A Concordance Between Ten-Digit U.S. Harmonized System Codes and SIC/NAICS Product Classes and Industries*

Justin R. Pierce†
Board of Governors of the Federal Reserve System

Peter K. Schott‡
Yale School of Management & NBER

December 2011

Abstract

While the relationship between international trade and domestic economic activity is an important topic in economics, research in this area has been slowed due to data limitations. In this paper we provide tools that improve the existing data in two ways. First, we develop an algorithm that yields concordances between the ten-digit Harmonized System (HS) codes used to classify products in U.S. international trade and the SIC and NAICS industry codes used to classify domestic economic activity. These concordances then yield novel time series of industry-level international trade data for the years 1989 to 2009. Second, we provide concordances between HS codes and the SIC and NAICS product classes used to classify U.S. manufacturing production, allowing for matching at a more disaggregated level than was previously available.

Keywords: International trade; industry classification

JEL classification: F1

*We thank Julie Linden of the Yale University Social Sciences Library for generous help in securing the publicly available U.S. trade data. We thank Kitjawat Tacharoen and Matt Flagge for research assistance. We thank Alvin Venning, Carol Ann Aristone, James Kristoff and Mendel Gayle of the U.S. Census Bureau for many enlightening conversations. Schott thanks the National Science Foundation (SES-0241474 and SES-0550190) for research support. The analysis and conclusions set forth in this paper are those of the authors and do not indicate concurrence by the Board of Governors, other members of the research staff or the National Science Foundation.

†Correspondence: 20th & C ST NW, Washington, DC 20551; email: justin.r.pierce@frb.gov; telephone: 202-452-2980; fax: 202-736-1937.

‡135 Prospect Street, New Haven, CT 06520, tel: (203) 436-4260, fax: (203) 432-6974, email: peter.schott@yale.edu.

1. Introduction

Empirical researchers in the fields of international trade and industrial organization are increasingly focused on examining the relationship between international trade and domestic economic activity. This research agenda was first pursued with industry-level data as in Revenga (1992) and Sachs and Shatz (1994). More recently, demand for linked trade and production data has increased along with the massive growth of research using highly disaggregated plant and firm-level data, as in Bernard, Jensen and Schott (2009) and Bernard, Redding and Schott (2010). Applied research in these fields has been slowed, however, due to an inability to create long time series of industry-level international trade and production data or to match trade data to detailed product-level domestic data.

In the U.S., international trade data have been classified since 1989 based on the World Customs Organization's Harmonized System (HS). In contrast, domestic economic activity has been classified using the North American Industrial Classification System (NAICS)—beginning with the 1997 economic census—and the Standard Industrial Classification (SIC), prior to the 1997 economic census.[1] This creates two potential difficulties when linking trade and production data. First, the HS classifies products solely on physical characteristics while SIC and NAICS classify products based on physical characteristics and the type of economic activity. Second, the switch from SIC to NAICS beginning with the 1997 economic census means that it has been difficult to construct a time series linking trade and production data for the entire period from 1989 to present.

This paper improves on currently available data in two ways. First, we provide an algorithm that generates concordances linking the ten-digit HS codes used by the United States to track international trade with the four-digit SIC and six-digit NAICS industry codes used to characterize domestic economic activity. These concordances are assembled from published U.S. Census Bureau ("Census") data, which provide a mapping of HS to SIC and NAICS industries from 1989 to 2001 and 2000 to 2009, respectively. Our contribution here is to extend these mappings to match HS codes with SIC industries after 2001, and to match HS codes with NAICS industries before 2000. As a result, applied economists will be able to create—for the first time—linked datasets of trade and domestic production in both SIC and NAICS over a long time series–1989-2009 for NAICS and 1989-2006 for SIC.[2]

Second, we provide a set of concordances linking ten-digit import and export HS codes to one or more five-digit SIC (SIC5) or seven-digit NAICS (NAICS7) *product classes*. These concordances are constructed using bridge codes known as "basecodes," which are created by the U.S. Census Bureau ("Census"). In each year of an economic census, Census constructs two mappings linking HS codes to basecodes and linking basecodes to SIC5 or NAICS7 product classes, respectively. We combine the two mappings to directly link HS codes to product classes. This set of concordances then allows researchers to match international trade and domestic production data at a more disaggregated level than has previously been available. Each of the contributions in this paper improves the ability of empirical researchers to calculate measures of trade and domestic economic activity that are more directly comparable and hence more accurate for research purposes.

[1] Each of these product classification systems is described in more detail in Section 2.

[2] The reason for the shorter time period for HS-SIC4 mappings is discussed in Section 3 below.

Finally, we briefly discusses how these concordances might be applied in current empirical international trade research. In particular, we provide background information useful for linking the firm-product-class domestic production data in the U.S. Census of Manufactures (CM) to the firm-product import and export data in the Longitudinal Firm Trade Transaction Database (LFTTD). For more detail on the former, see Bernard, Redding and Schott (2010). For more detail on the latter see Bernard, Jensen and Schott (2009).

The remainder of the paper is organized as follows. Section 2 provides a description of the HS, SIC and NAICS classification systems. Section 3 describes the HS to SIC4/NAICS6 industry concordance, while Section 4 describes the HS to SIC5/NAICS7 product-class concordance. Section 5 discusses how the latter can be used to link Census production and trade data. Section 6 concludes. Appendices provide the Stata code used to implement our algorithm and generate the concordances discussed in the paper and describe the key files used to construct the concordances.

2. A Description of the HS, SIC and NAICS Classification Systems

2.1. Classifying Products in U.S. International Trade - The Harmonized System

International trade data in all major trading countries—including the U.S.—is classified based on the Harmonized System developed by the World Customs Organization (WCO). The WCO begins by assigning products into 99 broad 2-digit categories such as chapter 72, "Iron and Steel." These chapters are then further broken out into 6-digit HS codes for categories of goods such as heading 851670, which is defined in the 2007 HS as "Coffee or tea makers." Individual countries are then free to maintain more disaggregated classifications beyond the 6-digit level.

The U.S. maintains separate HS classifications for imports and exports and classifies products at the ten-digit level. Import codes are provided in the Harmonized Tariff Schedule and maintained by the U.S. International Trade Commission (ITC). Export codes—formally known as "Schedule B" codes—are maintained by the Foreign Trade Division (FTD) of the U.S. Census Bureau. In this paper we refer to import and export codes generically as HS codes. For import HS codes, the ITC further aggregates the 99 chapters into 12 broad "sections," which are listed in Table 1. The full listing of HS chapters and 10-digit HS import and export codes are available at websites of the ITC and FTD, respectively.

2.2. Classifying U.S. Domestic Economic Activity - SIC and NAICS

In contrast to the HS, which classifies products based solely on their physical characteristics, SIC and NAICS are classifications of business activities that incorporate product characteristics as well as the type of economic activity. SIC codes were used to classify U.S. economic activity until the Census Bureau's 1997 economic census, with major revisions of the SIC occurring in 1972 and 1987. Starting with the 1997 census, U.S. economic activity is classified according to the NAICS, which is standardized for the first five digits across the U.S., Canada and Mexico.

Census refers to the first four digits of an SIC code, and the first six digits of a NAICS code, as an industry. It reserves the terms product class and product for the first five and

Table 1: Import HS Sections and Chapters

Section	Name	HS Chapters
1	Live Animals; Animal Products	1-5
2	Vegetable Products	6-14
3	Animal or Vegetable Fats and Oils	15
4	Prepared Foodstuffs; Beverages, Spirits, Tobacco	16-24
5	Mineral Products	25-27
6	Products of the Chemical or Allied Industries	28-38
7	Plastics, Rubber and Articles Thereof	39-40
8	Raw Hides, Skins, Leather	41-43
9	Wood and Articles of Wood	44-46
10	Pulp of Wood, Paper	47-49
11	Textile and Textile Articles	50-63
12	Footwear, Headgear, etc.	64-67
13	Articles of Stone, Plaster, Cement, Ceramics, Glass	68-70
14	Pearls, precious stones, precious metals	71
15	Base Metals and Articles of Base Metal	72-83
16	Machinery, Appliances, Electrical Equipment	84-85
17	Vehicles, Aircraft, Vessels	86-89
18	Precision Instruments	90-92
19	Arms and Ammunitions	93
20	Misc. Manufactured Articles	94-96
21	Works of Art	97
22	Special Classification Provisions	98-99

Notes: This table displays sections and chapters U.S. Import HS Codes. Section names have been shortened for brevity. See the website of the U.S. International Trade Commission for full section names.

seven digits of an SIC code, and the first seven and ten digits of a NAICS code, respectively. While the set of official U.S. industries is defined outside the Census Bureau, Census generally has discretion in defining product classes and products within these industries. The primary economic activity classifications for both SIC and NAICS are provided in Table 2.

There are a number of differences between SIC and NAICS. First, NAICS provides more granular industry definitions than SIC, with the movement from 1,004 industries in SIC compared to 1, 170 industries in NAICS in 1997. Second, some activities were completely reclassified in the switch from SIC to NAICS, such as printing and publishing, which was reclassified from manufacturing (SIC 27) to wholesale trade (NAICS 51).

2.3. Some Complications Associated With Mapping HS to SIC/NAICS

As mentioned above, the HS and SIC/NAICS systems are fundamentally different in that the HS classifies products based solely on physical characteristics, while SIC and NAICS incorporate physical product characteristics as well as the type economic activity. This difference between the two systems can perhaps be most easily seen through a specific example. In the 1992 Schedule B codes used to classify U.S. exports, HS code 7215200000 tracks exports of "other bars and rods of iron or nonalloy steel, cold-formed or cold-finished, less than 0.25 percent carbon." While this definition is based solely on physical characteristics, the SIC/NAICS product classes to which it matches also take into account the

Table 2: Import NAICS and SIC Categories

NAICS Categories	Description	SIC Categories	Description
11	Agriculture, Forestry, Fishing and Hunting	01-09	Agriculture, Forestry, Fisheries
21	Mining, Quarrying, and Oil and Gas Extraction	10-14	Mineral Industries
22	Utilities	15-17	Construction Industries
23	Construction	20-39	Manufacturing
31-33	Manufacturing	41-49	Transportation, Communication, Utilities
42	Wholesale Trade	50-51	Wholesale Trade
44-45	Retail Trade	52-59	Retail Trade
48-49	Transportation and Warehousing	60-67	Finance, Insurance and Real Estate
51	Information	70-89	Service Industries
52	Finance and Insurance	91-97	Public Administration
53	Real Estate and Rental and Leasing		
54	Professional, Scientific, and Technical Services		
55	Management of Companies and Enterprises		
56	Administrative and Support and Waste Management and Remediation Services		
61	Educational Services		
62	Health Care and Social Assistance		
71	Arts, Entertainment, and Recreation		
72	Accommodation and Food Services		
81	Other Services (except Public Administration)		
92	Public Administration		

Notes: Table displays the primary categories of economic activity in the NAICS and SIC classification systems.
Source: U.S. Census Bureau.

method of production. In particular, this HS10, maps to two separate SIC5 product classes, 33128—"cold-finished steel bars/bar shapes (made in mills)—and 33168—"cold-finished steel bars/bar shapes (not made in mills).

The switch from SIC to NAICS for classifying domestic production also complicates matters. Because international trade data are reported in SIC format only for the years 1989-2001 and in NAICS format only for the years 2000 to 2009, researchers have been unable to construct a long time series spanning SIC and NAICS years. The concordances provided in this table allow applied economists to construct these long time series for the years1989-2009 for NAICS and 1989-2006 for SIC.

Lastly, HS codes are continually revised over time. Changes to the U.S. import or export codes occur via three routes: changes by the World Customs Organization (WCO) to the official list of international six-digit prefixes; U.S. legislation that affects U.S. eight-digit codes (imports only); or changes by the Committee for Statistical Annotation of Tariff Schedules (known as the "484(f) Committee") to statistical ten-digit codes.[3] For more information on changes in HS codes over time, including a concordance tracking these changes, see Pierce and Schott (forthcoming).

[3]See http://www.census.gov/foreign-trade/aip/comb_seminar_pres.ppt, and www.census.gov/foreign-trade/faq/sb/sb0008.html for more detail.

3. Concording HS to SIC4/NAICS6 Industries

As described above, empirical researchers have been hampered by an inability to generate long time series of industry-level international trade data and domestic production spanning SIC and NAICS years. This section describes an algorithm and concordances that we create, which link international trade and domestic economic activity data for the years 1989-2009. The concordances can be used to construct comparable datasets of international trade and domestic production data for longer time series than have previously been available.

The source data for the concordances is found in the monthly trade data published in CD format by Census's Foreign Trade Division.[4] Each of the monthly CDs for imports and exports contains a dBase-formatted file (called concord.dbf) that separately matches the ten-digit import and export HS codes used in the month to four-digit SIC and/or six-digit NAICS codes. We refer to these four-digit SIC and six-digit NAICS codes as "baseroots" for reasons discussed in the next section, but they are almost always proper industries.[5] Note that the December CD for each year contains annual, as well as monthly totals.

From 1989 to 2001, the mappings provided by Census match ten-digit HS codes to four-digit SIC baseroots. From 2000 to the present, they match ten-digit HS codes to six-digit NAICS baseroots. But for certain applications, it might be useful to extend each set of mappings beyond the years for which these official concordances are available. That is, it may be useful to have an HS-NAICS6 concordance for years prior to 2000 or an HS-SIC4 concordance for years after 2001.

We extend the HS-NAICS6 mappings to cover the period from 1989-2009 and the HS-SIC4 mappings for the years 1989-2006 using a three-step algorithm based on the procedures used previously in Feenstra et al. (2002). The algorithm is implemented on a "master list" of concordances assembled by appending the HS-baseroot mappings contained in the annual December trade CDs for the years 1989-2009. Note that we do not provide HS-SIC4 mappings for years after 2006 because the number of SIC4 codes that need to be assigned by hand-step 3 in the algorithm-rises to a level that makes the mapping less reliable, in our view.

The Stata code for steps 1 and 2, and for incorporating the results of step 3, is available in Appendix 1 under filename schott_algorithm_20.do.[6] The Stata code for the algorithm was created using Intercooled Stata, version 9.2 on a 2.0 GHz T2700 Intel Core 2 CPU. The steps of the algorithm are described immediately below.

1. Step 1 (Mechanical Match 1): Examine all ten-digit HS within a nine-digit category. If all assigned ten-digit HS within this category have the same NAICS6 (SIC4) assignment, assign that NAICS6 (SIC4) to any unassigned ten-digit HS within that nine-digit

[4] CDs are available starting in December, 1989 for exports and January 1989 for imports. The CDs are available for purchase from Census and are often also available in university libraries. The copies used here are provided generously by the Yale University Social Sciences Library.

[5] Of the 461 NAICS baseroots in the HS-NAICS6 import concordance and 455 NAICS baseroots in the HS-NAICS6 export concordance, 10 are not real industries as defined in the NAICS. They are 11211X, 1123XX, 31131X 31181X, 31511X, 33631X, 910000, 920000, 980000, 990000. Of the 471 SIC baseroots in the HS-SIC import concordance, 5 are not real industries as defined in the SIC. They are 314X, 9100, 9200, 9800, 9900. Of the 470 SIC baseroots in the HS-SIC export concordance, 7 are not real industries. They are 314X, 3XXX, 9000, 9100, 9200, 9800, 9900.

[6] The file is also available electronically on Schott's website: http://www.som.yale.edu/faculty/pks4/sub_international.htm

category. Repeat for eight-, seven-, etc. digit HS categories.

2. Step 2 (Mechanical Match 2): Sort list by ten-digit HS code. Examine "gaps" consisting of HS codes, or groups of consecutive codes that have not been matched to a baseroot. If a gap is preceded and succeeded by the same NAICS6 (SIC4) code, use that NAICS6 (SIC4) code for all unassigned ten-digit HS codes in the gap.

3. Step 3 (Hand Matching): Hand match remaining unmatched HS codes where possible. Note that any remaining unmatched ten-digit HS codes account for a very small fraction of U.S. imports or exports.

Tables 3 and 4 summarize the number of HS codes assigned using this procedure with SIC4 codes for years after 2001 and NAICS6 codes for years before 2000, respectively. The descriptions in the "source" column match those provided by the variable "matchtype" in the files described in Appendix 2 below.

Table 3: Extending the HS-SIC4 Concordance

	Source	2002	2003	2004	2005	2006
Import HS Codes	From Census	16,043	15,989	15,915	15,854	15,805
	From Mechanical Match 1	1,101	1,184	1,289	1,377	1,469
	From Mechanical Match 2	209	215	216	218	235
	From Hand Match	293	300	308	308	355
Export HS Codes	From Census	7,912	7,886	7,883	7,856	7,853
	From Mechanical Match 1	752	768	773	839	843
	From Mechanical Match 2	132	132	134	134	134
	From Hand Match	151	151	150	150	150

Notes: This table displays the method used to assign SIC codes to HS codes for years after 2001, when Census stopped reporting HS-SIC matches.

Table 4: Extending the HS-NAICS6 Concordance

	Source	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Import HS Codes	From Census	10,464	11,293	11,630	11,874	12,057	12,834	14,964	16,345	16,853	16,926	17,110
	From Mechanical Match 1	2,991	2,981	2,862	2,711	2,599	2,407	1,237	394	354	132	66
	From Mechanical Match 2	340	317	297	263	258	220	137	68	58	23	2
	From Hand Match	607	623	625	582	588	519	292	75	80	18	1
Export HS Codes	From Census	6,750	6,874	7,117	7,117	7,200	7,338	7,468	8,250	8,478	8,592	8,620
	From Mechanical Match 1	827	769	676	673	658	624	592	228	69	14	6
	From Mechanical Match 2	143	139	137	137	137	120	99	54	19	9	0
	From Hand Match	188	189	180	180	172	157	149	61	43	5	0

Notes: This table displays the method used to assign NAICS codes to HS codes for years before 2000, the year in which Census began reporting HS-NAICS matches.

By aggregating the HS-baseroot mappings for all available years and extending them for the full period in which the HS was in existence, we create HS to SIC4 and HS to NAICS6 concordances for both imports and exports, for the period from 1989 to 2006 and 1989 to 2009, respectively. See Appendix 2 for a full description of the final concordance files available in the electronic appendix to this paper.

4. Concording HS to SIC5/NAICS7 Product Classes

4.1. Census's Procedure for Mapping HS to SIC and NAICS

Researchers in international trade and industrial organization have recently begun studying the role of changes in product mix on plant and firm-level performance, as well as examining how exposure to international trade can affect firms' product mix. Examples of this research include Bernard, Redding and Schott (2010), Pierce (2011), Bernard, Redding and Schott (2011) and Goldberg, Khandelwal, Pavcnik and Topalova (Forthcoming). With this growing interest in product-level data, it is increasingly important to be able to match international trade and domestic production data at a highly disaggregated level. This section describes the construction of concordances that match ten-digit HS codes to five-digit SIC and seven-digit NAICS product classes–a more disaggregated level than has previously been available to researchers.

The primary bridge between HS and SIC (NAICS) product classes is a code referred to by the Census Bureau as a "SIC-base" ("NAICS-base"), which we refer to generically as "basecodes."[7] Basecodes are eight-digit alphanumeric codes that can generally be thought of as describing product characteristics. The first four (six) digits of the SIC (NAICS) basecode represent the "root" industry of the basecode. We refer to basecode roots here as "baseroots" and use them in constructing the industry concordances, as described in the preceding section. The remaining digits are internal identifiers for whether the basecode encompasses one or more product classes, and, in the latter instance, whether those product classes are from different industries. For Census year 1992, enough data are available for us to also construct a HS to SIC5/NAICS7 concordance based on basecodes. For the 1997, 2002 and 2007 HS to SIC5/NAICS7 concordances, however, we are restricted by data limitations to matching HS and SIC/NAICS product codes through baseroots only. The differences between constructing concordances using basecodes versus baseroots are discussed in detail in the next sub-section.

We match HS product codes to SIC5/NAICS7 product classes via baseroots using two complementary mappings produced by Census. The first mapping, which we refer to here as an "HS-baseroot" concordance, assigns a single baseroot to each HS code. As noted above, these mappings are published in Census's monthly releases of U.S. trade data on CDs. The second mapping is known as the principle differences (PD) file, which is constructed for every economic census in years ending in 2 and 7. The PD file assigns a single baseroot to each product class in the SIC or NAICS. HS product codes can then be matched to SIC5/NAICS7 product classes through their baseroots. The HS-baseroot and PD mappings are discussed in detail in Appendixes 4 and 5 below, respectively.

At this point, an example may be useful for fixing ideas. In 1992, HS code 7215200000 was used to track exports of "other bars and rods of iron or nonalloy steel, cold-formed or cold-finished, less than 0.25 percent carbon." According to the 1992 HS-baseroot concordance, this HS code – and 222 others – maps into SIC baseroot 3312. This baseroot, in turn, maps into 11 different SIC product class codes from 3 different four-digit SIC industries in the 1992 PD file: 33121, 33122, 33123, 33124, 33126, 33127, 33128, 3312C, 33167, 33168 and

[7] A more detailed discussion of Census' SIC and NAICS concordance methods is available at www.census.gov/epcd/www/intronet.html.

33170.[8] We note that in the official Census ten-digit HS to four-digit SIC mapping discussed in Section 3, HS code 7215200000 maps uniquely to SIC industry 3312.

This example highlights the "many-to-many" nature of the HS-product-class concordances. While each HS code maps to a single baseroot, many HS codes (223 in this example) can map to a single baseroot. Similarly, while each five-digit SIC product class maps to a single baseroot, many product classes (from three different industries, in this example) may map to a single baseroot. As discussed in Section 5 below, the HS-baseroot and PD files can be used to match the product classes U.S. manufacturing firms produce in each CMF year to the products they import and export in those years.

4.2. Matching on Basecodes Versus Baseroots

Matching on baseroots is appealing because HS-baseroot mappings are available in all years, allowing us to create concordances for every economic census year since 1992. As noted above, however, we have access to more disaggregated HS-basecode and SIC5-basecode mappings for 1992. The primary advantage of concordances using basecodes is a "more precise" mapping between HS and SIC5.

To illustrate what we mean by "more precise", consider once again HS code 7215200000, which we used to illustrate matching through baseroots in the previous sub-section. As mentioned above, HS code 72150200000 and 222 other HS codes matched to 11 different SIC5 product classes through baseroot 3312. When HS and SIC5 codes are matched through full basecodes, rather than baseroots, however, we find that 7215200000 is one of only 10 HS codes that map to only two SIC5s – 33128 and 33168, defined under basecode 33128B00 – described as "cold-finished steel bars/bar shapes (made in mills)" and "cold-finished steel bars/bar shapes (not made in mills)."[9] Because HS code 7215200000 is described as "other bars and rods of iron or nonalloy steel, cold-formed or cold-finished, less than 0.25 percent carbon", it appears that assigning SIC5s based on a full basecode, rather than a baseroot, has provided a better match, by dropping unrelated SIC5 products like sheet and strip, pipe and tube and rails.

HS code 7215200000 was matched to 9 additional SIC5s when we matched HS codes to SIC5 codes with baseroots versus basecodes. This matching of HS codes to additional SIC5s when matching with baseroots is not uncommon, as illustrated in the following analysis of 1992, the only year for which we can do both types of mappings. Of the 16,022 import HS codes in use in 1992, 9,289 are matched to additional SIC5s when using baseroot matching. The mean number of additional SIC5s matched to each import HS is 2.35. Similarly, of the 8,054 export HS codes in use in 1992, 5,396 are matched to extra SIC5s under baseroot

[8] The product descriptions for these SIC5 product-classes are as follows: 33121 - Coke over and blast furnace products; 33122 - Steel ingot and semifinished shapes; 33123 - Hot-rolled sheet and strip including tin-milled products; 33124 - Hot-rolled bars and bar shapes, plates, structural; 33126 - Steel pipe and tubes (made in steel mills); 33127 - Cold-rolled steel sheet and strip (made in mills); 33128 - Cold-finished steel bars/bar shapes (made in mills); 3312C - Other steel mill products, including steel rails; 33167 - Cold-rolled steel sheet and strip (not made in mills); 33168 - Cold-finished steel bars and bar shapes (not made in mills); 33170 - Steel pipe and tubes.

[9] Note that this example provides a good illustration of how HS codes may match to more than one SIC5, since the SIC considers the method of production when assigning product classifications. The difference between these product classes is whether or not they are made in steel mills.

matching. The mean number of additional SIC5s matched to each export HS is 2.72. Table 5 displays the number of extra SIC5s associated with HS10 import and export codes for 1992.

Table 5: Additional SIC5s Associated With Each HS Under Baseroot Matching

Exports		Imports	
HS10	Additional SIC5	HS10	Additional SIC5
2,658	0	6,733	0
1,090	1	1,818	1
1,102	2	1,983	2
742	3	1,294	3
731	4	1,225	4
389	5	618	5
181	6	289	6
400	7	819	7
283	8	368	8
212	9	402	9
42	10	107	10
25	11	38	11
74	12	112	12
50	13	75	13
33	14	60	14
23	16	57	16
11	17	13	17
1	20	1	20
3	24	6	24
4	27	4	27

Notes: Table displays the number of "Additional" SIC5s associated with HS10 export and import codes in 1992. Additional SIC5s are SIC5 product codes that are associated with a particular HS10 when a concordance is constructed with 4-digit baseroots, rather than a full 8-digit basecode.

For some types of research, matching HS and SIC5/NAICS7 codes through full basecodes might be useful. Pierce (2011), for example, identifies U.S. manufacturing establishments that received antidumping protection by matching the HS10s used to classify products in antidumping investigations to the SIC5 product-classes that establishments reported producing in the CMF. In this case, matching on baseroots, rather than full basecodes, would likely lead to some unprotected plants being incorrectly identified as recipients of antidumping protection.

Unfortunately, Census published a full HS10-basecode mapping only for 1992. As a result, matching on basecodes can only be performed in a somewhat limited time period. In the electronic appendix, we provide HS10 to SIC5 concordances constructed with basecode matching for 1992 with filenames m_basecode_92.csv and x_basecode_92.csv for imports and exports, respectively. In the import concordance, 16,022 HS codes are matched to 1,564 SIC5 codes through 812 basecodes.[10] In the export concordance 8,053 HS codes are matched to 1,555 SIC5 codes through 806 basecodes.[11]

[10] Six-hundred twenty-five import HS codes have basecodes with no SIC5 match. Two SIC5 codes have basecodes with no import HS match.

[11] Four-hundred eight export HS codes have basecodes with no SIC5 match. Eleven SIC5 codes have basecodes with no export HS match.

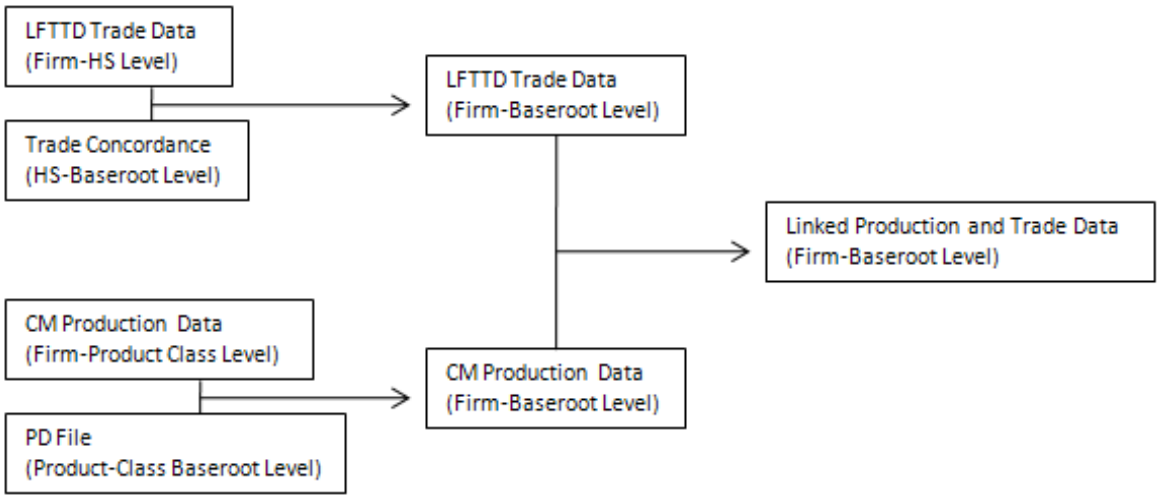

Figure 1: Linking the LFTTD to the CMF at the Firm-Baseroot-Level

5. Linking the LFTTD and CM

As mentioned above, a large new literature has grown around examining changes in product mix at firms and plants, and especially how those changes are related to international trade. This brief section illustrates how the concordances generated above can be used to created a firm-baseroot-level dataset of trade and production. Firm-level trade data for every U.S. importer and exporter are located in Census's Longitudinal Firm Trade Transactions Database, which is described in detail in Bernard, Jensen and Schott (2009). Firm-product-level domestic production data for every U.S. manufacturer are from the product trailer data of Census's Census of Manufactures (CM). Once these datasets are merged, researchers will possess a firm-baseroot-level dataset recording production, imports and exports at the same level of aggregation (i.e., according to SIC or NAICS baseroots) for a particular census year. This merged dataset will then greatly increase researchers' ability to understand changes in firms' product mix over time.

The merged trade and production dataset can be constructed relatively simply, as follows. First, the international trade data in the LFTTD are merged with the trade concordance described in Section 3 by HS code and year, yielding data on the full set of baseroots imported and exported by U.S. firms. Then, the product trailer files of the CM–which contain data on output by product for every U.S. manufacturing establishment–are merged with the PD file for the appropriate year, and aggregated to the firm-baseroot-level. Lastly, these two datasets are merged by baseroot. The resulting dataset contains information on the value of shipments, imports and exports for every U.S. manufacturing firm in SIC or NAICS format. This process is illustrated in Figure 1.

6. Conclusion

While empirical economists increasingly study the relationship between international trade and domestic economic activity, research has been slowed due to gaps in these datasets. This paper creates an algorithm and provides sets of concordances linking the ten-digit HS codes used by the United States to track international trade with the SIC and NAICS categories used to characterize domestic economic activity. Through the use of these concordances it is now possible to create linked datasets of trade and domestic production in both SIC and NAICS from 1989-2009 and to link trade and production data at a more disaggregated level than is typically available. In addition, we provide concordances linking ten-digit HS codes to five-digit SIC and seven-digit NAICS product classes. These concordnaces then allow researchers studying the product-switching behavior of U.S. firms to match trade and domestic production data at a more disaggregated level than was previously available.

References

A.B. Bernard, J.B. Jensen and P.K. Schott (2009), "Importers, Exporters and Multinationals: A Portrait of Firms in the U.S. that Trade Goods" in T. Dunne, J.B. Jensen and M.J. Roberts (eds.), Producer Dynamics: New Evidence from Micro Data (University of Chicago Press).

A.B. Bernard, S.J. Redding and P.K. Schott (2010), "Multi-Product Firms and Product Switching," *American Economic Review*, 100:70-97.

A.B. Bernard, S.J. Redding and P.K. Schott (2011), "Multi-Product Firms and Trade Liberalization," *Quarterly Journal of Economics*, 126(3), 1271-1318.

R.C. Feenstra. (1996), "U.S. Imports, 1972-1994: Data and Concordances," NBER Working Paper no. 5515.

R.C. Feenstra, J. Romalis and P.K. Schott. (2002), "U.S. Imports, Exports and Tariff Data, 1989 to 2001," NBER Working Paper 9387.

P. Goldberg, A. Khandelwal, N. Pavcnik and P. Topalova. (Forthcoming), "Imported Intermediate Inputs and Domestic Product Growth: Evidence from India," *Quarterly Journal of Economics.*

J.R. Pierce and P.K. Schott (forthcoming), "Concording U.S. Harmonized System and Schedule B Codes Over Time," *Journal of Official Statistics.*

J.R. Pierce (2011), "Plant-Level Responses to Antidumping Duties: Evidence from U.S. Manufacturers," *Journal of International Economics*, 85(2), 222-233.

A.L. Revenga (1992), "Exporting Jobs? The Impact of Import Competition on Employment and Wages in U.S. Manufacturing," *Quarterly Journal of Economics* 107(1), 255-284.

J.D. Sachs and H.J. Shatz (1994), "Trade and Jobs in U.S. Manufacturing," *Brookings Papers on Economic Activity*, 1994(1), 1-69.

A Appendix 1: Stata Code

Contents of schott_algorithm_20.do:

```
**0 Prelim
clear
set more off
set mem 500m
**1 SIC Mapping
foreach zzz in exp imp {
    **1.1 read in the hs-sic mappings provided by census in its monthly trade cd files
    cd "C:\Users\pks4\Documents\My Dropbox\research\concordances\production\for_schott\"
    *create list of mappings
    use 'zzz'_concord_89_106, clear
    *keep latest year for which sic is available
    keep if year==101
    keep commodity sic
    drop if sic==""
    duplicates drop commodity, force
    sort commodity
    save temp0, replace

    *read in the list of raw hs10 export codes
    use 'zzz'_concord_89_106, clear
    *only need to match years in which sic data are not provided
    keep if year>101
    keep commodity
    duplicates drop commodity, force
    sort commodity
    merge commodity using temp0, keep(sic)
    tab _merge
    drop _merge
    destring commodity, force g(hs)
    egen sic87=group(sic)
    save 'zzz'temp_01, replace
    *save group-sic mapping for below
    use 'zzz'temp_01, clear
    collapse (mean) sic87, by(sic)
    rename sic87 sic87_new1
    rename sic sic_new1
    drop if sic_new1=="" | sic87_new1==.
    sort sic87_new1
    save temp1, replace

    use 'zzz'temp_01, clear
    collapse (mean) sic87, by(sic)
```

```
rename sic87 sic87_new2
rename sic sic_new2
drop if sic_new2=="" | sic87_new2==.
sort sic87_new2
save temp2, replace
**1.2 First Mechanical Match
**Create new matches mechanically by looking to see what the already-matched sic look like.
**Look at all hs9 to see what sic87 the already-matched have; if unanimous, use that. If not,
**go up one level. and so on.
use `"zzz"'temp_01, clear
gen sic87_new1 = sic87
sum hs sic87*
quietly {
foreach x in 9 8 7 6 5 4 3 2 {
    noisily display [`"x"']
    local y = 10-`"x"'
    gen hs`"x"' = int(hs/(10^`"y"'))
    egen t1 = mean(sic87), by(hs`"x"')
    egen t2 = sd(sic87), by(hs`"x"')
    egen t3 = count(sic87), by(hs`"x"')
    gen sic87_`"x"' = t1 if t2==0 | t3==1
    replace sic87_new1 = sic87_`"x"' if sic87==. & sic87_new1==.
    drop t1 t2 t3
    drop hs`"x"' sic87_`"x"'
}
}
sum hs sic87 sic87_new1
sort hs
save `"zzz"'temp_02, replace

**1.3 Second Mechanical Match
**Look at gaps. If last known and next know are the same, use them to fill in.
use `"zzz"'temp_02, clear
gen sic87_new2 = sic87_new1
gen begin = 1 if sic87_new1==. & sic87_new1[_n-1]~=.
gen end = sic87_new1==. & sic87_new1[_n+1]~=.
gen bsum = sum(begin)
gen gap = sic87_new1==.
replace bsum=. if gap==0
gen sb = sic87_new1[_n-1]*begin
gen se = sic87_new1[_n+1]*end
egen tb = mean(sb), by(bsum)
egen te = mean(se), by(bsum)
```

```
        gen match = tb==te
        replace sic87_new2 = tb if match==1 & sic87_new1==.
        sum hs sic87*
        drop begin end bsum gap sb se tb te match
        sort hs
        save 'zzz'temp_03, replace

        *1.4 Recover groups from above
        use 'zzz'temp_03, clear
        sort sic87_new1
        merge sic87_new1 using temp1, keep(sic_new1)
        tab _merge
        drop _merge
        sort sic87_new2
        merge sic87_new2 using temp2, keep(sic_new2)
        tab _merge
        drop _merge
        sort hs
        gen t=sic87_new1~=.
        tab t
        drop t
drop sic87*
        format hs %15.0g
        drop if hs<100
        save 'zzz'_concord_89_106_sicfillin, replace
}
**2 naics
foreach zzz in exp imp {
        **2.1 read in the hs-sic mappings provided by census in its monthly trade cd files

        *create list of mappings
        use 'zzz'_concord_89_106, clear
        *keep earliest year for which naics is available
        keep if year==100
        keep commodity naics
        drop if naics==""
        duplicates drop commodity, force
        sort commodity
        save temp0, replace

        *read in the list of raw hs10 export codes
        use 'zzz'_concord_89_106, clear
        *Only need years for which there is no naics
        keep if year<100
        keep commodity
```

```
duplicates drop commodity, force
sort commodity
merge commodity using temp0, keep(naics)
tab _merge
drop _merge
destring commodity, force g(hs)
egen naics87=group(naics)
save ‘zzz’temp_01, replace
*save group-naics mapping for below
use ‘zzz’temp_01, clear
collapse (mean) naics87, by(naics)
rename naics87 naics87_new1
rename naics naics_new1
drop if naics_new1=="" | naics87_new1==.
sort naics87_new1
save temp1, replace

use ‘zzz’temp_01, clear
collapse (mean) naics87, by(naics)
rename naics87 naics87_new2
rename naics naics_new2
drop if naics_new2=="" | naics87_new2==.
sort naics87_new2
save temp2, replace
**2.2 First Mechanical Match
**Create new matches mechanically by looking to see what the already-matched naics look like.
**Look at all hs9 to see what naics87 the already-matched have; if unanimous, use that. If not,
**go up one level. and so on.
use ‘zzz’temp_01, clear
gen naics87_new1 = naics87
sum hs naics87*
quietly {
foreach x in 9 8 7 6 5 4 3 2 {
    noisily display [‘x’]
    local y = 10-‘x’
    gen hs‘x’ = int(hs/(10^‘y’))
    egen t1 = mean(naics87), by(hs‘x’)
    egen t2 = sd(naics87), by(hs‘x’)
    egen t3 = count(naics87), by(hs‘x’)
    gen naics87_‘x’ = t1 if t2==0 | t3==1
    replace naics87_new1 = naics87_‘x’ if naics87==. & naics87_new1==.
    drop t1 t2 t3
    drop hs‘x’ naics87_‘x’
```

```
	}
	}
	sum hs naics87 naics87_new1
	sort hs
	save 'zzz'temp_02, replace

	**2.3 Second Mechanical Match
	**Look at gaps. If last known and next know are the same, use them to fill in.
	use 'zzz'temp_02, clear
	gen naics87_new2 = naics87_new1
	gen begin = 1 if naics87_new1==. & naics87_new1[_n-1]~=.
	gen end = naics87_new1==. & naics87_new1[_n+1]~=.
	gen bsum = sum(begin)
	gen gap = naics87_new1==.
	replace bsum=. if gap==0
	gen sb = naics87_new1[_n-1]*begin
	gen se = naics87_new1[_n+1]*end
	egen tb = mean(sb), by(bsum)
	egen te = mean(se), by(bsum)
	gen match = tb==te
	replace naics87_new2 = tb if match==1 & naics87_new1==.
	sum hs naics87*
	drop begin end bsum gap sb se tb te match
	sort hs
	save 'zzz'temp_03, replace

	*2.4 recover groups from above
	use 'zzz'temp_03, clear
	sort naics87_new1
	merge naics87_new1 using temp1, keep(naics_new1)
	tab _merge
	drop _merge
	sort naics87_new2
	merge naics87_new2 using temp2, keep(naics_new2)
	tab _merge
	drop _merge
	sort hs
	gen t=naics87_new1~=.
	tab t
	drop t
drop naics87*
	format hs %15.0g
	drop if hs<100
	save 'zzz'_concord_89_106_naicsfillin, replace
}
```

```
**3 Add in hand matches to imports and exports, respectively, first for sic and then for naics
** Any missing matches after the last section were matched by hand by kitjawat. Add these
** hand matches into the data here and then also create a variable that identifies each
*** mapping according to whether it is from Census, mechanical match 1, mechanical match 2 or
** from kitjawat's hand matching.
**
** 2009.10.16 change sic 2612 to 2621 in kitjawat_handmatch_imports_sic_20080821 per Justin's email
** also add leading zero to sic's from handmatch and fix missing naics for 1605106000
**
use imp_concord_89_106_sicfillin, clear
sort hs
merge hs using kitjawat_handmatch_imports_sic_20080821
tab _merge
drop if _merge==2
replace kitjawat = 2621 if kitjawat==2612
drop _merge
gen sic_new3=sic_new2
tostring kitjawat, g(kitjawats)
replace kitjawats = "0"+kitjawats if kitjawat>=100 & kitjawat<=999
replace sic_new3=kitjawats if sic_new3=="" & kitjawats!=""
replace sic_new3="" if sic_new3=="."
sort hs
merge hs using sic_imp_jrp
tab _merge
replace sic_new3=sic_new4 if sic_new3=="" & sic_new4!=""
codebook sic_new3
gen id = "From Census"
gen newsic = sic
replace id = "From mechanical match 1" if sic==""
replace newsic = sic_new1 if sic==""
replace id = "From mechanical match 2" if newsic==""
replace newsic = sic_new2 if newsic==""
replace id = "From hand match" if newsic==""
replace newsic = kitjawats if newsic==""
label var id "SIC match type"
keep commodity hs newsic id
rename newsic sic
rename id sic_matchtype
rename sic new_sic
keep commodity new_sic sic_matchtype
order commodity new_sic sic_matchtype
```

```
sort commodity
save sic_imp_final, replace
use imp_concord_89_106_naicsfillin, clear
sort hs
merge hs using kitjawat_handmatch_imports_naics_20081016
tab _merge
drop if _merge==2
drop _merge
gen naics_new3=naics_new2
tostring kitjawat, g(kitjawats)
replace kitjawats = "311711" if commodity=="1605106000"
replace naics_new3=kitjawats if naics_new3=="" & kitjawats!=""
replace naics_new3="" if naics_new3=="."
sort hs
merge hs using naics_imp_jrp
tab _merge
replace naics_new3=naics_new4 if naics_new3=="" & naics_new4!=""
codebook naics_new3
gen id = "From Census"
gen newnaics = naics
replace id = "From mechanical match 1" if naics==""
replace newnaics = naics_new1 if naics==""
replace id = "From mechanical match 2" if newnaics==""
replace newnaics = naics_new2 if newnaics==""
replace id = "From hand match" if newnaics==""
replace newnaics = kitjawats if newnaics==""
label var id "NAICS match type"
drop naics
rename newnaics naics
rename id naics_matchtype
rename naics new_naics
keep commodity new_naics naics_matchtype
order commodity new_naics naics_matchtype
sort commodity
save naics_imp_final, replace
use exp_concord_89_106_sicfillin, clear
sort hs
merge hs using kitjawat_handmatch_exports_sic_20080821
tab _merge
drop if _merge==2
drop _merge
gen sic_new3=sic_new2
tostring kitjawat, g(kitjawats)
replace kitjawats = "0"+kitjawats if kitjawat>=100 & kitjawat<=999
replace sic_new3=kitjawats if sic_new3=="" & kitjawats!=""
```

```
replace sic_new3="" if sic_new3=="."
sort hs
merge hs using sic_exp_jrp
tab _merge
replace sic_new3=sic_new4 if sic_new3=="" & sic_new4!=""
codebook sic_new3
gen id = "From Census"
gen newsic = sic
replace id = "From mechanical match 1" if sic==""
replace newsic = sic_new1 if sic==""
replace id = "From mechanical match 2" if newsic==""
replace newsic = sic_new2 if newsic==""
replace id = "From hand match" if newsic==""
replace newsic = kitjawats if newsic==""
label var id "SIC match type"
drop sic
rename newsic sic
rename id sic_matchtype
rename sic new_sic
keep commodity new_sic sic_matchtype
order commodity new_sic sic_matchtype
sort commodity
save sic_exp_final, replace
use exp_concord_89_106_naicsfillin, clear
sort hs
merge hs using kitjawat_handmatch_exports_naics_20081016
tab _merge
drop if _merge==2
drop _merge
gen naics_new3=naics_new2
tostring kitjawat, g(kitjawats)
*replace kitjawats = "0"+kitjawats if kitjawat>=100 & kitjawat<=999
replace naics_new3=kitjawats if naics_new3=="" & kitjawats!=""
replace naics_new3="" if naics_new3=="."
sort hs
merge hs using naics_exp_jrp
tab _merge
replace naics_new3=naics_new4 if naics_new3=="" & naics_new4!=""
codebook naics_new3
gen id = "From Census"
gen newnaics = naics
replace id = "From mechanical match 1" if naics==""
replace newnaics = naics_new1 if naics==""
replace id = "From mechanical match 2" if newnaics==""
replace newnaics = naics_new2 if newnaics==""
```

```
replace id = "From hand match" if newnaics==""
replace newnaics = kitjawats if newnaics==""
label var id "NAICS match type"
drop naics
rename newnaics naics
rename id naics_matchtype
rename naics new_naics
keep commodity new_naics naics_matchtype
order commodity new_naics naics_matchtype
sort commodity
save naics_exp_final, replace
**4 Reassemble HS-SIC data for all years
*Imports
use imp_concord_89_106, clear
sort commodity
merge commodity using sic_imp_final
tab _merge
drop _merge
replace sic_matchtype="From Census" if sic!=""
replace sic=new_sic if sic=="" & new_sic!=""
sort commodity
merge commodity using naics_imp_final
tab _merge
drop _merge
replace naics_matchtype="From Census" if naics!=""
replace naics=new_naics if naics=="" & new_naics!=""
drop new* descrip*
destring commodity, g(hs) force
append using imp_107_concord
append using imp_108_concord
append using imp_109_concord
order commodity hs year sic sic_matchtype naics naics_matchtype
sort commodity year
save hs_sic_naics_imports_89_109_20111004, replace
outsheet using hs_sic_naics_imports_89_109_20111004.csv, replace
*Exports
use exp_concord_89_106, clear
sort commodity
merge commodity using sic_exp_final
tab _merge
drop _merge
replace sic_matchtype="From Census" if sic!=""
replace sic=new_sic if sic=="" & new_sic!=""
sort commodity
merge commodity using naics_exp_final
```

```
tab _merge
drop _merge
replace naics_matchtype="From Census" if naics!=""
replace naics=new_naics if naics=="" & new_naics!=""
drop new* descrip*
destring commodity, g(hs) force
*This drops several special classification codes for U.S. goods returned from Puerto Rico
drop if hs<10
append using exp_107_concord
append using exp_108_concord
append using exp_109_concord
order commodity hs year sic sic_matchtype naics naics_matchtype
sort commodity year
save hs_sic_naics_exports_89_109_20111004, replace
outsheet using hs_sic_naics_exports_89_109_20111004.csv, replace
```

Contents of hs_sic5_basecodes_02.do:

```
clear
capture log close
set more off
set mem 1000m
log using full_conc_92.log, replace
use appndxd, clear
keep sicbase92 pc5
rename pc5 sic5
drop if sic5=="N/A"
sort sicbase92
save t1, replace
use hs_sic_m_allsources_1989_2006, clear
keep hs sicbase92
drop if sicbase92==""
sort sicbase92
joinby sicbase92 using t1, unmatched(both)
tab _merge
keep if _merge==3
drop _merge
rename sicbase92 basecode
sort hs
save m_basecode_92, replace
outsheet using m_basecode_92.csv, replace
use hs_sic_x_allsources_1989_2006, clear
keep hs sicbase92
drop if sicbase92==""
sort sicbase92
joinby sicbase92 using t1, unmatched(both)
tab _merge
```

```
keep if _merge==3
drop _merge
rename sicbase92 basecode
sort hs
save x_basecode_92, replace
outsheet using x_basecode_92.csv, replace
capture log close
```

Contents of hs_sic5_naics7_baseroots_04.do:

```
clear
set more off
set mem 1000m
cd "C:\Users\Justin\Documents\RA Work\Jensen_Schott_Bernard\hs_sic_naics_concordance"
capture log close
log using baseroot_conc_create.log, replace
foreach x in imports exports {
      use hs_sic_naics_'x'_89_106_20091016, clear
      keep commodity sic sic_matchtype
      order commodity sic
      keep commodity sic
      rename commodity hs
      rename sic sicbaseroot
      sort sicbaseroot
      save hs_sic_'x', replace
}
foreach x in imports exports {
      use hs_sic_naics_'x'_89_106_20091016, clear
      keep commodity naics naics_matchtype
      order commodity naics
      keep commodity naics
      rename commodity hs
      rename naics naicsbaseroot
      sort naicsbaseroot
      save hs_naics_'x', replace
}
foreach x in imports exports {
      use pd92, clear
      keep sicbase92 pc5
      drop if pc5=="N/A"
      gen sicbaseroot=substr(sicbase92,1,4)
      rename pc5 sic5
      keep sicbaseroot sic5
      sort sicbaseroot
      joinby sicbaseroot using hs_sic_'x', unmatched(both)
      tab _merge
      keep if _merge==3
```

```
        drop _merge
        order hs sic5
        save hs_sic5_‘x’_92, replace
        outsheet using hs_sic5_‘x’_92.csv, replace
    }
    foreach y in 97 02 {
        foreach x in imports exports {
            noisily display "‘x’‘y’"
            use pd‘y’, clear
            keep baseroot pc7
            drop if baseroot=="N/A"
            drop if pc7=="N/A"
            rename pc7 naics7
            rename baseroot naicsbaseroot
            sort naicsbaseroot
            joinby naicsbaseroot using hs_naics_‘x’, unmatched(both)
            tab _merge
            keep if _merge==3
            drop _merge
            order hs naics7
            save hs_naics7_‘x’_‘y’, replace
            outsheet using hs_naics7_‘x’_‘y’.csv, replace
        }
    }
    capture log close
```

B Appendix 2: Downloads

Downloads

All files described here are available in a zip archive accompanying this paper on Schott's website.[12]

B1. HS-SIC4/NAICS6 Concordance Files

The HS-NAICS6 (SIC4) industry concordances for 1989 to 2009 (1989 to 2006) are available in two files for exports and imports, named, respectively, hs_sic_naics_imports_89_109_20111004.dt and hs_sic_naics_exports_89_109_20111004.dta, where 89 represents the beginning year of 1989, 109 represents the ending year of 2009 and 20101220 represents the version date.

1. HS: ten-digit HS import or export code
2. SIC: corresponding four-digit SIC code
3. NAICS: corresponding six-digit NAICS code

[12] See http://www.som.yale.edu/faculty/pks4/sub_international.htm.

4. SIC_MATCHTYPE: description of match origin (see Table 3)
5. NAICS_MATCHTYPE: description of match origin (see Table 3)
6. COMMODITY: a string version of HS, with leading zeroes, where applicable

The Stata do-file used to create these concordances are also available in the electronic appendix with filename schott_algorithm_20.do.

B2. HS-SIC5 (1992, Using Full Basecodes)

The HS-SIC5 (basecode) concordances for 1992 are available in two files named m_basecode_92.csv for imports and x_basecode_92.csv for exports. These files contain the following variables:

1. HS: ten-digit HS import or export code
2. Basecode: eight-character basecode associated with HS
3. SIC5: The SIC5s associated with a particular HS and basecode.

Note that there may be multiple entries for a single HS code when it matches to more than one SIC5. The Stata do-file used to create these concordances is also available in the electronic appendix with hs_sic5_basecodes_02.do.

B3. HS-SIC5 (1992, Using Baseroots)

The HS-SIC5 (baseroot) concordances for 1992 are available in two files named hs_sic5_imports_92.csv for imports and hs_sic5_exports_92.csv for exports. These files contain the following variables:

1. HS: ten-digit HS import or export code
2. SICBASEROOT: four-character SIC baseroot associated with HS
3. SIC5: The SIC5s associated with a particular HS and basecode.

Note that there may be multiple entries for a single HS code when it matches to more than one SIC5. The Stata do-file used to create these concordances is also available in the electronic appendix with filename hs_sic5_naics7_baseroots_04.do.

B4. HS-NAICS7 (1997 and 2002, Using Baseroots)

The HS-NAICS7 (baseroot) concordances for 1997 and 2002 are available in four files named hs_naics7_imports_yy.csv for imports and hs_naics7_exports_yy.csv for exports, where yy is the last two digits of the year. These files contain the following variables:

1. HS: ten-digit HS import or export code
2. NAISBASEROOT: six-character NAICS baseroot associated with HS

3. NAICS7: The NAICS7 associated with a particular HS and basecode.

Note that there may be multiple entries for a single HS code when it matches to more than one NAICS7. The Stata do-file used to create these concordances is also available in the electronic appendix with filename hs_sic5_naics7_baseroots_04.

B5. HS-SITC Concordance Files

Census's mapping of HS and SITC codes from its published trade data are available in two files named hs_sitc_imports.csv for imports and hs_sitc_exports.csv for exports. These files contain the following variables:

1. HS: ten-digit HS import or export code

2. Corresponding five-digit revision 3 SITC code.

C Appendix 3: Other Concordances

This appendix discusses the relationship between the concordances developed above and two other HS-SIC/NAICS concordances that can be found on the web.

C1. The Feenstra (2002) Concordance

Feenstra et al. (2002) provide background for U.S. HS10-level trade data for 1989 to 2001. Those data have subsequently been extended to 2006 and are available on Feenstra's website. Of the 26,277 ten-digit HS codes used to track U.S. imports (exports) in the Feenstra et al. (2002) 1989 to 2001 dataset, Census provided a baseroot concordance for all but 1,222. Of these 1,222 HS codes, 898 were assigned to a four-digit SIC category using a HS to 1987-revision MSIC concordance from Feenstra (1996). Though in principle MSIC codes differ from SIC codes, a number of MSIC codes map directly into regular SIC codes. The remaining 324 products were assigned to industries via an algorithm similar to that described in Section 3 above.

The set of HS codes found in the Feenstra et al. concordances differs slightly from that of the master list described in Section 3. Of the 25,329 (11,509) unique import (export) HS codes that result from merging Feenstra et al.'s concordances with our own, we find that 24,947 (11,472) are in common while 382 (37) are only in the Feenstra et al (2002) concordance. We don't have an explanation for the codes unique to the Feenstra et al (2002) dataset though we suspect they may be due to Census' periodic revisions of the trade data.

C2. The EIIT Concordance

A five-digit SIC to ten-digit HS concordance of unknown origin is posted to the EIIT website.[13] This concordance does not distinguish between import or export HS categories and it does not note the years to which either its HS codes apply.

[13] See www.macalester.edu/research/economics/page/haveman/Trade.Resources/Concordances/FromHS/10hs5sic87.txt.

The EIIT concordance contains 17,436 HS codes and maps them to 805 five-digit SIC categories, 741 of which are in manufacturing. If collapsed to the four-digit SIC level, this list comprises 439 four-digit SIC codes, 386 of which are in manufacturing. This compares with the 1,440 five-digit and 459 four-digit manufacturing SIC codes contained in the 1987 revision of the SIC, and the 1,462 five-digit and 459 four-digit manufacturing SIC codes in the 1992 revision of the SIC. The 386 unique manufacturing codes in the EIIT concordance are similar to the 386 "super-sic" codes described in Feenstra et al (2002).

The EIIT concordance appears to be a close cousin of the concordance described in Section 4. Of the 8,215 (15,120) export (import) HS codes which appear in both concordances, 6,058 (10,762) have the same four-digit root.

D Appendix 4: Census's HS-Baseroot Concordances

Census produces an HS-basecode concordance only for the years in which there is an economic census. However, it provides more aggregate, HS-baseroot concordances with its monthly published trade statistics. Census constructs the HS-to-basecode and HS-baseroot concordances so that the Foreign Trade Division can publish trade statistics using the same industry categories it uses to publish domestic production statistics. As alluded to above and as discussed in more detail at www.censusbureau.biz/epcd/oei/view/appenda.txt, the HS to basecode mappings often make more sense for exports than for imports: "It is somewhat easier to find a reasonable statistical basis for comparing domestic output with exports than with imports. This is because there are substantial numbers of imported commodities which are not produced in the United States or are produced in very small quantities. On the other hand, the merchandise exported from the United States is ordinarily produced in this country and reflects items important to output."

As discussed above, we assemble a "master list" of these mappings by appending the HS-baseroot concordances contained in the December trade CD-roms. The Stata files containing these lists are discussed in Section B They are available on Schott's website.

D1. HS-SIC

Census's HS-baseroot concordances virtually always map HS codes to a single four-character SIC root. As noted above, these roots are the first four characters of an eight-character SIC basecode.[14] For the most part, these baseroots are proper industries, but there are some (e.g., 3XXX) that reflect the difficulties noted in Sections 3 and 4 above. We note the following:

- As indicated in Table 6, the number of unique HS export (import) codes in the master list that have SIC basecodes associated with them in at least one year ranges from 7,908 (14,402) in 1989 to 8,629 (17,183) in 2001. 2001 is the final year in which SIC codes appear in the concordance.

[14] Though the concordance files included with the monthly trade data do not include the full, internal-to-Census basecode, that mapping is available for 1992 at http://www.census.gov/epcd/www/intronet.html (see second paragraph).

- The number of unique SIC codes to which these export (import) HS codes match ranges from 429 (443) in 1989 to 449 (450) in 2001.[15]

- Some of the SIC basecodes to which HS codes are assigned are incomplete (e.g., 23XX), while others are outside manufacturing (e.g., 0273). As noted in the third column of each panel in Table 6, the number of manufacturing SIC basecodes to which these export (import) codes match ranges from 371 (386) in 1989 to 391 (392) in 2001. The fact that there are fewer than the official number of 459 manufacturing SIC codes in the concordance files is consistent with the discussion in Sections 3 and 4 above.

Table 6: HS and Four-Digit SIC Codes in the "Master List"

	Exports			Imports		
	HS10	SIC4	Man SIC4	HS10	SIC4	Man SIC4
1989	7,908	429	371	14,402	443	386
1990	7,971	447	387	15,214	446	387
1991	8,110	448	387	15,414	446	387
1992	8,107	448	387	15,430	448	388
1993	8,167	449	391	15,502	447	389
1994	8,239	449	391	15,980	447	389
1995	8,308	449	391	16,630	447	389
1996	8,593	449	391	16,882	447	389
1997	8,609	449	391	17,345	447	389
1998	8,620	449	391	17,099	447	389
1999	8,626	449	391	17,179	450	392
2000	8,635	449	391	17,215	450	392
2001	8,629	449	391	17,183	450	392

Notes: Table displays number of ten-digit HS codes, four digit SIC codes, and four-digit manufacturing SIC codes appearing in the concordance files accompanying the U.S. monthly trade statistics sold by the U.S. Census Bureau.

D2. HS-NAICS

As with the SIC, Census's concordance files virtually always map HS codes to a unique six-digit NAICS baseroot. For the most part, these baseroots are proper NAICS industries, but there are some that reflect the difficulties noted in Sections 3 and 4 above. We note the following:

- As summarized in Table 7, the number of HS export (import) codes in the master list that have NAICS basecodes associated with them in at least one year ranges from 8,628 (16,897) in 2000 to 8,882 (17,745) in 2009. 2000 is the first year that NAICS codes appear in the concordance files.

- The number of NAICS basecodes to which these export (import) codes match ranges from 454 in 2000 to 456 in 2009 for imports and switches between 453 and 454 for exports.[16]

[15] There are 459 "official" four-digit SIC manufacturing codes in 1992 and 1997 economic censuses. For a complete list, see http://www.censusbureau.biz/epcd/oei/view/sic-sht2.txt.

[16] There are 473 "official" six-digit NAICS manufacturing codes in the 2002 economic census. For a complete list of the six-digit codes, see http://www.census.gov/epcd/naics02/naico602.txt.

- Some of the NAICS basecodes to which HS codes are assigned are incomplete, while others are outside manufacturing. As noted in the third column of each panel of Table 7, the number of manufacturing NAICS industry codes to which these export (import) codes match ranges from 387 (387) in 2000 to 386 (388) in 2009. As with the SIC, these numbers of manufacturing codes are lower than the 473 official manufacturing industries in the NAICS.

Table 7: HS and Six-Digit NAICS Codes in the "Master List"

	Exports			Imports		
	HS10	NAICS6	Man. NAICS	HS10	NAICS6	Man. NAICS
2000	8,628	454	387	16,897	454	387
2001	8,622	453	386	16,910	453	386
2002	8,940	453	386	17,351	453	386
2003	8,930	454	386	17,390	454	386
2004	8,933	454	386	17,382	454	386
2005	8,971	453	386	17,717	453	386
2006	8,972	453	386	17,746	453	386
2007	8,878	453	385	17,665	455	387
2008	8,883	453	385	17,728	455	387
2009	8,882	454	386	17,745	456	388

Notes: Table displays number of ten-digit HS codes, six-digit NAICS codes and six-digit manufacturing NAICS codes appearing in the concordance files accompanying the U.S. monthly trade statistics sold by the U.S. Census Bureau.

E Appendix 5: Census's Principle Differences (Product Class-Basecode) Concordances

This section summarizes Census' PD files for 1992, 1997, 2002 and 2007.

E1. 1992 Economic Census

The 1992 PD file maps five-digit SIC product classes to eight-digit (SIC-based) basecodes and is available in the electronic appendix with filename pd92.csv. We note the following:

- 814 unique basecodes match to a product class (PC) in the 1992 PD file, 768 of which are in manufacturing. Table 8 summarizes the distribution of these basecodes according to the number of five-digit SIC product classes into which they map. As a group, the eight-digit basecodes contain 418 unique four-character basecode roots, 391 of which are in manufacturing. Note that there are 459 unique four-digit SIC manufacturing industries in 1992.[17]

[17] The set of four-digit SIC manufacturing industries in 1992 is identical to the set used in 1987. See www.census.gov/prod/2/manmin/mc92-r-1.pdf.

Table 8: Number of Product Classes per Basecode and Basecode Root (1992)

	Overall		Manufacturing	
Product Classes	Basecodes	Basecode Roots	Basecodes	Basecode Roots
1	549	117	520	111
2	109	60	103	54
3	59	76	53	69
4	36	50	32	46
5	20	39	19	37
6	17	25	17	24
7	9	13	9	13
8	2	7	2	7
9	2	9	2	8
10	2	5	2	5
11	1	2	1	2
12	2	4	2	4
14		3		3
15		1		1
16	1		1	
18	1	3	1	3
20	1		1	
21		1		1
22	1		1	
23	1	1	1	1
25	1	1	1	1
28		1		1
Total	814	418	768	391

Notes: Table displays distribution of basecodes and basecode roots according to the number of product classes into which they map, overall and for manufacturing.

- 1,566 unique five-digit SIC product classes are matched to an eight-digit basecode in the 1992 PD file. The official list of SIC categories for the 1992 CMF encompasses 1,462 five-digit product classes for manufacturing.[18]

  - A merge of the unique five-digit SIC codes from the PD concordance into the official list from Census (1992) reveals that 1400 codes match exactly and that they are all in manufacturing. The largest portion (24) of the 62 in the official list but not in the PD concordance end in "9", and their descriptions indicate they are generally receipts for contract work on the good categorized by the first four digits. Code 22579 in the PD file, for example, is "contract and commission receipts for knitting only or knitting and finishing weft (circular) knit fabrics". Code 22573, which appears in both the PD and the official list, by comparison, is "finished weft (circular) knit fabrics, excluding hosiery".

  - There are 166 five-digit SIC codes that are matched to HS codes in the PD concordance but do not appear in the official SIC list. Of the 166, 102 end in "0" and 95 are in manufacturing. We suspect that the 102 codes ending in "0" are used to facilitate the matching of SIC and basecodes by capturing a range of

[18] See Census (1992) at http://www.census.gov/prod/2/manmin/mc92-r-1.pdf.

goods spread across five-digit codes with the same four-digit root. For example, 20220 is in the PD file but not on the official list, and is described as "cheese, natural and processed, not specified as to kind", versus 20223 and 20224, both of which are in both the PD and the official list but which break cheese down into natural and processed cheese, respectively. All three of these codes map into the same basecode, 20223B00, which maps to HS codes beginning with 0406, i.e., "cheese and curd".

E2. 1997 Economic Census

The 1997 PD file maps seven-digit NAICS product classes to eight-digit (NAICS-based) basecodes and is available in the electronic appendix with filename pd97.csv.[19] We note the following:

- 841 unique basecodes are matched to a product class (PC) in the 1997 PD file, 763 of which are in manufacturing (i.e., begin with a "3"). Table 9 summarizes the distribution of these basecodes according to the number of seven-digit NAICS product classes into which they map. As a group, the eight-digit basecodes contain 451 unique six-character basecode roots, 388 of which are in manufacturing.

Table 9: Number of Product Classes per Basecode and Basecode Root (1997)

	Overall		Manufacturing	
Product Classes	Basecodes	Basecode Roots	Basecodes	Basecode Roots
1	576	143	518	105
2	130	91	120	80
3	55	71	50	65
4	24	44	21	40
5	25	38	23	35
6	7	16	7	15
7	6	12	6	12
8	5	7	5	7
9	4	6	4	6
10	3	4	3	4
11		4		4
12		4		4
13	1	1	1	1
14		2		2
16		1		1
17		1		1
18	2	2	2	2
19		1		1
24	1	1	1	1
30	1		1	
36		1		1
44	1	1	1	1
Total	841	451	763	388

Notes: Table displays distribution of basecodes and basecode roots according to the number of product classes into which they map, overall and for manufacturing.

[19] We thank Alvin Venning of the U.S. Census Bureau for providing us with a copy of the 1997, 2002 and 2007 PD files.

- 1559 unique seven-digit NAICS product classes are matched to an eight-digit basecode in the 1997 PD file, of which 1418 are in manufacturing. The official list of NAICS categories for the 1997 CMF encompasses 1469 seven-digit product classes in manufacturing.

E3. 2002 Economic Census

The 2002 PD file maps seven-digit NAICS product classes to eight-digit (NAICS-based) basecodes and is available in the electronic appendix with filename pd02.csv. We note the following:

- 832 unique basecodes are matched to a product class (PC) in the 2002 PD file, 754 of which are in manufacturing (i.e., begin with a "3"). Table 10 summarizes the distribution of these basecodes according to the number of seven-digit NAICS product classes into which they map. As a group, the eight-digit basecodes contain 450 unique six-character basecode roots, 388 of which are in manufacturing.

Table 10: Number of Product Classes per Basecode Root

	Overall		Manufacturing	
Product Classes	Basecodes	Basecode Roots	Basecodes	Basecode Roots
1	567	143	509	105
2	132	90	122	80
3	53	73	48	67
4	25	42	22	39
5	25	39	23	36
6	7	19	7	17
7	6	11	6	11
8	4	4	4	4
9	4	6	4	6
10	2	5	2	5
11		2		2
12		4		4
13	2	2	2	2
14		1		1
15		2		2
17		2		2
18	1		1	
19	1	2	1	2
23	1	1	1	1
31	1		1	
37		1		1
43	1	1	1	1
Total	832	450	754	388

Notes: Table displays distribution of basecodes and basecode roots according to the number of product classes into which they map, overall and for manufacturing.

- 1,547 unique seven-digit NAICS product classes are matched to an eight-digit basecode in the 1997 PD file, of which 1,406 are in manufacturing. The official list of NAICS categories for the 2002 CMF encompasses 1,450 seven-digit product classes in manufacturing.

E4. 2007 Economic Census

The 2007 PD file maps seven-digit NAICS product classes to eight-digit (NAICS-based) basecodes and is available in the electronic appendix with filename pd07.csv. We note the following:

- 799 unique basecodes are matched to a PC in the 2007 PD file, 724 of which are in manufacturing (i.e., begin with a "3"). Table 11 summarizes the distribution of these basecodes according to the number of seven-digit NAICS product classes into which they map. As a group, the eight-digit basecodes contain 454 unique six-character basecode roots, 390 of which are in manufacturing.

Table 11: Number of Product Classes per Basecode Root (2007)

	Overall		Manufacturing	
Product Classes	Basecodes	Basecode Roots	Basecodes	Basecode Roots
1	558	170	498	123
2	115	82	107	75
3	48	67	46	65
4	23	42	19	37
5	25	34	24	32
6	8	19	8	18
7	7	9	7	9
8	5	5	5	5
9	3	5	3	5
10	1	4	1	4
11		3		3
12	1	5	1	5
13	1	1	1	1
14		2		2
15		1		1
18		1		1
20	1	1	1	1
22	1	1	1	1
29	1	1	1	1
30	1		1	
36		1		1
Total	799	454	724	390

Notes: Table displays distribution of basecodes and basecode roots according to the number of product classes into which they map, overall and for manufacturing.

- 1,496 unique seven-digit NAICS product classes are matched to an eight-digit basecode in the 2007 PD file, of which 1,383 are in manufacturing. The official list of NAICS categories for the 2007 CMF encompasses 1,435 seven-digit product classes in manufacturing.