Guide to constructing QFR time series

Cooper Howes

June 2023

Disclaimer: These data are created from public releases of the Quarterly Financial Report for Manufacturing Corporations (QFR). While I have done my best to ensure their accuracy, it is possible that mistakes remain, and neither the historical digitized data nor the cleaning program have been verified or endorsed in any capacity by the Census Bureau or any other organization. If you use these data, please cite my paper "Financial Constraints, Sectoral Heterogeneity, and the Cyclicality of Investment". Questions, comments, and suggestions are welcome.

The raw QFR data come from the Excel file "QFR_raw_data.xlsx". Data prior to 1988 are manually digitized from original publications. Physical versions of the original QFR releases can be found in research libraries.[1] Each tab includes all published variables for the durable, nondurable, and total manufacturing sectors. Each publication reports data for that quarter as well as the prior four quarters. The tab names correspond to the reference year most closely corresponds to the data in each publication, while the specific dates covered in each tab are shown in cell C1. For example, the "77" tab contains the data used to calculate values for 1977 and covers observations from 1976Q4 through 1977Q4.

Column A reports the variable names, which are taken directly from the publications. Due to changes in methodology, the precise definitions for each variable and its availability across size/sector can change over time. The tab colors correspond to major changes, including changes in survey scope/methodology and industry reclassifications. For example, from 1966 through 1974, the QFR reports one measure for total operating expenses that is available for all firm sizes (row 3 in the purple tabs). From 1974 through 1980, the QFR reports separate entries for operating expenses (row 7 in the red tabs) and non-operating expenses (row 10 in the red tabs) for all firms. Starting in 1981, the QFR reports only operating expenses for companies with <$25 million in assets (row 6, columns A-BU in the orange tabs), but reports breakdowns for operating and non-operating expenses for firms with >$25 million in assets (rows 6 and 9, columns BV and beyond in the orange tabs).

The data also include splits by size and sector (durable, nondurable, and total). Data are aggregated to "buckets" that include all firms whose assets fall within a given range.

[1] Some digital versions can also be obtained through sources such as Hathi Trust. For example, a selection of QFR releases can be found here: https://catalog.hathitrust.org/Record/000517669.

The size ranges for the smallest buckets change over time; for example, in 1966, there were separate buckets for firms with <$1 million and $1-5 million in assets, while the lowest threshold was raised to $5 million in 1977. These sector and size classifications are shown in rows 1 and 2. In theory, firms can move between buckets over time as they grow and shrink, which suggests caution in treating these series as a pseudo-panel.[2] To minimize these concerns and allow for potential use of the data split by size, the cleaning program will also generate series for "small" and "large" firms within each sector. By default, it uses the classification from Crouzet (2017), which treats all firms with $1bn or more in assets as "large" and combines all other buckets into the "small" category.

For 1988 and beyond, data come from the Census website (https://www.census.gov/econ/qfr/historic.html). Each tab corresponding to a different quarter, with naming format "YYYYQQ". In addition, in 2001, the QFR switched its industry classification from SIC to NAICS, causing a discontinuity across the Census series. To address this, the "2001Q3_sic" tab manually digitized the report from 2001Q3, which allowed for calculating growth rates under the SIC methodology.

These methodological changes mean that the unrevised historical series available through the Census will not allow for consistent comparisons across time, and will instead exhibit large jumps driven by non-economic factors. To generate a consistent time series, I calculate percent changes for each variable *within the same report* to ensure consistent measurement. I then retroactively apply these percent changes to the most up-to-date level series provided by the Census. This process can be summarized as follows:

Step 1: Calculate growth rates for 1966-1987 based on manually digitized QFR data.

Step 2: Calculate growth rates for 1988-2001 from Census historical data.

Step 3: Calculate growth rates from manually digitized 2001Q3 release to handle SIC/NAICS conversion.

Step 4: Combine the steps above to get a consistent series of growth rates from 1966Q1-2001Q3

Step 5: Collect observations for 2001Q4 through 2021Q4 directly from Census historical data.

Step 6: Retroactively apply the growth rate series from step 4 to the 2001Q4 level from step 5 to generate a consistent time series for 1966-2021.

[2] In practice, however, past work using QFR micro data suggests that this kind of switching is relatively infrequent; see Crouzet and Mehrotra (2020).

The program “QFR_data_cleaner.R” performs these steps by converting the raw Excel file into the quarterly time series used in the main paper. Given the varied structure of the QFR reports over time, this program requires specifying which rows correspond to which variables for each format. By default, the program calculates the six variables used in the main paper: NPPE, sales, total liabilities, current (<1 year maturity) liabilities, dividends, and total equity. These are shown in lines 59-64, with a description about how each time period is broken down shown in lines 49-57. For example, line 61 specifies the rows used to calculate total liabilities. From 1966-1973 (the purple tabs), total liabilities will be shown in row 40. From 1974-1980 (the red tabs), total liabilities are in row 70. Starting in 1980, the survey was stratified into two pieces based on a firm’s size, so for this period there are two separate rows that need to be reported; one for the smallest firms (row 54), and one for larger firms (row 70). For 1988 and beyond, the descriptions of each variable and their rows can be found in the “KEY” tabs.[3]

By default, the program will create nine series for each variable specified in “qfr_var_names” (shown in line 67): three industry groupings (total manufacturing, durable manufacturing, and nondurable manufacturing) times three size categories (all firms, “large” firms with $1bn+ in assets, and “small” firms with <$1bn in assets). The steps for adding a new series are outlined below:

Step 1: Find the row corresponding to the variable that you want for each of the time windows shown in lines 44-52.

Step 2: Create a new list that specifies which rows to use following the format of those shown in lines 54-59. This object should be named “X_rows”, where “X” is the name given to the variable you want to include.

Step 3: Add the prefix name “X” from step 2 to the list of variables to be processed in the “qfr_var_names” (line 67).

Step 4: Run the entire program, which will generate series for all of the variable prefixes included in the “qfr_var_names” object.

The QFR data are all reported in nominal terms. In lines 439-441, the program allows you to specify which deflators (if any) to use for each series to create real versions. By default, I use the BEA’s nonresidential fixed investment price deflator to deflate NPPE, and the GDP price index to deflate all other variables. To use a different deflator, first add it to the “Aggregate_data.csv” file found in the replication package, and then specify it in lines 439-441.

[3] The rows shown in each “key” tab for each variable are the same. The differences between the three arise solely from differences in industry codes. These keys are included for completeness, though the program only considers the total durable and nondurable sectors, for which the codes do not change over time.

After the program runs, the output consists of the following for each variable prefix “X” specified in line 67:

1. A data frame titled “X_data” which reports the adjusted nominal levels of that variable by size and industry.
2. A data frame titled “X_data_real” which reports the real series of that variable by size and industry, deflated using the price indices specified in lines 439-441.
3. A data frame titled “X_data_raw”, which reports the unadjusted nominal level of that variable by size and industry. These series should not be used on their own, because they will not take into account methodological changes and can lead to incorrect growth rates, but they *should* be used to calculate ratios.

These data frames are also combined into several objects named “type_X_size_sector”. The type classification will be either “real” or “nom” depending on whether that series is deflated. Size classifications are s/l/t for small/large/total, and sector classifications are d/n/a for durable/nondurable/all, so that the series “nom_nppe_l_n” corresponds to the nominal capital stock (NPPE) for large nondurable producers. All of these series of each type are combined into objects titled “real_data”, “nom_data”, and “raw_data” that include all of the variables of each type for all series.

The program also calculates ratios of raw variables. As mentioned above, these use the unadjusted series to facilitate consistent comparisons of the level of each ratio over time. By default, the program calculates three ratios: the current liability ratio (liabilities with maturity of less than one year divided by total liabilities), the cashflow ratio (net income divided by NPPE), and the dividend ratio (dividend payments divided by the book value of equity). These ratios, along with the real series for sales and NPPE, are saved into a file called “qfr_cleaned.csv”. This file is combined with the other macro series in the “Aggregate_data.csv” file used in the main regressions.