**“Import Competition and the Great U.S. Employment Sag of the 2000s”**

**by Daron Acemoglu, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price**

**Data and Programs**

May 13, 2015

*These data and programs are provided for use in replicating the particular analyses reported in the paper. We make no warrant as to their suitability for other purposes.*

**NOTES ON THE REPLICATION PACKAGE**

**Basic overview:**

* The folder “do/build/national” contains the programs used to construct the industry-level sample used in our national analysis.
* The folder “do/build/regional” contains the programs used to construct the commuting zone-level sample used in our analysis of local labor markets.
* The folder “do/estimate” contains scripts that perform all of the analyses reported in the paper (tables, figures, and appendix).
* The folder “dta” contains unedited source data as well as cleaned data files prepared by the scripts in “do/build’.
* The folder “out” contains all analytical output.
* The folder “xwalks” contains the geographic and industry crosswalks used in the paper.

**A note on Stata:** All .dta files are saved in Stata 13 format. If you are using an older version of Stata, you may load these files using the user-written command “use13” (available on SSC). The .do files were prepared using Stata version 12.1, and this version is thus invoked at the beginning of each script.

**A note on source data:** To conserve space, for most datasets we provide only cleaning scripts and the cleaned data: we omit the (typically larger) source files. In these cases, we indicate where users may download the source data. The cleaning scripts refer to source files by their original names when downloaded.

For information on the provenance of the datasets and crosswalks used in this paper, please refer to the numerous readme files contained within the “dta” and “xwalks” folders.

**A note on terminology:** The version of this paper published in the *Journal of Labor Economics* reverses the input-output terminology used in earlier versions (notably NBER working paper no. 20395). The new terminology has been implemented both in the manuscript’s text and tables and in the underlying Stata code. We urge all researchers employing our code to abide by the updated terminology.

Previously, we used the term “downstream (upstream) import exposure” to denote the impact on a given industry stemming from import shocks experienced by its customers (suppliers). We now use the term “upstream import exposure” to denote the impact on a given industry that propagates upstream from its import-exposed customers, and “downstream import exposure” to denote the impact that propagates downstream from its import-exposed suppliers.

We believe this new terminology better accords with conventions used in the existing literature. We apologize for any confusion.

**PREPARATION OF THE MAIN DATASETS**

In this section, we briefly describe the preparation of the samples used in our analysis. For further details, please consult the Stata .do files provided in this package.

**Preparation of the industry-level sample:** Our industry-level analyses are conducted at the level of 392 manufacturing and 87 non-manufacturing industries, identified using 4-digit “SIC87DD” codes.

For most industry-level specifications, employment data are derived from the national County Business Patterns files for the years 1991, 1999, 2007, and 2011.[[1]](#footnote-1) We begin by calculating employment within 1987 SIC codes. Broadly, this process involves five steps (additional, non-employment outcomes derived from the CBP are computed similarly):

1. *Imputation of missing cells*. Certain industry-level employment counts are suppressed in the CBP for confidentiality reasons. In these cases, we nonetheless observe the number of establishments disaggregated by employment categories. Because only a small number of cells are suppressed in the national-level file, we impute missing cells on the basis of these establishment counts, using information from non-suppressed cells to determine the average number of employees per establishment within each establishment size class.
2. *Exclusion of employment in auxiliary establishments*. Treatment of auxiliary establishments changes greatly under the NAICS regime, and we know of no reliable way to attribute auxiliary employment to detailed industries. We therefore exclude auxiliary establishments from the analysis.
3. *Allocation of higher-level totals to detailed industries*. The CBP reports employment counts at multiple levels of industry aggregation. In particular, employment within a higher-level industry (such as a 3-digit industry) often exceeds the sum of employment within the corresponding lower-level industries (the underlying 4-digit industries). We allocate this excess (or “residual”) employment across lower-level industries proportionately on the basis of these industries’ employment shares within the higher-level aggregate. By applying this procedure recursively, we obtain employment counts entirely at the 4-digit industry.
4. *Crosswalking into 1987 SIC codes*. Over time, the CBP switches from 1987 SIC codes (used through 1997) to 1997 NAICS codes (used over 1998-2002) to 2002 NAICS codes (used 2003-2007) to 2007 NAICS codes (used from 2008 onwards). We crosswalk all employment into 1987 SIC codes (crosswalks are provided in this package).
5. *Aggregation to SIC87DD codes*. Following Autor, Dorn, and Hanson (2013) and Acemoglu et al. (2014), we aggregate manufacturing codes modestly into “SIC87DD” codes. We perform this aggregation both to facilitate matching with the available trade data and to minimize the risk of classification error.[[2]](#footnote-2) When analyzing downstream and upstream linkages, we aggregate non-manufacturing industries to the level of 87 industries that can be consistently identified with the commodity flows reported in the Bureau of Economic Analysis’s 1992 Benchmark Input-Output Table.

Additional specifications use industry-level employment, payroll, output, and price deflators derived from the NBER-CES Manufacturing Industry Database, which is based largely on the Annual Surveys of Manufacturing (Becker and Gray, 2009).[[3]](#footnote-3) Because this database straddles the 1997 conversion from SIC industry codes to NAICS codes, the providers of the NBER-CES publish both a version expressed in 1987 SIC codes and a version expressed in 1997 NAICS codes. Following Acemoglu et al. (2014), we combine information from both of these files. First, we use a weighted crosswalk to convert the NAICS-based data from the years 1997-2009 into 1987 SIC codes, and append the SIC-based data from the years 1977-1996. Second, we map SIC codes into the same SIC87DD codes used in cleaning the CBP data.

Industry-level trade data are adapted from Autor, Dorn, and Hanson (2013) and are ultimately derived from the UN Comtrade Database. See “dta/trade/trade\_readme.txt” for further details.

Data on upstream and downstream linkages derive from the Bureau of Economic Analysis’s 1992 Benchmark Input-Output Tables. First, we combine the “make” and “use” tables to determine each industry’s usage of commodities produced by each other industry. Second, we crosswalk from BEA industries into SIC87DD codes. For manufacturing industries that do not match one-to-one, we allocate input and output flows proportionately on the basis of shipments and materials usage measured in the NBER-CES data. For non-manufacturing industries, such an allocation procedure is not possible, so we aggregate non-manufacturing industries as needed to ensure a one-to-one match.

**Preparation of the commuting zone-level sample:** Our regional analysis encompasses 722 commuting zones comprising the whole of the mainland United States (excluding Alaska and Hawaii). Commuting zones are aggregations of counties (Tolbert and Sizer, 1996; Autor, Dorn, and Hanson 2013).

Employment within each commuting zone x industry cell is computed using the county-level County Business Patterns files for the years 1991, 1999, 2007, and 2011. In most respects, our preparation of the county-level CBP files parallels that of the national-level CBP files described above. The main difference is that, because the county-level file has considerably more suppressed cells (reflecting the greater risk of deanonymization in regional data), we employ a fixed-point algorithm to estimate employment within suppressed cells. The algorithm is similar to that used in Autor, Dorn, and Hanson (2013).

Commuting zone population is calculated by aggregating county-level population counts in the Census Population Estimates. Following standard practice, we use so-called “intercensal” estimates whenever these are available (in particular, for 1991-2010) and “postcensal” estimates otherwise (for 2010-2011).

**REFERENCES:**

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. 2014a. “Import Competition and the Great U.S. Employment Sag of the 2000s.” NBER Working Paper no. 20395.

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. 2014b. “Return of the Solow Paradox? IT, Productivity, and Employment in US Manufacturing.” *American Economic Review Papers & Proceedings*, 104(5), 394-399.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. “The China Syndrome: Local Labor Market Effects of Import Competition in the United States.” *American Economic Review*, 103(6), 2121-2168.

Tolbert, Charles M., and Molly Sizer. 1996. “US Commuting Zones and Labor Market Areas: A 1990 Update.” Economic Research Service Staff Paper 9614.

**For more information or to report errors:** Please contact Brendan Price at bmprice@mit.edu.

1. The CBP data are available at http://www.census.gov/econ/cbp/download. [↑](#footnote-ref-1)
2. The SIC87DD manufacturing codes used in this paper differ slightly from those used in Autor, Dorn, and Hanson (2013) and Acemoglu et al. (2014). In particular, Acemoglu et al. (2014) exclude a small number of SIC manufacturing industries that were reclassified into non-manufacturing under NAICS, since the core outcomes studied in that paper cannot be computed for these industries after the NAICS conversion. Because we can calculate employment in these industries over the full sample period using the CBP, we retain these problematic industries except for specifications whose outcome variables are derived from the NBER-CES Manufacturing Industry Database. [↑](#footnote-ref-2)
3. While most of the variables in the NBER-CES are taken from the Annual Surveys of Manufacturing, price deflators and depreciation rates are derived from other data published by the Census Bureau, the Bureau of Economic Analysis, the Bureau of Labor Statistics, and the Federal Reserve Board. NBER-CES data and documentation are available at http://www.nber.org/nberces. [↑](#footnote-ref-3)