Leibniz Institute for Financial Research SAFE
Theodor-W.-Adorno-Platz 3 I 60629 Frankfurt am Main I Germany

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Information on the preprocessed merged CRSP-Compustat data set

Alexander Hillert
SAFE Research Datacenter*

Last Update: March 15, 2025

Contents

1 Contact

If you encounter any difficulties or just want general information, do not hesitate to contact us.

SAFE Research Datacenter: datacenter@safe-frankfurt.de

More information about the SAFE Research Datacenter, and further guides can be found here, and here.

*datacenter@safe-frankfurt.de

2 Introduction

The Center for Research in Security Prices (CRSP) database, which provides stock market data, and the Compustat (CS) database, which provides firm-level accounting data, are the backbone of empirical U.S. finance research. Even though there is a linking table available on the Wharton Research Data Services (WRDS), researchers still must preprocess the CRSP and Compustat data sets and then merge them to obtain the combined data set that they can use for their analyses. Our preprocessed merged CRSP-Compustat data set seeks to fill this gap.

The preprocessed file is based on the CRSP data from February 2025 and on Compustat data from March 2025 and includes the period from December 1949 (start of Compustat's annual data; Compustat's quarterly data only start in March of 1961; CRSP data start already in 1926) until December 2024 (most recent available period).

The preprocessed file is a csv file named

›› *CRSP_monthly_Compustat_quarterly_merged_1961_2024.csv*

It uses comma as delimiter. String variables are enclosed by double quotes. It has been saved as a zip file to improve efficiency.

3 Data processing steps

Note that below, variable names are indicated in brackets.

- In CRSP, restrict the sample to ordinary shares of U.S. firms, i.e., keep securities with share codes [shrcd] 10 and 11.
- In CS, remove financial reports that are in Canadian dollar, as mixing numbers in Canadian dollar with numbers in US dollar can cause problem, for example, when computing book-to-market ratios. (CRSP's variables like share price are always in US dollars.)
  Note that Compustat contains not only U.S. firms but also firms from Canada.
- In CS, eliminate duplicates in terms of firm [gvkey] and report date [datadate]. When firms change their fiscal year end, there are sometimes two observations for the same firm at the same report date. For example, firm A's financial report on September 30 could be their Q2 report according to their old fiscal year end (March 31) but their Q3 report according to their new fiscal year end (December 31).
- Preprocess the CRSP-Compustat linking table
  - Remove entries with missing permno, i.e. CS's gvkey without any match in CRSP.
  - Keep only the following four link types (definitions are from the data provider's manual):
    * LC: "Link research complete. Standard connection between databases."
    * LU: "Unresearched link to issue by CRSP."
    * LS: "Link valid for this security only. Other CRSP PERMNOs with the same PERMCO will link to other GVKEYs."
    * LN: "Primary link exists, but Compustat does not have prices."

  - This filtering for the four link types is equivalent to eliminating links with types LX ("Link to a security that trades on another exchange system not included in CRSP data.") and LD ("Duplicate link to a security. Another GVKEY/IID is a better link to that CRSP record.").

- Using the preprocessed linking table to merge the CRSP and Compustat data by gvkey and the year-month of the date variables ([date] for CRSP and [datadate] for Compustat).

  (1) We do not merge by date directly but instead by the year-month of the date because CRSP's dates refer to the last trading day of the month (e.g., December 30), whereas Compustat's date refers to the last calendar day of the month (e.g., December 31).

  (2) We use each permno-gvkey links from the preprocessed CRSP-Compustat linking table only for the period for which the links are valid. The period is indicated by it start date [LINKDT] and end date [LINKENDDT]. If the end date is a missing value, the link is still active.

4 CRSP and Compustat variables included

Below you will find the list of CRSP and Compustat variables included in the data set. The data source is indicated in brackets after the variable name. For Compustat, the annual variables names (e.g., AT) are displayed and the quarter variable names (e.g., ATQ) are shown in parentheses.
The list is split into three categories (1) firm/stock IDs, (2) dates and time IDs, and (3) actual accounting/stock market information including self-constructed variables. If variables are self-computed, it is indicated.
To give researchers the possibility to determine how outliers should be handled, none of the variables is winsorized or truncated. For some variables like, for example, the book-to-market ratio winsorizing is recommended.

4.1 Firm/stock identifiers

- **Permno** [CRSP]: unique security-level (=stock-level) identifier in the CRSP stock database.

- **Permco** [CRSP]: unique firm-level identifier in the CRSP stock database.
  One company (=a single permco) can have multiple share classes outstanding (=multiple permnos). A prominent example is Alphabet Inc. (the parent company of Google). Alphabet (permco== 45483) has permno==14542 (class C shares) and permno== 90319 (class A shares) outstanding.

- **Gvkey** [CS]: unique firm-level identifier in the Compustat accounting database.
  Like the permco, the gvkey is a unique company identifier. However, permcos and gvkeys may change differently in response to corporate events like restructurings and M&As.
  To add accounting variables to the data set, use gvkey and time (e.g., reporting date ("datadate")) as identifiers in the merge.

- **Comnam** [CRSP]: historical company name from CRSP.
  This variable contains abbreviations like "Mgmt" for "Management" or "Intl" for "International".

- **Conm** [CS]: the current (most recent, NOT historical) company name from Compustat.
  This variable contains abbreviations like "Mgmt" for "Management" or "Intl" for "International".

4.2 Dates and time identifiers

- **Date** [CRSP]: the date when the information was recorded. In the monthly CRSP data, it is the last trading day of the month. CRSP's point-in-time data (e.g., share price, number of shares outstanding) represent the information on that day.

- **Month_id** (self-computed): numerical time identifier on the monthly frequency based on the date from CRSP (January 1960 is month_id==0, February 1960 is month_id==1, January 1961 is month_id==12, February 1961 is month_id==13, October 2023 is month_id==765; the variable has been constructed using Stata's mofd()-function).
  This variable is helpful to define a panel data set (e.g., in Stata "xtset permno month_id").

- **Datadate** [CS]: the reporting date of the quarterly/annual accounting data. The reporting date is the last day of the fiscal year/quarter. The annual/quarterly report and its accounting information are only published several weeks after the fiscal year/quarter end. So, the information is not yet available to investors at the fiscal quarter/year-end date [datadate] but only at the report date [rdq] (see next item).

- **Rdq** [CS]: reporting date of the fiscal quarter's/year's earnings.
  This variable is only available in Compustat quarterly database.

4.3 Accounting and stock market variables

a) Accounting variables

- **Exchcd** [CRSP]: the code for the exchange the stock is listed at
  - Exchcd==1: NYSE
  - Exchcd==2: Amex
  - Exchcd==3: Nasdaq

- **Market_cap** (self-computed from CRSP): the stock's market capitalization in million USD. It is calculated as the number of shares outstanding times the closing price (i.e., shrout * abs(prc) / 1000).

- **Prc** [CRSP]: closing price of a stock on the given date. In the monthly data, it is the closing price on the last trading day of the month.
  Negative numbers indicate that the price is not an actual closing price but the average of the bid and the ask. Recommendation: use the absolute value.

- **Ret** [CRSP]: the holding period return. In CRSP daily (monthly), it the return from yesterday's (last month's last trading day's) closing price to today (this month's last trading day's) closing price. The variable is adjusted for dividends and stock splits.

- **Shrout** [CRSP]: number of shares outstanding in thousands of shares.

- **Siccd** [CRSP]: historical Standard Industry Classification Code
  Based on this variable, on can create Fama and French (1997) industry groups (go to https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html and scroll down to "Industry Portfolios").
  Compustat also provides information on the standard industry classification code. However, it shows today's industry classification, and it is not historical information.

- **Vol** [CRSP]: the trading volume of a stock on the day (daily CRSP data) or during the month (monthly CRSP data). It is recorded in the number of shares traded. In the monthly CRSP files, volume is expressed in units of hundreds (100).

b) Stock market variables

- **At (atq)** [CS]: the firm's total assets according to the most recent annual (quarterly) report. Numbers are expressed in millions of USD.
- **Book_market** (self-constructed from CRSP and CS): the book-to-market ratio computed as Ceq over market cap.
  The variable can have negative values. It is common to drop observations with a negative book-to-market ratio.
- **Ceq (ceqq)** [CS]: book value of the firm's common equity in million USD at the end of the fiscal quarter/year. It is defined as the common/ordinary stock plus capital surplus plus retained earnings minus treasury stock.
- **Cogs (cogsq)** [CS]: the costs of the goods/services sold in the most recent fiscal year (quarter); subtracting Cogs from Sale is a firm's gross profitability. Numbers are expressed in millions of USD.
- **Dlc (dlcq)** [CS]: the firm's debt in current liabilities at the end of the most recent fiscal year (quarter). Numbers are expressed in millions of USD.
- **Dltt (dlttq)** [CS]: the firm's long-term debt at the end of the most recent fiscal year (quarter). Numbers are expressed in millions of USD.
- **Epsfi (epsfiq)** [CS]: earnings per share including extraordinary items on a diluted share basis during the fiscal year.
- **Epsfx (epsfxq)** [CS]: earnings per share excluding extraordinary items on a diluted share basis during the fiscal year.
- **Epspi (epspiq)** [CS]: earnings per share including extraordinary items on a basic share basis during the fiscal year.
- **Epspx (epspxq)** [CS]: earnings per share excluding extraordinary items on a basic share basis during the fiscal year.
- **Lt (ltq)** [CS]: the firm's total liabilities at the end of the most recent fiscal year (quarter). Numbers are expressed in millions of USD.
- **Ni (niq)** [CS]: the firm's net income or net loss (negative number) during the most recent fiscal year (quarter). Numbers are expressed in millions of USD.
- **Oibdp (oibdpq)** [CS]: operating income before depreciation. Numbers are expressed in millions of USD.
- **Sale (saleq)** [CS]: the firm's total sales during the most recent fiscal year (quarter). Numbers are expressed in millions of USD.

5 Information on data frequency

CRSP data are available for **each day (daily data) or month (monthly data)** in which a stock is listed.
Compustat data are typically only available **once per year ("Compustat annual")** or **once per quarter ("Compustat quarterly")** when the firm has its fiscal year end or fiscal quarter end, respectively.
It is common to use Compustat-based variables like, for example, the book-to-market ratio or the gross profitability for twelve months (based on the annual Compustat data) or for three months (based on the quarterly Compustat data). To ensure that the information was actually available at a point in time, Compustat data are only used with a lag of several months. I.e., variables using Compustat information as not used directly at after the fiscal year/quarter end (datadate) but only some months later.
A popular approach is to follow Fama and French (1993): to construct their factor portfolios, Fama and French (1993) use the accounting data from year t only starting at July of year t+1. Other approaches include using a four-month gap between the fiscal year end month and the time when the accounting information is used.

6 Structure of the combined CRSP-Compustat data set

The table below illustrates the structure of the combined CRSP-Compustat data set using Microsoft Corp. (MSFT) as an example. MSFT has its fiscal year end at the end of June. You see that the observations from June show that the datadate (i.e., the last day of the fiscal year/quarter) is June 30 and that it is the fourth fiscal quarter (variables fqtr and datafqtr). The earnings are released a few weeks after the fiscal quarter end (see variable rdq). For example, the results of the fourth quarter of 2022 (ending on June 30, 2022) were announced on July 26, 2022. In the months, which are not a fiscal year or quarter end, all Compustat variables (see, e.g., the total asset column "atq") are missing and only CRSP variables (see, e.g., the return column "ret") are available.

permco	comnam	date	ret	datadate	fyearq	fqtr	datafqtr	rdq	atq
8048	MICROSOFT CORP	30jun2021	.084988795	30jun2021	2021	4	2021Q4	27jul2021	333779
8048	MICROSOFT CORP	30jul2021	.051716536	.	.	.		.	.
8048	MICROSOFT CORP	31aug2021	.061528206	.	.	.		.	.
8048	MICROSOFT CORP	30sep2021	-.066118956	30sep2021	2022	1	2022Q1	26oct2021	335418
8048	MICROSOFT CORP	29oct2021	.176291078	.	.	.		.	.
8048	MICROSOFT CORP	30nov2021	-.001236351	.	.	.		.	.
8048	MICROSOFT CORP	31dec2021	.017332681	31dec2021	2022	2	2022Q2	25jan2022	340389
8048	MICROSOFT CORP	31jan2022	-.075344898	.	.	.		.	.
8048	MICROSOFT CORP	28feb2022	-.03720497	.	.	.		.	.
8048	MICROSOFT CORP	31mar2022	.031861804	31mar2022	2022	3	2022Q3	26apr2022	344607
8048	MICROSOFT CORP	29apr2022	-.099867046	.	.	.		.	.
8048	MICROSOFT CORP	31may2022	-.018124798	.	.	.		.	.
8048	MICROSOFT CORP	30jun2022	-.055320591	30jun2022	2022	4	2022Q4	26jul2022	364840
8048	MICROSOFT CORP	29jul2022	.093096621	.	.	.		.	.
8048	MICROSOFT CORP	31aug2022	-.066431537	.	.	.		.	.
8048	MICROSOFT CORP	30sep2022	-.109266862	30sep2022	2023	1	2023Q1	25oct2022	359784
8048	MICROSOFT CORP	31oct2022	-.003306093	.	.	.		.	.
8048	MICROSOFT CORP	30nov2022	.102054857	.	.	.		.	.
8048	MICROSOFT CORP	30dec2022	-.060045436	31dec2022	2023	2	2023Q2	24jan2023	364552
8048	MICROSOFT CORP	31jan2023	.033316612	.	.	.		.	.
8048	MICROSOFT CORP	28feb2023	.009240953	.	.	.		.	.
8048	MICROSOFT CORP	31mar2023	.155881599	31mar2023	2023	3	2023Q3	25apr2023	380088
8048	MICROSOFT CORP	28apr2023	.065764904	.	.	.		.	.
8048	MICROSOFT CORP	31may2023	.07098224	.	.	.		.	.
8048	MICROSOFT CORP	30jun2023	.036998671	30jun2023	2023	4	2023Q4	25jul2023	411976
8048	MICROSOFT CORP	31jul2023	-.013566674	.	.	.		.	.
8048	MICROSOFT CORP	31aug2023	-.022267217	.	.	.		.	.

References

Fama, Eugene F. and Kenneth R. French (1993) "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 33 (1), 3–56.