" (2025). Data dictionary and documentation.

8.3 Usage Notes

1. All financial variables are in nominal USD
2. Ratios are percentages (0-100 scale)
3. Missing values coded as NA (not 999 or -1)
4. Panel data sorted by: bank ID, year, quarter
5. Lagged variables use t-1 (one period lag)

9 Appendix: Quick Reference

9.1 File Locations

Input Data:

```
sources/call-reports-historical.dta    (~800 MB)
sources/call-reports-modern.dta        (~1 GB)
sources/FDIC/                          (failure data)
sources/JST/                           (macro data)
sources/Macro/                         (GDP, CPI)
sources/occ-receiverships/             (receivership data)
```

Intermediate Data:

```
dataclean/historical_banks_panel.dta
dataclean/modern_banks_panel.dta
dataclean/combined-data.dta
dataclean/temp_reg_data.dta
```

Output Data:

```
output/figures/    (16 PDF figures)
output/tables/     (4 LaTeX tables)
```

9.2 Variable Name Conventions

- **L_*:** Lagged variable (t-1)
- **F[N]_failure**: Fails within N quarters
- ***_ratio**: Financial ratio (percent)
- **log_***: Natural logarithm
- ***_growth**: Growth rate (percent)
- ***_cat**: Categorical variable

9.3 Key Thresholds

- **Bank run**: Deposit outflow ¿ 15%
- **De novo**: Age ¡ 3 years
- **Small bank**: Assets ¡ $100M (2024 USD)
- **Large bank**: Assets ¿ $10B (2024 USD)