Online Data Appendix: New Measures of Imported Intermediate Inputs

Douglas L. Campbell
dcampbell@nes.ru
New Economic School

Lester Lusher
lrlusher@ucdavis.edu
UC Davis

May, 2016

Abstract

Given the rise of China, which has been linked to the decline in American Manufacturing, it is understandable that there is considerable academic and public interest in "offshoring". It is surprising, then, that there are no publicly-available, annual measures of offshoring at the detailed sector level (generally proxied by manufactured imported intermediate goods) of which we are aware. Thus, we have filled this gap by providing new estimates of imported intermediate goods at the 4-digit SIC level from 1972 to 2009, and for NAICS from 1992 to 2010. In this portion of the online Appendix, we detail the construction of our indices, and provide a detailed user-guide.

JEL Classification: F14, F16, L60
Keywords: Intermediate Inputs, Offshoring, Input-Output Tables

1 Creation of SIC Series

- In the first step, we downloaded the raw Input-Output Use tables for the benchmark years, 1972, 1977, 1982, 1987, and 1992 from the BEA. Then we created crosswalks between the IO SIC codes and the SIC codes used by the Annual Survey of Manufactures (ASM). We then combined the IO data with sectoral data from the ASM on materials inputs and import data by sector from WITS. Often we needed to apportion data for one IO SIC sector to several ASM SIC sectors. When we did this for the using sectors, we apportioned it based on the relative size of materials usage based on the ASM. For commodity (providing) sectors, we apportioned intermediate imports based on the ratios of imports in each ASM SIC sector.

- We then make use of the "proportionality" assumption which is used by the BEA in their later estimates of imported intermediate inputs, and also by Feenstra and Hanson (1999) – hereafter FH. While this assumption is not perfect, Feenstra and Jensen (2012) showed that this assumption is mostly correct, as they find that at the 3-digit level, direct estimates of imported intermediate inputs at the three digit level from the Linked/Longitudinal Firm Trade Transaction Database (LFTTD), which do not rely on the proportionality assumption, have a correlation of .68 with data that does rely on this assumption, or a correlation of .87 when the shares are value-weighted. Thus we assume that a sector consumes the same fraction imports of a commodity as it does of domestic consumption. From FH, for each industry i, its sum of intermediate inputs from sector j is computed as:

$$\sum_j (input\,purchases\,of\,j\,by\,industry\,i) * \frac{imports\,of\,j}{consumption\,of\,j} \tag{1.1}$$

where consumption of good j is measured by: $shipments + imports - exports$.

- There were also a small number of cases where the right-hand term, also known as import penetration, is either less than zero or greater than one. This could happen if, for example, imports and exports were equal and greater than shipments, or if exports were greater than shipments plus imports (perhaps indicating a an inconsistency with the data). One option would be to relax the implicit assumption that imports are not re-exported. While our import data is ostensibly imports for domestic production, and not for re-export, for some industries, this assumption is clearly not met. At the same time, for 1992, only 5 out of 462 sectors yield problematic figures for import penetration, and so for these sectors, we made a second "proportionality" assumption, assuming that the share of imports that are re-exported are equal to $\frac{imports}{imports+shipments}$. Thus, the imports for domestic use will be given by the following formulation:

$$M_{Domestic} = M\left(1 - \left(\frac{X}{M + Shipments}\right)\right) \tag{1.2}$$

Where M = Imports, X = Exports, and $M_{Domestic}$ are imports for domestic consumption. The intuition for this formula is that total imports are multiplied by the share of imports plus domestic shipments which are consumed at home, equal to one minus the share of imports plus shipments which are exported. Thus, in these cases, we recalculated equation (3) replacing "imports of j" with $M_{Domestic}$

from equation 4, and consumption using the formula:

$$shipments + M_{Domestic} - X_{Domestic}, \tag{1.3}$$

where $X_{Domestic}$ are exports produced domestically, equal to exports times shipments plus imports. This construction of import penetration has the benefit that it always varies between 0 and 1. However, it also has the downside that if, in reality, a larger share of exports come from domestic production than from imports, then it will underestimate intermediate imports. If we use this formulation for only the 5 "problem cases" then we will be changing the rank order of import penetration across sectors. (One solution might be to just assume an import penetration ratio of 1 in these cases, but given that domestic production was substantial in each of these 5 cases, this would appear to be counterfactual). Thus, the last step is to make a rank-order adjustment, assuming that, for each of these 5 sectors, their rank order in this alternative calculation, which yields generally lower estimates for import penetration, is their true rank, and then adjusting upwards by the ratio of the average import penetration using equation 1.1 with that based on equation 1.3, which in practice is a 20% upwards adjustment for these observations. In this way, their rank in terms of import penetration based on equation 1.3 is roughly preserved. The original and reformed series are compared below in Figure 1.

- We did find that, for a small handful of sectors, the imported intermediate inputs calculated this way are too large. For some commodity sector-using sector combinations, the amount of imported intermediate imports we record are larger than the total imports recorded in that particular commodity sector (and this does not seem to depend on the data source – as we tried USITC data, and also WITS data). In addition, when we sum intermediate import uses by commodity sectors (thus, for each commodity sector we add up all of the uses across using sectors), we find that for roughly 12% of sectors, the intermediate imports are again "too large", in that their value is greater than total recorded imports. This is true both for the imported intermediate inputs provided at the NAICS level by the BEA itself, and for our our projections for imported intermediate inputs which were taken from the raw data. While the IO tables are constructed using data from the ASM, there are a variety of reasons why this might be the case. The ASM data are meant to be annual, as are the IO data, but one can imagine that if a good is produced in January, then the intermediates purchased likely came in the previous year, while intermediates purchased at the end of the year will likely go towards

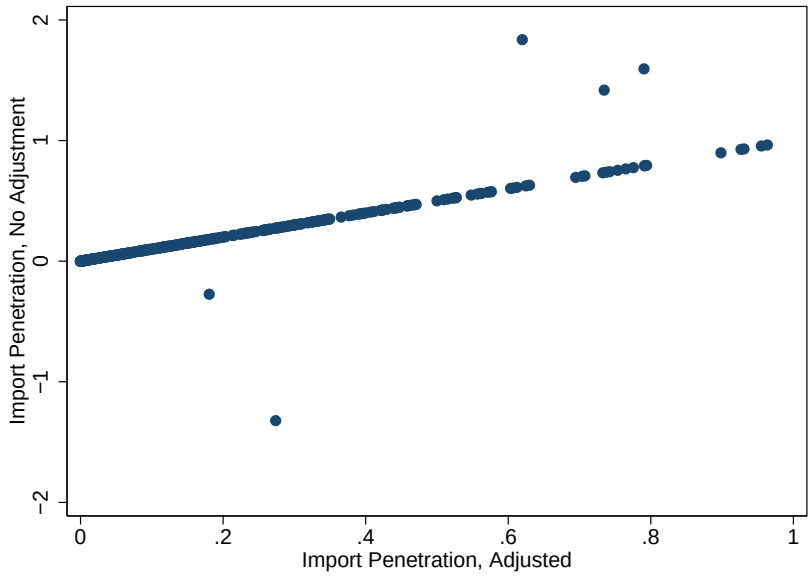


Figure 1: Import Penetration vs. Adjusted Import Penetration, 1992

Note: "Adjusted Import Penetration" adjusts sectors with implausible values for re-exports. With this adjustment, import penetration is forced to vary between zero and one.

production in subsequent years. Given that some volatility in manufacturing is to be expected, this could be one factor which generates implausibly large (or small) estimates for some years. It is also the case that the "Make" table is not entirely consistent with the variable "shipments" from the ASM, although the two variables are very highly correlated. We tried using materials shipments from the ASM, and then multiplying this amount by the ratio of using sector i's make to each use of commodity j, but this led to even more inconsistent results. We also considered capping intermediate imports for any commodity sector at the total recorded imports, but eventually, we decided against this on the grounds that it is arbitrary, and for our purposes we are mostly interested in the degree of reliance on intermediate imports, and so we decided to preserve this feature of the data.

- Equation 1.1 can be used to derive the intermediate import matrix for the benchmark years. To extrapolate for the other years, we first modeled the evolution of intermediate inputs based on changes in materials usage of the using sectors and changes in import penetration of commodity sectors. Thus, we estimate:

$$ln(MPI)_{ijt} = \rho L5.ln(MPI)_{ij,t-1} + \beta_1 ln\Delta ImportPen._{jt} + \beta_2 ln\Delta Materials_{it} + \epsilon_{ijt} \tag{1.4}$$

where MPI_{ijt} = imported intermediate imports of commodity j used by sector i at time t, $ImportPen.$ is import penetration, "Materials" is materials used by sector i, and we have suppressed the constant. The results are displayed in Column 1 of Table 1, which show that lagged intermediate imports enter with a coefficient equal to one, and that changes in import penetration and materials inputs are highly predictive of change in imported intermediate inputs between the benchmark years. Given the lagged coefficient of 1, we now estimate the model in log changes:

$$ln\Delta MPI_{ijt} = \beta_1 ln\Delta ImportPen._{jt} + \beta_2 ln\Delta Materials_{it} + \epsilon_{ijt} \tag{1.5}$$

We run this model for the full period (column 2), and for each of the years individually in Table 1, and find that the coefficients are roughly the same, and have considerable predictive power in terms of r-squared, of .44 on the full sample.

Table 1: Modeling the Evolution of Intermediate Imports: SIC

	(1) ln(MP Inputs)	(2) ln5yr. Δ MPI	(3) 1992	(4) 1987	(5) 1982	(6) 1977
L5.ln(MP Inputs)	1.00***					
	(0.0049)					
ln 5yr. Δ Import Pen.	1.13***	1.14***	1.15***	1.16***	1.08***	1.05***
	(0.060)	(0.053)	(0.084)	(0.076)	(0.100)	(0.12)
ln 5yr. Δ Matcost	0.44***	0.46***	0.57***	0.72***	0.34***	0.45***
	(0.037)	(0.053)	(0.14)	(0.045)	(0.11)	(0.091)
Observations	110148	110148	19964	32965	30570	26649
r2	0.96	0.44	0.54	0.52	0.34	0.27

Notes: Standard errors clustered by commodity-using sector pair in parenthesis. $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the log of imported intermediate inputs in column (1), and the log 5 year change of imported intermediates in columns (2)-(6). Column (2) is run on the full sample, while columns (3)-(6) are run on individual years. These regressions were run without a constant.

To test this model out of sample, we ran the panel while excluding the year 1992, and then we generated out-of-sample predictions using realized changes in imports, shipments, and materials costs of the using sector. Below in Figure 2 we show the out-of-sample results. On the whole, it looks like our model validates fairly well. The mean absolute error is .73, while regressing our predictions on the actual data of log changes in imported intermediate inputs yield a coefficient of 1.02 (with an error of .07), and an R-squared of .54. Thus, these results appear to be a method we can use to extrapolate to years in which there is no benchmark

data. While there is significant error in this method, note that when we do our actual extrapolation, since there are 5 years in-between benchmarks, we'll generally never have to extrapolate more than 3 years, and even then, we can form multiple estimates derived from "backward" and "forward" estimation.

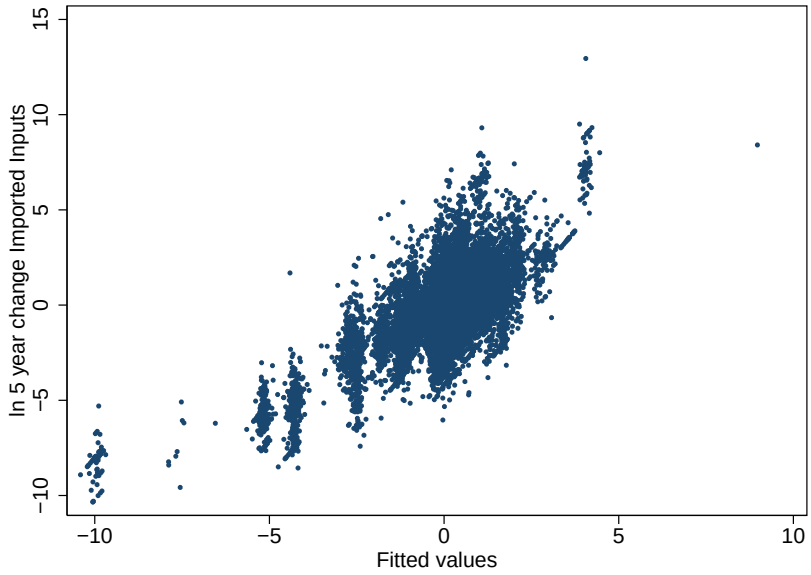


Figure 2: Out-of-Sample Test of Intermediate Imports, 1992

Notes: On the x-axis we have plotted the model predictions for intermediate imports (used by sector i of commodity j), which is based on changes in import penetration of the commodity and in total materials usage of the using sector.

- Using these regression coefficients, we first fill in missing data in the benchmark years simply by using the regression predictions from column (2). (Except for 1972, there we run the same regression backwards, where the dependent variable is now the change in imported inputs from the subsequent benchmark, and predict the missing values for the initial benchmark that way.)

- Then, we extrapolate forward and backward from the base years. For the years after 1992, we use the formula:

$$MPI_{ijt} = MPI_{ij,t-1} * exp(1.14 * ln\Delta ImportPen. + .46 * ln\Delta MaterialsCost) \quad (1.6)$$

  For years in between benchmark years, we use a weighted average of the forward extrapolation from the previous benchmark year (using formula 1.6, and the backwards extrapolation from the subsequent benchmark year (which also uses a

formula similar to 1.6 to extrapolate backwards using the regression coefficients).

$$MPI_{ij,t+s} = \frac{k-s}{k} MPI^{F}_{ij,t+s} + \frac{s}{k} MPI^{B}_{ij,t+s} \tag{1.7}$$

where t is a benchmark year, k is the number of years between benchmarks, and s is the number of years after the last benchmark. Thus, for 1988, which is one year after the 1987 benchmark, this estimate gives the forward estimate a weight of .8, and the backwards estimate from the 1992 benchmark a weight of .2. However, in some cases, there may be 10 years between any two benchmarks, such as between the 1972 and the 1982 benchmark. In this case, in 1973, the forward estimate from 1972 will be given a weight of .9 and the backward estimate extrapolating from 1982 will be given a weight of .1.

- Comparing our results to the Feenstra and Hanson (1999) estimates for 1990 in Figure 3, the correlation between the two is pretty good, albeit not perfect. Regressing the log of the Feenstra estimates on our own, we get an R-squared of .82, while our estimates are bit smaller on average.

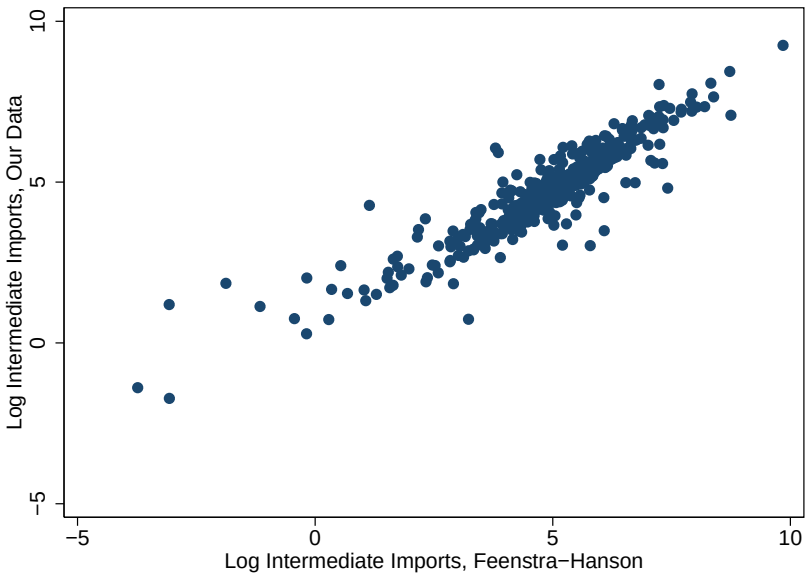


Figure 3: Feenstra-Hanson Comparison, 1990

- We then summed across commodities to arrive at the total amount of imported intermediate inputs for each SIC using sectory by year. Following 1999, we also provide estimates of "narrow" offshoring, defined as total intermediate inputs within

the same 2-digit SIC sector. Lastly, we converted these estimates to the MORG version of SIC for the 1979 to 2002 period.

2 Creation of NAICS Series

- Intermediate import data were downloaded from the BEA: (including from here: http://www.bea.gov/industry/iedguide.htm) for the years 1997, 2002, and 2007. These were the only years in which intermediate imports were computed directly by the BEA. The 2012 data will become available in 2017.[1]

- These data used NAICS codes which are specific to the IO database (we call it IONAICS), which differ slightly each year. Thus, we created a crosswalk between the IONAICS codes in each year and NAICS codes from the Annual Survey of Manufactures (in order to match this data to ASM data and use it as a panel). As, in many cases there is one IONAICS sector which matches to several ASM NAICs sectors. Thus, it is necessary to partition the intermediate inputs data to the multiple NAICS sectors based on (1) intermediate materials consumption for the using sectors, and (2) imports for the commodity sectors. To make a concrete example, in 2002, the IONAICS sector 315100 is matched to two ASM NAICs sectors: 314991 and 314999. Thus, if a given using sector used 10 million worth of inputs from this sector, then we divided those imports based on the relative share of imports of each of those sectors. Thus, if 314991 had total imports of 400 and sector 314999 had imports of 600, then we would do a 40/60 split. For the using sectors, we would do the same thing, only using materials input usage.

- Next, as we did with SIC, we tested to see if we could predict changes in intermediate inputs in the data provided by the BEA in order to extrapolate out of sample. Column (1) of Table 3 shows the simple model where we regress log imported inputs (commodity j used by sector i) on its 5 year lag and changes in import penetration measured by commodity and changes in materials cost measured by the using sector. Although the point estimate of the lag is not so close to one, the point of this exercise is to get coefficients which can help us predict the evolution of intermediate inputs solely based on changes in other variables so that we can extend the series to additional years. Thus, in column (2), we replace the left-hand-side variable with the log change in imported intermediate imports (MPI).

1. We thank Robert Correa, and Economist at the BEA, for providing this information.

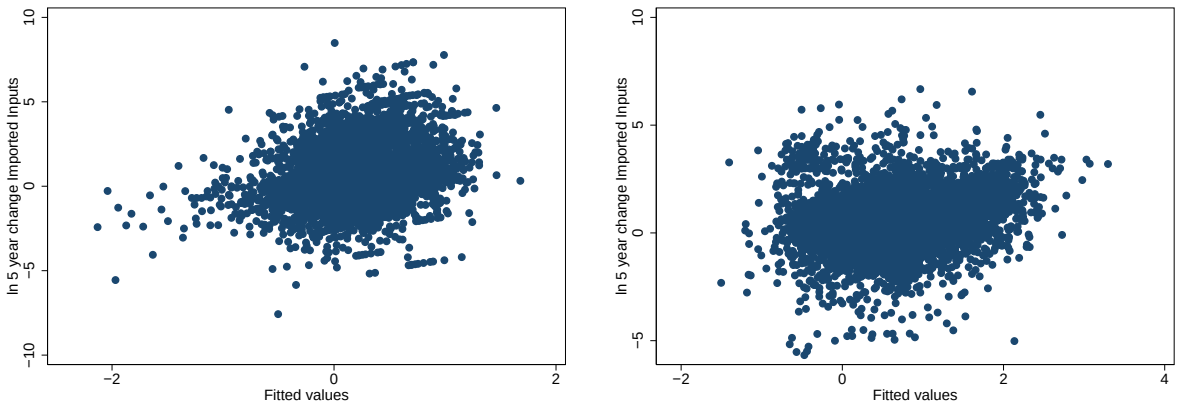
Again, the coefficients look broadly similar to what we had previously in Table 1, although the coefficient on materials usage is a bit higher and the coefficient on import penetration is a bit lower. Unfortunately, when we run this regression on invdividual years, it falls apart. Perhaps this is due to the massive volatility experienced by the manufacturing sector in this period, which experienced a collapse, or due to the fact that, since we are now using the BEA's estimates for imported intermediates, we now have less control over the exact data generating process. In column (4), we try instead an alternate model, in which which swap out import penetration for simply the log change in imports. This tends to do better, as in this case, there is at least a small amount of out-of-sample predictive ability (see 4 below).

Table 2: Modeling the Evolution of Intermediate Imports: NAICS

	(1)	(2)	(3)	(4)	(5)	(6)
	ln(MP Inputs)	ln5yr. Δ MPI	2007	Full	2007	2002
L5.ln(MP Inputs)	0.75***					
	(0.0047)					
ln 5yr. Δ Import Penetration	0.84***	0.83***	-0.035			
	(0.032)	(0.035)	(0.050)			
ln 5yr. Δ Matcost	1.40***	1.17***	1.39***	0.76***	0.99***	0.71***
	(0.034)	(0.037)	(0.038)	(0.037)	(0.040)	(0.077)
ln 5yr. Δ Imports				0.99***	0.64***	1.35***
				(0.024)	(0.029)	(0.038)
Observations	12452	12452	6419	12452	6419	6033
r2	0.71	0.11	0.18	0.19	0.23	0.18

Notes: Standard errors in parenthesis. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable in column is the log of imported intermediate inputs in column (1), and the log 5 year change of imported intermediates in columns (2)-(6). Columns (2) and (4) were run on the full sample, while columns (3) and (5) were run just on 2007 and column (6) was run only using data from 2002. The constant was suppressed in each of these regressions.

- Given the somewhat rough out-of-sample test results in Figure 4, one option would certainly just to do a linear extrapolation. If the period from 1997 to 2010 did not include any big events, this might be advisable. However, this period also includes a major financial crisis and recession, and we suspect that, in this case, the collapse in both materials used and in imports in the 2009-2010 period would have had to have been reflected in fewer intermediate inputs. In addition, our own out-of-sample tests do both have a lower mean absolute error as compared to either a random walk or an extrapolation of the previous trend, both for 2002 and for 2007. In addition, part of the reason the out-of-sample tests here may be worse



(a) 2002 (b) 2007

Figure 4: Out-of-Sample Tests of Intermediate Input Growth

Notes: The fitted values are derived from a regression of log changes in imported inputs on log changes in imports and materials costs.

is that the in-sample portion of the model is only one year in each case. When we actually do the extrapolation, we'll be using twice as much data, which means that the models performance should be improved.

Table 3: Out-of-Sample Tests

	Mean Absolute Error	
	2002	2007
Random Walk	1.27	1.11
Random Walk with Drift		1.66
Our Model	1.17	.95
N	6419	6033

Notes: The Mean Absolute Error is compared for each model for 2002 and for 2007. The "Random Walk" simply uses the estimate for imported intermediate inputs from the previous benchmark. The "Random Walk with Drift" uses the time trend from 1997 to 2002 to predict imported intermediates in 2007. "Our model" uses our regression results to test out-of-sample.

- In the first part of the extrapolation, we once again fill in the missing observations for the benchmark years using the regression in column (4) of Table 3.

- Then, we fill in the remainder of the years in the exact same way as with the SIC indices. For the years after the 2007 benchmark, we use the formula:

$$MPI_{ijt} = MPI_{ij,t-1} * exp(.99 * ln\Delta Imports + .76 * ln\Delta MaterialsCost) \tag{2.1}$$

For years in between benchmark years, we use a weighted average of the forward extrapolation from the previous benchmark year (using formula 2.1, and the backwards extrapolation from the subsequent benchmark year (which also uses a formula similar to 2.1 to extrapolate backwards using the regression coefficients).

$$MPI_{ij,t+s} = \frac{k-s}{k} MPI^{F}_{ij,t+s} + \frac{s}{k} MPI^{B}_{ij,t+s} \tag{2.2}$$

where t is a benchmark year, k is the number of years between benchmarks, and s is the number of years after the last benchmark. Thus, for 1998, which is one year after the 1997 benchmark, this estimate gives the forward estimate a weight of .8, and the backwards estimate from the 1992 benchmark a weight of .2. However, in some cases the 2002 benchmark data is simply missing, in which case there are 10 years between benchmark observations. In this case, in 1998, the forward estimate from 1997 will be given a weight of .9 and the backward estimate extrapolating from 2007 will be given a weight of .1.

- We then summed across commodities to arrive at the total amount of imported intermediate inputs for each NAICS using sectory by year. Once again, we also included "narrow" estimates of offshoring by summing up imported inputs for each using sector for all the commodities within the same 3-digit NAICS classification. Lastly, we converted these estimates to the MORG version of NAICs for the 2003 to 2010 period.