NBER-CES Manufacturing Industry Database: Technical Notes

Randy Becker, Wayne Gray, Jordan Marvakov
February 2013

1 Introduction

The NBER-CES Manufacturing Industry Database contains annual data from the United States manufacturing sector for the period from 1958 to 2009. The data used for the development of the database come from various sources, but chiefly from four government agencies: the Bureau of the Census, the Bureau of Economic Analysis, the Bureau of Labor Statistics, and the Federal Reserve Board. The goal is to provide a long time-series of data for a large number of industries, adjusting for changes in industry definitions, and creating price deflators, capital stocks, and productivity estimates. This paper describes issues related to the most recent update of the database (1997 to 2009). For a more detailed discussion of an older version of the data see Bartelsman and Gray (1996).

The key feature (and complication) of recent updates is the adjustment to the NAICS industry definitions. It is published in two versions: one based on the 1987 Standard Industrial Classification (SIC), containing 459 industries, and the other based on the 1997 North American Industrial Classification System (NAICS), containing 473 industries.[1] The boundaries of the manufacturing sector shifted between SIC and NAICS, with some SIC industries leaving manufacturing and other NAICS industries entering manufacturing. The exiting SIC industries don't have data after 1996, and the entering NAICS industries don't have data before 1997. To enable the calculation of price deflators in each version, the benchmark year for deflators in the NAICS version is 1997, while the SIC version continues to use 1987 as a benchmark. This version of the database recalculates all the data post-1996, and earlier updates of the database (1958-2002 and 1958-2005) are superseded and should not be used.

The database is available for download at http://www.nber.org/data/nberces5809.html in several different formats, along with concordances between SIC and NAICS industries and other related files.

2 Data and Sources

2.1 ASM/Census Data

The majority of the variables provided in the database were extracted from the Annual Survey of Manufactures (ASM). The ASM is a representative sample of approximately 50,000 establishments selected from the approximately 330,000 establishments included in the Census of Manufactures (CMF). The ASM is conducted annually, except for years ending in 2 and 7, when it is part of the CMF.

ASM/CMF data is used for eleven of the nineteen variables in the database: total employment (*emp*), total payroll (*pay*), production workers (*prode*), production worker hours (*prodh*), production worker wages (*prodw*), total value of shipments (*vship*), total cost of materials and energy (*matcost*), total value

[1] Earlier versions of the database were available on a 1972 SIC basis, with 448-450 industries.

added (*vadd*), total capital expenditure (*invest*), end-of-year inventories (*invent*), and the cost of fuels and electricity (*energy*).[2] Definitions for these ASM/CMF variables are presented in Appendix A.

The above variables are denominated in millions of nominal dollars, except for labor variables that are denominated in thousands of workers or millions of worker hours. In order to convert them to real dollars and calculate productivity factors, four different deflators are used (*piship, pimat, pien,* and *piinv,* for *vship, matcost, energy,* and *invest*).

One variable that underwent a substantial change in definition is *invest.* The ASM reported new capital expenditures before 1997 (with used capital expenditures in an additional table), but switched to reporting the sum of new and used expenditures in 1997, with no separate breakout of new and used expenditures (even in Census years). In order to maintain consistency with the earlier data, we adjusted the post-1996 data to reflect new expenditures. The data on new and used expenditures from 1987, 1992, and 1996 were combined to derive a "fraction new" (new/(new+used)) for each industry, which was then used to adjust the post-1996 capital expenditure numbers to reflect only new expenditures.

.2.2 Industry Capital Stocks and the Investment Deflator (PIINV)

The ASM provides annual data for gross nominal capital expenditures on equipment and structures. However, it lacks information on depreciation rates and investment deflators, which are necessary in order to construct the real capital stocks in the database.

Depreciation rates are derived from Federal Reserve Board data (FRB). FRB series contain equipment and structures stocks and expenditures in real and nominal terms for 4-digit NAICS industries, providing less industry detail than the 6-digit NAICS Census data. Depreciation rates are calculated for each industry, backing them out of a standard perpetual inventory equation:

$$K_t = (1 - d_{t-1}) K_{t-1} + I_t \tag{1}$$

In equation (1), d is the rate of depreciation, K is the real capital stock in years t and $t-1$, and I is the real capital investment in year t.

To calculate the investment deflator, price indexes for equipment and structures are computed from the ratios of nominal and real investment for the 4-digit NAICS industries in the FRB data. The final deflator values are created as the weighted mean of the two types of capital expenditures. The shares used for the weights are averaged over years t and $t-1$:

$$\Delta P_t = \sum_{n=1}^{2} \frac{1}{2}(S_{n,t} + S_{n,t-1}) \Delta P_{n,t} \tag{2}$$

In equation (2), ΔP is the annual change in the price level, S is the share of total capital investment for each category, and the summation occurs over n=2 categories (equipment and structures). Note that the data on investment shares come from the Census data, so they are available at the 6-digit level. The

[2] Note that the *energy* variable is included in *matcost*; to treat energy and materials as distinct inputs requires calculating non-energy materials (*matcost* - *energy*).

depreciation rates and the investment deflators are at the 4-digit NAICS industry level, so we assign each 4-digit "parent" to all of its 6-digit NAICS "offspring".

2.3 The Shipments Deflator (PISHIP)

The price indexes used in the calculation of the shipments deflator come from the Bureau of Economic Analysis (BEA) GDP-by-Industry data, found at http://www.bea.gov under the "Industry" tab.

The BEA price index data contain several "rolled-up" 6-digit codes that combine two or more 6-digit NAICS industries. For example, NAICS 311411 (Frozen Fruit, Juice, and Vegetable Manufacturing) and NAICS 311412 (Frozen Specialty Food Manufacturing) are "rolled up" into a single code, 311410. This problem is addressed by assuming that industries within each roll-up have the same price structure, and assigning the values of the parent to all associated 6-digit NAICS offspring codes.

2.4 The Materials Cost Deflator (PIMAT)

The materials cost deflators are calculated using data from the "use-make" (input-output) tables and the GDP-by-Industry data of the Bureau of Economic Analysis (BEA). The raw data are available at the BEA website (http://www.bea.gov) under the "Industry" tab.

The use-make tables disaggregate each industry's total materials cost into the amount spent on each specific material. The share of each material in the industry's total materials cost is then used to weight each material's price index, resulting in a weighted-average price index for the industry's total materials cost, similar to equation (2). We use the benchmark versions of the use-make tables, which are generated every 5 years. We use a moving average of the weights from the two surrounding benchmark years to interpolate the weights for the use-make tables in the intervening years. Inputs coming from manufacturing industries use the price index in our database for that industry's shipments. Inputs coming from other (non-manufacturing) sectors use that sector's price index for gross output.

The main difficulty with implementing the above procedure is the lack of a clear one-to-one correspondence between the industry codes used in the use-make tables and those used in the ASM/CMF data. The use-make tables contain many fewer industries than the NAICS industries. BEA does provide concordance tables, which are supplemented with some manual matching based on industry definitions. We assume that all NAICS industries which map into the same use-make industry have the same structure of inputs.

2.5 Energy Usage Data and the Energy Deflator (PIEN)

The energy deflator is based on each industry's expenditures on seven types of energy: electricity, residual fuel oil, light oils, liquefied petroleum, coal, coke, and natural gas. Previous updates used information for six types of energy (the "distillates" category was recently split into light oils and liquefied petroleum). As with the materials cost deflator, we combine price indices for the different energy types with information on each energy type's share in the industry's total energy expenditures, in a calculation similar to that shown in equation (2).

The data on energy expenditures for six of the fuels in the current update (excluding electricity) come from the 2002 Manufacturing Energy Consumption Survey (MECS), available at the Bureau of the

Census website (http://www.census.gov). The ASM data provide information on electricity and non-electricity energy spending, and the MECS data are used to disaggregate each industry's non-electricity spending share into the share of each of the six non-electricity energy types.

There are a number of missing values in the MECS data. "Extremely small" values (less than 0.5), as well as values with high relative standard error (more than 50%) are withheld from the official MECS release by Census. Additionally, some values are withheld to avoid disclosing information for individual establishments in cases where an industry relatively few plants. The "extremely small" values are replaced by 0.1. Since MECS reports totals by industry and type of energy, it is possible to impute the few remaining missing values at the 3-digit NAICS level by running a cross-check against the unassigned residuals from those totals.

MECS data contain a full set of 3-digit NAICS industry codes, but little industry detail beyond that. Only a handful of 4-, 5-, and 6-digit NAICS industries are included in the MECS, and the cross-check approach cannot be used to fill missing values in those series. Therefore, all 4-to-6 digit NAICS industries with missing values have to be discarded, and the next higher level of detail is used in the construction of the energy deflator.

Price indexes for each non-electricity energy type are obtained from the Bureau of Labor Statistics database (http://www.bls.gov), using the industrial price indexes. The codes of the extracted BLS variables are:

WPU051	Coal
WPU052	Coke oven products
WPU0532	Liquefied petroleum gas
WPU054321	Industrial electric power
WPU055321	Industrial natural gas
WPU0573	Light fuel oils
WPU0574	Residual fuels

Since the 1996 update, the *WPU052 Coke oven products* series has been discontinued. Thus, the coke price data used here are extrapolated from coal prices. Electricity prices are calculated from the industry-specific electricity consumption and expenditures data in the ASM tables, where available, and otherwise taken from the national-average values.

3 Industry Classification Changes

The system of industrial classification used by Census to report manufacturing data has changed several times over the lifespan of the NBER-CES database. The original SIC system underwent two substantial modifications in 1972 and 1987 before it was finally replaced with the NAICS industry codes in 1997. Earlier versions of the NBER-CES database were published using both SIC 1972 and SIC 1987 classifications. The 2009 version is provided in NAICS 1997 and SIC 1987 codes.

The change in industry classification from SIC to NAICS and the necessary conversion of existing data presents challenges to the extension of the NBER-CES dataset. In addition to the changes in industry classification, in 2003 the ASM data stopped providing full industry detail in non-CMF years. Several 6-digit NAICS industries are only reported in "rolled-up" form, with two or more 6-digit industries within the same 5-digit NAICS combined into a single value (designated by the 5-digit number plus one letter).

These roll-ups are broken back down into their component 6-digit NAICS industries, using CMF year values.

The biggest hurdle arising from the SIC to NAICS transition is the entry, exit, and realignment of manufacturing industries. While there is direct correspondence between many 4-digit SIC industries and the new 6-digit NAICS industries, there are also many cases in which SIC industries are split or multiple pieces from SIC industries are joined to form a new NAICS industry. For example, SIC 2077 (Animal and Marine Fats and Oils) is split into NAICS 311613 (Rendering and Meat Byproduct Processing), 311711 (Seafood Canning), and 311712 (Fresh and Frozen Seafood Processing), while NAICS 311712 also includes a part of SIC 2092 (Prepared Fresh or Frozen Fish and Seafoods).

The 1997 CMF concordance table linking NAICS 1997 and SIC 1987 codes is instrumental in the conversion process. This table provides a concordance based on three variables: employment, payroll, and shipments.[3] The employment split is used in the calculation of labor-related variables (*emp*, *prodh, prode).* The payroll split is used in the calculation of compensation related variables (*pay*, *prodw*). The shipments split is used for all remaining variables.

As pointed out by Bartelsman and Gray (1996), the use of this type of concordance is not without its pitfalls. Most notably, the composition of industries tends to change over time, and the use of a fixed concordance matrix may distort the data, especially in rapidly-evolving industries. While there is no real solution for this problem, it should be recognized when analyzing long time-series of data for industries that have been redefined over time.

While most manufacturing industries can be reconstructed under either SIC or NAICS definitions, there are some cases in which entire industries (or large fractions of them) were reclassified out of (or into) manufacturing. Eleven industries were classified as manufacturing under NAICS that had been primarily considered non-manufacturing under SIC (examples include Tire Retreading, Dental Laboratories, and Software Reproducing). The initial capital stocks of these new industries are obtained from 1997 CMF data, using the total book value of assets, though it should be noted that unlike the starting capital stocks used for 1958, these values have not been examined for sensitivity to accounting methods. The entry of these new industries forces us to shift the benchmark year to 1997 for the NAICS version of the database, so that deflator series can be properly defined (starting as 1.000 in 1997). These new manufacturing industries in the NAICS database have missing values for all years before 1997. On the other hand, seven SIC manufacturing industries move (in whole or part) to non-manufacturing under NAICS, and therefore have missing values for all years after 1996. Most notably, in the new NAICS system, the publishing industries are reclassified in their entirety from Manufacturing to Information (NAICS 51). A few industries are split between manufacturing and non-manufacturing in the SIC-NAICS concordance. In such cases, we retain only industries which keep 95+% of their activity in manufacturing, to avoid misleading shifts in 1997. This adjustment results in our dropping SIC 3732 (Boat Building and Repairing) after 1996, since only 87% of the industry's shipments remain in the NAICS manufacturing industry 336612 (Boat Building), with the other 13% moving to NAICS 811490 (Boat Repairing). Table 1 lists the industries that were part of this manufacturing realignment.

[3] We made some adjustments to the original 1997 CMF concordance table, based on comparisons to other concordances and manual review.

Table 1. SIC-NAICS Manufacturing Realignment in 1997.

Entering Industries

NAICS Code	*Name*
311330	Confectionery Manufacturing from Purchased Chocolate
311612	Meat Processed from Carcasses
311811	Retail Bakeries
313311	Broadwoven Fabric Finishing Mills
314121	Curtain and Drapery Mills
315999	Other Apparel Accessories and Other Apparel Manufacturing
326212	Tire Retreading
334611	Software Reproducing
334612	Prerecorded Compact Disc (except Software), Tape, and Record Reproducing
337110	Wood Kitchen Cabinet and Countertop Manufacturing
339116	Dental Laboratories

Exiting Industries

SIC Code	*Name*
2411	Logging
2711	Newspapers: Publishing, or Publishing and Printing
2721	Periodicals: Publishing, or Publishing and Printing
2731	Books: Publishing, or Publishing and Printing
2741	Miscellaneous Publishing
2771	Greeting Cards
3732	Boat Building and Repairing[4]

In addition, some NAICS industry definitions in the manufacturing sector went through changes during the 2002 and 2007 NAICS updates. Several of these changes cause the movement of sizable fractions between industries, making the data inconsistent over time without additional adjustments. Values for NAICS 327992 (Ground/Treated Mineral and Earth) and NAICS 339912 (Silverware/Hollowware) are adjusted upwards, since a substantial share of each was reassigned to different industries in the published numbers. NAICS 339111 (Laboratory Apparatus and Furniture) was discontinued in 2007, with most of its contents moving to NAICS 339113 (Surgical Appliances and Supplies). There were also changes between 2002 and 2007 which caused parts of NAICS 339112 and 339113 to move to NAICS 339111. We reverse these latter changes, and impute post-2007 NAICS 339111 values based on the last available ratio of the sizes of 339111 and 339113. Other industries adjusted for "moving parts" are: NAICS 331422, 335929, and 331491; NAICS 333311 and 333313; NAICS 334220 and 334419; NAICS 339912, 332812 and 332999.[5]

4 Productivity Measures

There are two versions of TFP in the NBER-CES Manufacturing Database: 4-factor and 5-factor. The 5-factor version separates out energy from non-energy materials; the 4-factor uses a single materials input

[4] Only 87% are retained in NAICS 336612 (Boat Building), which is insufficient to calculate post-1997 SIC estimates.
[5] For more details of the specific magnitude of these adjustments, please consult the concordance tables at http://www.census.gov/eos/www/naics/concordances/concordances.html.

(which includes energy). The TFP calculation requires definitions of the cost shares, the factors, the factor changes, and the output changes.

The five cost shares (α_i) vary by industry by year, defined using the variable names in the dataset:

(α_1) Non-production workers: *(pay-prodw)/vship*	[i.e., their pay divided by shipments]
(α_2) Production workers: *prodw/vship*	[i.e., their pay divided by shipments]
(α_3) Energy: *energy/vship*	[i.e., energy expenditure divided by shipments]
(α_4) Materials: *(matcost-energy)/vship*	[i.e., non-energy materials divided by shipments]
(α_5) Capital: 1 - (sum of the above shares)	[i.e., the residual]

In calculating TFP growth from one year to the next, we use the average of the two years' cost shares:

$(\alpha'_{it} = 0.5\alpha_{it} + 0.5\alpha_{it-1})$

The 5 factors (X_i) are defined as follows, using the variable names in the dataset:

(X_1) Non-production workers: *(emp-prode)*	[i.e., the number of non-production workers]
(X_2) Production workers: *prodh*	[i.e., production worker hours, not employees]
(X_3) Energy: *(energy/pien)*	[i.e., real energy expenditures]
(X_4) Materials: *((matcost/pimat) - (energy/pien))*	[i.e., real non-energy materials]
(X_5) Capital: *cap*	[i.e., total capital stock, already in real terms]

The change in factor usage between one year and the next is defined as the change in natural logs, (for example):

$dX_{it} = \ln(X_{it}) - \ln(X_{it-1})$

We also need the change in real output:

(Q) Real output: *vship/piship*

As with factor usage, we express output change in terms of natural logs, hence:

$dQ_t = \ln(Q_t) - \ln(Q_{t-1})$

The change in 5-factor TFP (*dTFP5*) between this year and last is thus defined as:

$$dTFP5_t = dQ_t - \Sigma_i\, \alpha'_{it}\, dX_{it}, \quad i = 1, ..., 5 \tag{3}$$

Given the series of *dTFP5* values, one can then "roll up" these changes to form a TFP index (*TFP5*), by setting the index equal to 1.0 in some initial year *t* and then growing the index forward by the following equation:

$$TFP5_{t+1} = \exp[\ \ln(TFP5_t) + (dTFP5_{t+1})\ ] \tag{4}$$

The values of 4-factor TFP growth (*dTFP4*) and the corresponding TFP index (*TFP4*) are calculated similarly, but using total materials cost spending rather than separating it into energy and non-energy materials.

5 Conclusions

This document discusses issues related to the 1958-2009 version of the NBER-CES productivity database. Further updates will follow as more years of data become available.

If you have questions, comments or problems to report about the database, please contact Dr. Wayne Gray at wgray@clarku.edu or Dr. Randy Becker at randy.a.becker@census.gov. Feel free to use the data in your research, with a citation of the appropriate NBER technical paper - we hope it will be useful.

6 References

Bartelsman, E and W. Gray. 1996. "The NBER Manufacturing Productivity Database". Technical Working Paper 205, National Bureau of Economic Research.