THE EFFECTS OF INDUSTRY CLASSIFICATION CHANGES ON US EMPLOYMENT COMPOSITION*

Teresa C. Fort[†]
Tuck School of Business, Dartmouth College

Shawn D. Klimek[‡]
Center for Economic Studies, US Census Bureau

Preliminary Draft–Comments Welcome
This Version: March 2016

Abstract

This paper documents the extent to which compositional changes in US employment from 1976 to 2009 are due to changes in the industry classification scheme used to categorize economic activity. In 1997, US statistical agencies began implementation of a change from the Standard Industrial Classification System (SIC) to the North American Industrial Classification System (NAICS). NAICS was designed to provide a consistent classification scheme that consolidated declining or obsolete industries and added categories for new industries. Under NAICS, many activities previously classified as Manufacturing, Wholesale Trade, or Retail Trade were re-classified into the Services sector. This re-classification resulted in a significant shift of measured activities across sectors without any change in underlying economic activity. Using a newly developed establishment-level database of employment activity that is consistently classified on a NAICS basis, this paper shows that the change from SIC to NAICS increased the share of Services employment by approximately 36 percent. 7.6 percent of US manufacturing employment, equal to approximately 1.4 million jobs, was reclassified to services. Retail trade and wholesale trade also experienced a significant reclassification of activities in the transition.

JEL codes: E24

[*] Comments are welcome. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the US Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. We thank Linh Vu for excellent research assistance.

[†] 100 Tuck Hall, Hanover, NH 03755, USA, *tel:* +1 603 646 8963, *email:* teresa.fort@tuck.dartmouth.edu

[‡] *email:* shawn.d.klimek@census.gov

1 Introduction

Industrial classification systems are critical for obtaining accurate measures of economic activity. They provide information about the types of goods and services produced in an economy, the kinds of jobs available, and countries' comparative advantage. Industry classification systems also play an important role in economic research. They are used across all fields to restrict a study to a specific type of activity (e.g., manufacturing), to analyze the industrial composition of a sample, to provide appropriate controls, and even to construct sources of exogenous variation.[1]

Industrial classification systems are also essential for analyzing how economic activity has changed over time. In order to produce reliable time series information, however, the classification system itself must be consistent across time. This need for consistency poses a significant hurdle to analyzing economic change, since economies naturally evolve, with some industries becoming obsolete and the need for new categories arising. Statistical agencies face a tradeoff between maintaining the same system to facilitate time series analysis and updating the existing system to ensure the most accurate depiction of activity at a particular point in time. In 1997, US statistical agencies transitioned from using the Standard Industrial Classification System (SIC) to using the North American Classification System (NAICS). NAICS was implemented to provide a consistent classification methodology across different industries, to facilitate comparisons across the US, Canada, and Mexico, and to introduce new flexibility so that revisions could be made to the system as the economy grows and evolves. While the transition from SIC to NAICS helps address the challenges of clearly and correctly classifying economic activity in an evolving economy, it also poses new challenges. In particular, it makes it difficult to conduct any time series analysis that relies on industry classifications and that spans years prior and post 1997.[2]

This paper has two main goals. First, it provides an accurate and consistent measure of the composition of US economic activity from 1976 to 2009. We find that the transition from SIC to NAICS led to significant reclassification of activity from manufacturing and wholesale into other sectors, particularly services. A naive analysis of the changes in US employment composition over this period would overstate the rise in services employment by approximately 36 percentage points and overstate the fall in manufacturing employment by almost 11 percentage points. A second goal of the paper is to illustrate potential sources of bias that may arise in economic research that spans this time period and relies on industry classifications. We therefore identify key distinctions between SIC and NAICS and discuss how ignoring them may bias existing research.

[1] For example, Rajan and Zingales (1998); Autor et al. (2013); Pierce and Schott (2015)all use industry-level information as a source of exogenous variation to identify a causal effect.

[2] It is challenging to conduct a times series analyses with any industry component whenever the classification system changes. SIC underwent revisions in 1972, 1977, and 1987. The transition from SIC to NAICS involved reclassifying a significantly larger share of activity, and the affected activity was more likely to be classified not just in different industries, but also in different sectors.

The Standard Industrial Classification (SIC) system was first implemented in the US in the 1930s. Although it was updated numerous times, by the 1990s the SIC system faced several important limitations. First, it did not provide adequate categories new types of activities, especially those in services. Second, it classified activity based on a number of different concepts, including both production and demand-based definitions. According to Triplett et al. (1995), by the 1990s the SIC system was sufficiently heterogeneous in its classification of economic activity that it was necessary to qualify interpretations of results from research based on the SIC system, especially cross-sectional and time-series studies making inter-industry comparisons. Third, it was not easily comparable to international classification systems. In 1997, after years of discussion and analysis, US statistical agencies addressed these issues by replacing SIC with the new NAICS.

Unlike SIC, all industry categories in NAICS are based on production concepts. In other words, establishments are classified into industries based on the activities performed at the establishment, rather than as a function of the products they sell or the customers they serve. The NAICS production-process methodology was chosen to provide consistency across different industries, as well as to facilitate comparisons across the US, Canada, and Mexico. In addition, NAICS was designed to be flexible so that revisions could be made to the system as the economy grows and evolves.

We analyze the transition from SIC to NAICS using data from the US Census Bureau's Longitudinal Business Database (LBD) from 1977 to 2009. The LBD has information on every private, non-farm establishment in the US. We address the transition from SIC to NAICS by developing a new methodology that exploits the time series and establishment nature of the LBD to assign a consistent NAICS 2002 industry code to every establishment in the LBD. Using this consistent classification across the entire time period, we can calculate more accurate measures of the changes in US employment composition over time. Our methodology allows for differences in the evolution of economic activity at the establishment level.

The main contribution of the paper is to develop and implement a new methodology that assigns a consistent industry code to all establishments in the LBD. These codes allow us to calculate the extent to which the transition from SIC to NAICS changed the sectoral composition of US employment, both in the aggregate and at finer levels of aggregation. Although statistical agencies and researchers have been careful to address the potential biases that might arise from the transition, their efforts have been limited to using aggregate concordances. Our approach exploits the time series and establishment-level features of the LBD to reduce the need for random assignments and to allow for industry mappings to vary by establishment size. Because every establishment is assigned a NAICS code, it is possible to exploit industry variation across geography and across the firm size distribution. This flexibility is important since researchers' identification strategies often rely on geographic variation, or since specific mechanisms may have very heterogeneous effects across different firms.

This paper also contributes to two important literatures. First, it provides a new perspective on a recent literature that examines the decline of US manufacturing employment. A number of papers have studied this decline and attributed a significant portion of it to increased competition from Chinese imports (e.g., Acemoglu et al. (forthcoming); Pierce and Schott (2015); Autor et al. (2013)). We show that the change from SIC to NAICS in the LBD, which is concurrent with China's accession to the WTO, is also associated with a large decrease in manufacturing employment. The paper also adds to an existing literature analyzing the methodological impact of various classification systems (Fertuck (1975), Clarke (1989), and Guenther and Rosman (1994)). Those papers focus on analyzing SIC codes at different levels of aggregation and determining how informative they are in describing a firm's activities. In contrast, we show how the change from SIC to NAICS affects measures of US employment composition. These changes may bias research on time series analyses that span the transition if the transition is not explicitly addressed.

The rest of the paper proceeds as follows. In the next section, we describe the various classification schemes and the public and confidential data used in this project to map between them. Section 3 provides an overview of the change from SIC to NAICS and summarizes the main differences between the two schemes. In section 4 we describe the methodology used to assign a NAICS 2002 code to all establishments from 1976 to 2009. Section 5 shows how the change from SIC to NAICS affects measured employment differences across sectors. The last section provides concluding remarks and suggestions for future research.

2 Data and Industry Classification Systems

In this section, we describe the various datasets used in our analysis. We also describe the industry classification systems used in these datasets and how they change over time.

2.1 Data description

The main analysis in the paper is based on establishment-level data from the US Census Bureau. An establishment denotes a single physical location where business transactions take place and for which payroll and employment records are kept. The primary dataset we use is the Longitudinal Business Database (LBD). The LBD is an establishment-level dataset of every non-farm, private, employer establishment in the US from 1976 to the present. The LBD data are based on administrative tax data and are designed to provide longitudinal links of the same establishment over time. See Jarmin and Miranda (2002) for a description of these data. The LBD is constructed from the Business Register (BR). The BR is a comprehensive dataset compiled from administrative tax records and augmented with information from various additional sources. The widely used Country Business

Patterns (CBP) data are constructed from the BR. In instances in which the BR contains more industry information than the LBD, we incorporate the additional information into our analysis.[3]

We also supplement the LBD with information from the Economic Censuses (ECs). The EC data are establishment-level data collected by comprehensive surveys that are conducted in years that end in 2 and 7. There are nine Economic Censuses (Census of Auxiliaries; Census of Construction; Census of Finance, Insurance and Real Estate; Census of Manufacturing; Census of Mining; Census of Retail Trade; Census of Services; Census of Transportation, Warehousing and Utilities; Census of Wholesale Trade). The EC data also provide industry information at the establishment level. Because the EC data provide information collected directly from establishments, they are considered more reliable than administrative data and are therefore used to update industry information from the BR.

2.2 Industry classification systems

Industrial classification systems are critical for understanding the types of economic activity taking place a particular moment in time, and to analyzing how this activity evolves over time. Classification systems generally define categories in which to group different types of economic activity. Activity can be categorized based on a production concept, in which case establishments that share a common production process are assigned to the same industry. An alternative market or demand-based concept classifies establishments together if their primary products are better substitutes for each other than the products of other establishments.

The SIC system was first implemented in the 1930s. Over time, economic activity evolved so that the SIC categories no longer provided the requisite accuracy or consistency necessary for rigorous analysis. According to a comprehensive 1995 study of the 1987 SIC system undertaken by the US Economic Classification Policy Committee described in Triplett et al. (1995), SIC reflected "a potpourri of conflicting classification concepts" (p. 148). Estimates from the study suggested that approximately 25 percent of SIC industries were based on market or demand criteria, just under 20 percent were based on production criteria, and about 20 percent were defined by no recognizable concept whatsoever. The authors caution that these shortcomings mean that qualified interpretations must be applied to any research results based on the 1987 SIC system, especially cross-section or time-series studies that make inter-industry comparisons. They further note that the lack of a production-based classification concept implies that the SIC system is particularly problematic for productivity analyses that inherently assume the same production function across plants in the same industry.

NAICS addresses these issues and provides more accurate information on new and emerging industries. According to Murphy (1998), NAICS was developed with four main goals. First, it

[3] This occurs in 2004 when the BR data classify establishments on both a NAICS 1997 and a NAICS 2002 basis.

classifies economic activity based solely on a production process concept. As such, it is the first industrial classification system in the US that uses a unique economic concept to define industries. Second, it includes many new industries to accommodate the growing importance of services and the proliferation of new categories within this broad sector. The new industries were designed to yield better measures of the types of services being provided and necessitated by the introduction of many new types of activities. For example, cell phone and internet service providers did not exist when the SIC industries were first implemented. In other cases, existing industries were changed to meet the NAICS production process definitions.[4] Third, it attempts to minimize breaks in the time-series properties of industries by maintaining as much consistency as possible with SIC, while still categorizing activity based on a production concept. Fourth, it aims to be compatible with the two-digit level of the International Standard Industrial Classification of All Economic Activities.

The US Census Bureau uses industry classification systems to assign individual establishments to a primary industry. While workers within a single establishment may perform different activities, perhaps within different industries or even sectors, data products based on establishment-level data classify all workers within an establishment to that establishment's industry. These establishment-level data, along with their industry classification, are used in a large number of publications on US employment, productivity, and input-output usage released by the US Census Bureau, the Bureau of Economic Analysis (BEA), and the Bureau of Labor Statistics. While the two main classification systems are SIC and NAICS, there are several different vintages for each of these systems. The data used in this paper include establishments that were classified according to SIC 1972, SIC 1977, and SIC 1987 systems. In addition, the Census created internal bridge codes in 1997 that were designed to map to a unique NAICS 1997 code. In this paper, we refer to these internal bridge codes as SIC 1997 codes. The SIC 1997 codes were created because a number of Census data products continued to be published under the SIC system, even after NAICS was officially implemented in 1997. In reality, these SIC 1997 codes do not always map uniquely to a single NAICS code. NAICS also contains several vintages, including 1997 NAICS, 2002 NAICS, 2007 NAICS, and 2012 NAICS. This paper focuses exclusively on NAICS 1997 and 2002 codes.[5]

Table 1 shows the industry classification systems used in this paper from each of the data sources described above. The second column reports the industry classification system in the LBD for each time period. For example, the LBD classifies establishments using the 1972 industry classification system in the years 1976 through 1978. An important warning to researchers, however, is that there is always some fraction of establishments in a given year that is assigned an industry code

[4] NAICS also introduces additional manufacturing categories to provide accurate measures for new industries. For example, SIC industry code 3559 “Special industry machinery, not elsewhere classified” included semi-conductor machinery manufacturing and broom making machinery. Semi-conductor machinery manufacturing is classified under NAICS 333295.

[5] The Center for Economic Studies at the Census Bureau currently has efforts underway to update the 2002 codes to a NAICS 2012 basis.

from an older vintage of the current system. It should also be clear from Table 1 that assigning a consistent NAICS 2002 code to all establishments involves not only addressing the change from SIC to NAICS, but also all the transitions across different vintages within a system. Columns 3 and 4 report additional information from the BR and the ECs that are used in the analysis.

An important contribution of the EC data is that, as evident in Table 1 , the 1997 EC includes both SIC 1997 and NAICS 1997 information for the same establishment. This information is useful not only for assigning a NAICS code to establishments in the 1997 LBD, but also for constructing a detailed concordance between SIC and NAICS. We use the EC data to construct a concordance that shows the fraction of establishments that maps to any given NAICS code for each SIC code, within establishment size quintiles. In 2002, the EC data provide a 1997 NAICS code as well as a 2002 NAICS code for all establishments in the census. We construct a similar concordance between NAICS 2002 and NAICS 2007 using these data.

Table 1: Data sources and classification systems

Year	LBD original codes	BR codes	Census Codes
1976-1978	SIC72		SIC72, SIC77
1979-1986	SIC77		SIC72, SIC77, SIC87
1987-1996	SIC87		SIC72, SIC77, SIC87
1997	SIC97		SIC97, NAICS 1997
1998	SIC97		
1999	SIC97		
2000	SIC97		
2001	SIC97		
2002	NAICS97		NAICS97, NAICS02
2003	NAICS97		
2004	NAICS97	NAICS97, NAICS02	
2005	NAICS02	NAICS02	
2006		NAICS02	
2007		NAICS02	NAICS02, NAICS07

Notes: It is important to note that the industry codes in the data in a given year often include a mix of multiple vintages. For example, most SIC codes in 1987 are SIC87 codes, but there are always some establishments with older vintage codes. There is no way to distinguish industry code vintages in the data unless there is a particular code that did not exist in a given vintage.

We also employ publicly available data from the 1997 EC. Specifically, the Census Bureau provides data on the number of establishments, employment, payroll, and sales both by SIC industry and by NAICS industry. We use these data to construct concordances that show the fractions of establishments, employment, and sales within each SIC industry that are allocated to different

NAICS industries. We compile this information into a user-friendly format to facilitate its use by a broad audience. The public data are also useful for calculating differences in SIC versus NAICS in 1997 at an aggregate level.[6]

3 The change from SIC to NAICS

As describe in section 2, the SIC system classified establishments into industries using information on the establishments' customers, their production processes, and even whether they were vertically integrated.[7] In contrast, NAICS classifies establishments into industries based on their production process. The production-based methodology used to define NAICS industries has three main implications for how US activity is classified. First, there are significant changes in the industries and sectors in which establishments are classified. Second, establishments that perform headquarter and other support services for a firm are now classified in separate sectors that describe this activity, whereas they were classified in the same sector as the establishments they supported under SIC. Third, establishments that fragment their production process across other establishments pose new classification challenges under NAICS. The rest of this section describes these issues in more detail.

3.1 Industry and sectors changes

The change from SIC to NAICS entailed a change in the definitions of existing industries as well as the creation of new industries, most of which are in services. In this section, we use information from the publicly available data of the 1997 EC to summarize the main industry and sector differences between SIC and NAICS. These definitional changes are not easy to address with simple concordances since it is often the case that one SIC industry has establishments that were re-classified into multiple different NAICS industries and even sectors. There are 93 SIC industries out of 891 industries for which at least some fraction of activity was re-classified into other sectors: 3 industries in mining, 1 in construction, 7 in manufacturing, 19 in transportation warehousing and utilities (TWU), 23 in wholesale trade (WT), 9 in retail trade (RT), 9 in finance, insurance and real estate (FIRE), and 22 in services. Among these, 33 four-digit SIC industries transferred to different sectors in their entirety. For example, newspapers and periodicals, which were classified as manufacturing under SIC, were reclassified as services under NAICS. The remaining 66 industries had some portion of their establishments re-classified into other sectors.

Even when the implementation of NAICS does not result in sectoral-level changes, it still induces

[6] The concordances are available here: http://faculty.tuck.dartmouth.edu/teresa-fort/data. The raw 1997 EC data on a SIC and a NAICS basis are available here: http://www.census.gov/epcd/ec97brdg/index.html.

[7] For example, chain made from purchased wire (SIC 3496) was distinct from chain made from forged steel (SIC 3462) because of the difference in the physical production process. Meat packing establishments that sold meat products from animals they slaughtered (SIC 2011) were distinct from establishments that made the same types of products from purchased carcasses (SIC 2013).

significant reclassification of activity within sectors. Specifically, 321 SIC industries map to more than one NAICS industry. Manufacturing was particularly affected by these types of changes, with 134 SIC industries (out of 458 total SIC manufacturing industries) that map to multiple NAICS codes. We provide additional information on all the changes between SIC 1987 and NAICS 1997 codes in an online concordance. The concordance is based on public data from the 1997 EC and provides the fraction of establishments, employment, and sales for each SIC-NAICS relationship.[8]

3.2 Auxiliary Establishments

Another important distinction between SIC and NAICS is the treatment of headquarter and auxiliary establishments. Auxiliary establishments are defined as those establishments primarily serving other establishments of the same enterprise. Examples of auxiliary establishments include management, warehousing, data processing, and R&D. Under SIC, auxiliary establishments were classified in the primary industry of the establishments that they served. In contrast, NAICS classifies these establishments in a number of different industries and sectors, depending upon the types of services the establishments actually provide. Table 2 lists all the possible NAICS industries to which an establishment classified as an auxiliary under SIC could be assigned. As an example of the change between SIC and NAICS, an R&D establishment that primarily engages in research for the manufacturing establishments in its firm would be classified in those manufacturing establishment industry under SIC, but in *Scientific Research & Development* (5417) under NAICS.

A significant amount of employment was affected by this change. In 1997, there were approximately 48,000 Auxiliary establishments that employed 3.3 million workers. Of these auxiliaries, about 11,000 establishments and 1.2 million workers were classified in manufacturing under SIC. One of the most important new auxiliary sectors under NAICS is the *Management of Companies and Enterprises* (sector 55). This sector had about 47,000 establishments and employed 2.6 million employees in 1997. In 1997, 35,000 establishments and 2.5 million employees were reclassified into *Corporate, subsidiary, and regional managing offices* (551114). These employees had previously been classified in a number of different sectors, including manufacturing, wholesale and retail.[9]

3.3 Outsourcing and production fragmentation with NAICS

A conceptual issue with applying NAICS is how to treat establishments that fragment their production process across locations. The contracting out of some or parts of the production process

[8] The online data are available here http://faculty.tuck.dartmouth.edu/teresa-fort/data. The raw data from the 1997 Economic Census are downloadable here http://www.census.gov/epcd/ec97brdg/.

[9] See http://www.census.gov/epcd/ec97brdg/E97B1551.HTM#5511 for a breakdown. See http://www.census.gov/epcd/ec97sic/E97SusL.HTM#LF for more details on auxiliaries generally. Prior to 2002, auxiliaries are generally assigned a partial (e.g., two or three digit code) in the LBD. In addition, the TOC variable often identifies them as auxiliaries. Use of the FK codes should address these issues. Note that the 1987 and 1992 CM both contain the auxiliaries, though they are assigned partial codes in the FIRE sector.

Table 2: Auxilliary establishments under NAICS and 1997 establishments, shipments, payroll, and employment

NAICS code	NAICS description	Estabs	Shipments	Payroll	Empl
484	Truck transportation	1,084	928	4,008	74
4931	Warehousing & storage	4,800	4,109	17,749	327
514210	Data processing services	387	331	1,431	26
5411	Legal services	68	58	251	5
5412	Accounting, tax returns, payroll services, etc	1,285	1,100	4,751	88
5417	Scientific research & development services	1,048	897	3,875	71
5418	Advertising & related services	409	350	1,512	28
551114	Corporate, subsidiary, & regional managing offices	35,263	30,184	130,390	2,403
5613	Employment services	196	168	725	13
56161	Investigation, guard, & armored car services	46	39	170	3
5617	Services to buildings & dwellings	76	65	281	5
811	Repair & maintenance	712	609	2,633	49
949999	Unclassified auxiliary establishments	2,819	2,413	10,424	192

Notes: Based on information from the publicly available 1997 Economic Census data. Shipments and payroll in millions us$s. Employment in 1000s.

presents a challenge under NAICS since it generally entails a change in the production process being performed by individual establishments. A non-trivial fraction of establishments has some or all of its physical transformation activities performed by other establishments in another location. This type of fragmentation is often done via Contract Manufacturing Services (CMS) purchases. CMS purchases entail an arrangement in which the purchaser provides design and production criteria to a manufacturer who performs the physical transformation activities, generally on materials or inputs specified by the purchaser. See Fort (2014) for a detailed description of CMS purchases and what they entail.

When CMS are purchased by manufacturing establishments, measures of the inputs and employment that an establishment used to produce its final products may be incorrect. Specifically, the CM asks establishments to list only those expenditures on contract work that are done on materials owned by the establishment. The cost of contract work performed using inputs that were not supplied by the purchasing establishment is therefore not included in input costs. When CMS are purchased by non-manufacturing establishments, it is even more difficult to ascertain the value of those inputs. This is because non-manufacturing Census surveys collect limited or no information on input use. Non-manufacturing establishments that purchase CMS, often referred to as factoryless goods producers (FGPs), generally specify the design and production specifications for their products, but do not perform the majority of the physical transformation activities. The existence of FGPs in the wholesale sector is well-documented (see, for example, Bernard and Fort (2013);

Bayard et al. (2013); Bernard and Fort (2015)). The extent to which FGPs exist in other sectors of the economy remains an open question. The Census Bureau has efforts underway to identify FGPs in the wholesale sector and may transfer them to manufacturing eventually, with a flag variable to distinguish them from more traditional manufacturers. An important implication of this transfer is that it would provide richer information on the types of inputs used to produce goods in the US economy. For example, the input-output tables constructed by the Bureau of Economic Analysis rely on the detailed information collected in the CM about use of inputs by plants in different industries. If FGPs were re-classified from wholesale to manufacturing, this would provide statistical agencies and researchers much more detailed information about the values and types of inputs used to produce goods sold by FGPs.

4 Classifying activity on a consistent basis

The change from SIC to NAICS makes it difficult to compare differences in activity across time periods spanning the two classification systems. The same establishment performing the same activities may be classified in entirely different sectors, depending upon whether that classification is made based on a SIC or a NAICS code. These differences represent a significant hurdle to efforts to depict an accurate understanding of secular changes in US activities across sectors and industries. They also complicate any time series analysis in which industry differences matter. Although the first NAICS codes were introduced in 1997, NAICS was not incorporated into the LBD until 2002. Prior to 2002, the LBD is classified on a four-digit SIC basis. As a result, any analysis that uses these data pre and post 2002 and relies on industry information must address this change in the classification system.

The LBD does include “Best” SIC and “Best” NAICS codes. These codes are the modal SIC and modal NAICS code for a particular establishment, or the SIC or NAICS code from the most recent Economic Census. While the “Best” codes may be appropriate in some studies, they are clearly not in others. First, they do not allow for any industry switching since every establishment will have at most one “Best” SIC and one “Best” NAICS code. Second, the “Best” SIC will only be populated for establishments that existed prior to 2002, while “Best” NAICS will only be populated for establishments that exist post 2001. These “Best” variables therefore do not make any adjustments to address the SIC-NAICS transition, so that it continues to pose the same challenge to any time series analysis.

To address these issues, we develop a methodology to assign a consistent, NAICS 2002 industry code to every establishment in the LBD from 1976 to 2009. The codes that result from this methodology, called the Fort-Klimek (FK) codes are designed to: 1) improve the accuracy of the industry codes in the LBD; 2) replace missing and partial industry codes; 3) provide a continuous industry code basis for the entire LBD and; 4) to minimize “industry switching” that might be induced by random assignments of NAICS codes that do not map uniquely from SIC codes.

To assign a consistent industry code, we first use the longitudinal information in the LBD to fill in all missing or incomplete codes. Second, we use detailed concordances to assign all NAICS codes that map uniquely from an establishment's SIC code. Third, we use the longitudinal structure of the LBD to assign NAICS code to an establishment with an SIC code that maps to multiple NAICS codes, but for which the NAICS code assigned to that establishment in a post 2001 year is consistent with one of the mappings. Fourth, when the longitudinal information is unavailable or insufficient, we use detailed concordances and random assignment techniques to assign a NAICS code to an establishment with an SIC code that maps to multiple NAICS codes. Note that our goal *is not* to identify cases that might be considered errors (e.g. establishment A is classified in activity 1 then 2 then back to 1 where all are legitimate industries, though assignment to 2 is possibly an error). In that sense industry codes are allowed to change overtime for reasons that include changes in activity, as well as errors in the underlying data. Additional details of the assignment methodology are in appendix section A as well as in Fort and Klimek (2014).

5 Aggregate Implications

The differences between SIC and NAICS described in the previous section have significant implications for the measured share of aggregate activity across sectors. In this section, we describe how the change affected the distribution of US employment across the following broad categories: a) construction, b) finance, insurance and real estate, c) manufacturing, d) wholesale trade, e) retail trade, f) services, g) transportation, warehousing and utilities, and h) mining. To gain an initial sense of the magnitudes, we first report results based on the publicly available concordance constructed from the 1997 EC. These data suggest that 7.6 percent of US manufacturing employment, equal to approximately 1.4 million jobs, was reclassified to services under NAICS in 1997. More striking is retail trade, in which approximately 8.68 million jobs were reclassified into services, almost 40 percent of the size of the sector under SIC. Wholesale trade also saw significant changes, with approximately 1.8 million jobs re-classified into other sectors. The largest fraction of these jobs were reclassified into retail trade, although a significant fraction also went to services.

These statistics provide a general sense of the implications of the change from SIC to NAICS in a static setting. The dynamic implications, however, may be quite different. For example, the types of activities (and establishments that perform them) re-classified from manufacturing to services may grow and evolve quite differently from those that remain in manufacturing. To gain a sense of these effects, we use the FK codes to track establishments over time using a consistent industry classification system. These codes allow us to measure activity across entering, exiting, and continuing establishments.

Figure 1 illustrates the differences between SIC and NAICS by plotting the level of employment in a panel for broad categories over time. The dashed line in each panel represents the level of employment in a sector based on the contemporaneous industry code in the LBD. As described earlier, the LBD is on a SIC basis from 1976 to 2001 and on a NAICS basis from 2002 to the present. The dashed line therefore depicts employment in each sector on a SIC basis between 1976 and 2001 and on a NAICS basis from 2002 to 2007. The solid line in each panel depicts the level of employment in each sector when using the consistent FK 2002 NAICS codes. In some sectors, the transition from SIC to NAICS did not lead to significant changes. This is evident for construction and finance, insurance and real estate, where the two series largely overlap. In contrast, the change from SIC to NAICS resulted in a large drop in the measure of employment in manufacturing, wholesale trade, and retail trade that is offset by a large increase in measured employment in services. In sum, a significant portion of establishments were re-classified from manufacturing, wholesale, and retail into services.

How large are these changes? Focusing first on manufacturing, the contemporaneous codes suggest that employment dropped by 18.8 percent from 2001 to 2002. In contrast, the FK codes show a decline of 7.86 percent. While still a significant drop, this difference implies that the change from SIC to NAICS led to an additional 10.9 percentage point decline in manufacturing employment–a large effect equal to 140 percent of the actual employment change. Similar patterns are evident in other sectors. In wholesale, the contemporaneous codes suggest a 17 percent decline in employment, while the FK codes show that employment increased by five percent. In retail trade, the contemporaneous codes depict a 35.7 percent drop in employment, while the FK codes show only a 2.8 percent decline. The majority of jobs "lost" from these sectors as a result of the industry classification change simply represent a reclassification into services. The contemporaneous industry codes suggest that services employment increased by 30.8 percent from 2001 to 2002, while the FK codes show that services employment actually decreased by 1.2 percent for that period.

6 Research and Measurement Implications

In this section, we discuss potential implications of the change from SIC to NAICS for measurement and research. The methodology employed in this paper exploits the time series and establishment nature of the data to ensure that information on both the aggregate transition probabilities and the individual establishment-level information are incorporated into NAICS assignments in years prior to 1997. Most importantly for research, our approach allows for heterogeneity in the transition from SIC to NAICS by establishment size and by geography.

While the implications of the transition from SIC to NAICS are generally known in US statistical agencies, there is no clean way to address the issues. All the agencies have been careful to modify time series publications to correct for the differences implied by the change.[10] However, the problem

[10] For example, see Yuskavage (2007) for a description of the conversion of timeseries data from SIC to NAICS by

may still affect the time series properties of the data. For example, Pierce and Schott (2015) use BLS data to note that US manufacturing employment dropped 18 percent from March 2001 to March 2007. In contrast, the consistently coded establishment-level data based on the FK codes suggest the employment decline over this period was 15.4 percent. The difference in these calculated changes corresponds to over 400,000 manufacturing jobs. This is just one example, yet it is illustrative of the complexities involved in the transition from SIC to NAICS and the need to remember that comparisons across the classification systems likely involve some measurement error. When we say US manufacturing has declined and that services are increasingly dominant, an important caveat is that our industry classification system now classifies a significant portion of economic activity as non-manufacturing that was formerly classified as manufacturing under the SIC system.

The transition from SIC to NAICS may also lead to measurement error that biases research estimates. The most obvious potential issue is a failure to capture economic activity simply because it was reclassified into another sector outside the scope of a particular study. Figure 1c clearly shows how this could lead to significant bias in a study of US manufacturing. For example, if a researcher assessed the impact of an event that took place between 1997 and 2002 without addressing the reclassification from SIC to NAICS, she might incorrectly conclude that the event led to a significant decline in manufacturing activity. The most straightforward way to address this source of bias is for researchers to convert all the data to one classification system before beginning the analysis (e.g., as is done in Autor et al. (2013)). The aggregate concordances we provide with this paper can be used for this procedure. Unfortunately, this approach may still lead to measurement error since it relies on aggregate information to reclassify the data. In reality, the transition from SIC to NAICS, and the shares of establishments, sales, and employment that were re-classified both within and across sectors may vary significantly by firm size and/or by geography. To the extent that there is heterogeneity within industries, relying on the aggregate concordances may lead to over or under estimates of the parameters of interest.

Another approach, employed by Pierce and Schott (2015), is to exclude all industries with transitions across sectors from a particular analysis. This approach ensures that the statistics derived reflect the economic activity of the included industries, but it omits those industries that may have very different and also important dynamics. The new industry codes developed in this paper allow researchers to conduct a comprehensive study that captures all types of reallocation and churn. They also allow for an establishment-level analysis in which there is potential heterogeneity in establishments' reclassification status across the different systems.

Although not yet studied, another difference between SIC and NAICS with potential research implications is due to the within-industry changes across the two systems. Because NAICS groups establishments together based on the activities that the establishment actually performs, within

the Bureau of Economic Analysis.

industry dispersion in characteristics such as size and productivity may be smaller under NAICS relative to SIC. We plan to explore this possibility across industries and over time.

7 Conclusion

This paper develops and implements a new methodology to assign consistent NAICS 2002 codes to every establishment in the LBD from 1976 to 2009. We use these codes to show that the transition from SIC to NAICS entailed a significant reclassification of establishments from manufacturing, wholesale trade, and retail trade into services, without any change in the underlying economic activity performed by those establishments. This reclassification has clear implications for how we measure US employment composition. It also reveals the potential for bias in time series research of the US economy that spans the transition from SIC to NAICS. Although this transition was officially implemented in 1997, it took multiple years to implement, with a number of data products reflecting the change only in 2002. Studies that span this time frame should be careful to address this change in industry classification systems. At a minimum, researchers should convert their data to a single system-a process that can be done using the aggregate concordances that accompany this paper. Those with access to the establishment-level Census data should consider using the consistently classified fk codes.

References

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price, "Import Competition and the Great U.S. Employment Sag," *Journal of Labor Economics*, **forthcoming. 1**

Autor, David H., David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition," *American Economic Review*, **2013,** *103* **(6), 2121–2168. 1, 1, 6**

Bayard, Kimberly, David Byrne, and Dominic Smith, "The Scope of U.S. Factorlyess Manufacturing," mimeo, Federal Reserve Board 2013. 3.3

Bernard, Andrew B. and Teresa C. Fort, "Factorlyess Goods Producers in the US," Working Paper 19396, NBER June 2013. 3.3

_ and _ , “Factoryless Goods Producing Firms,” *American Economic Review: Papers & Proceedings*, 2015, *105* (5), 518–23. 3.3

Clarke, R., “SICs as Delineators of Economic Markets,” *Journal of Business*, 1989, *62*, 17–31. 1

Fertuck, L., “A Test of Industry Indices Based on the SIC Codes,” *Journal of Financial and Quantitative Analysis*, 1975, *10*, 837–848. 1

Fort, Teresa C., “Technology and Production Fragmentation: Domestic versus Foreign Sourcing,” mimeo, Dartmouth Tuck School of Business 2014. 3.3

_ and Shawn D. Klimek, “Detailed Methodology for the Fort-Klimek NAICS codes,” mimeo, Tuck School of Business at Dartmouth 2014. 4

Guenther, D. A. and A. J. Rosman, “Differences between Compustat and CRSP SIC Codes and Related Effects on Research,” *Journal of Accounting and Economics*, 1994, *18*, 115–128. 1

Jarmin, Ron S. and Javier Miranda, “The Longitudinal Business Database,” CES Working Paper 02-17 2002. 2.1

Murphy, John, “Introducing the North American Industry Classification System,” *Monthly Labor Review*, 1998, pp. 43–47. 2.2

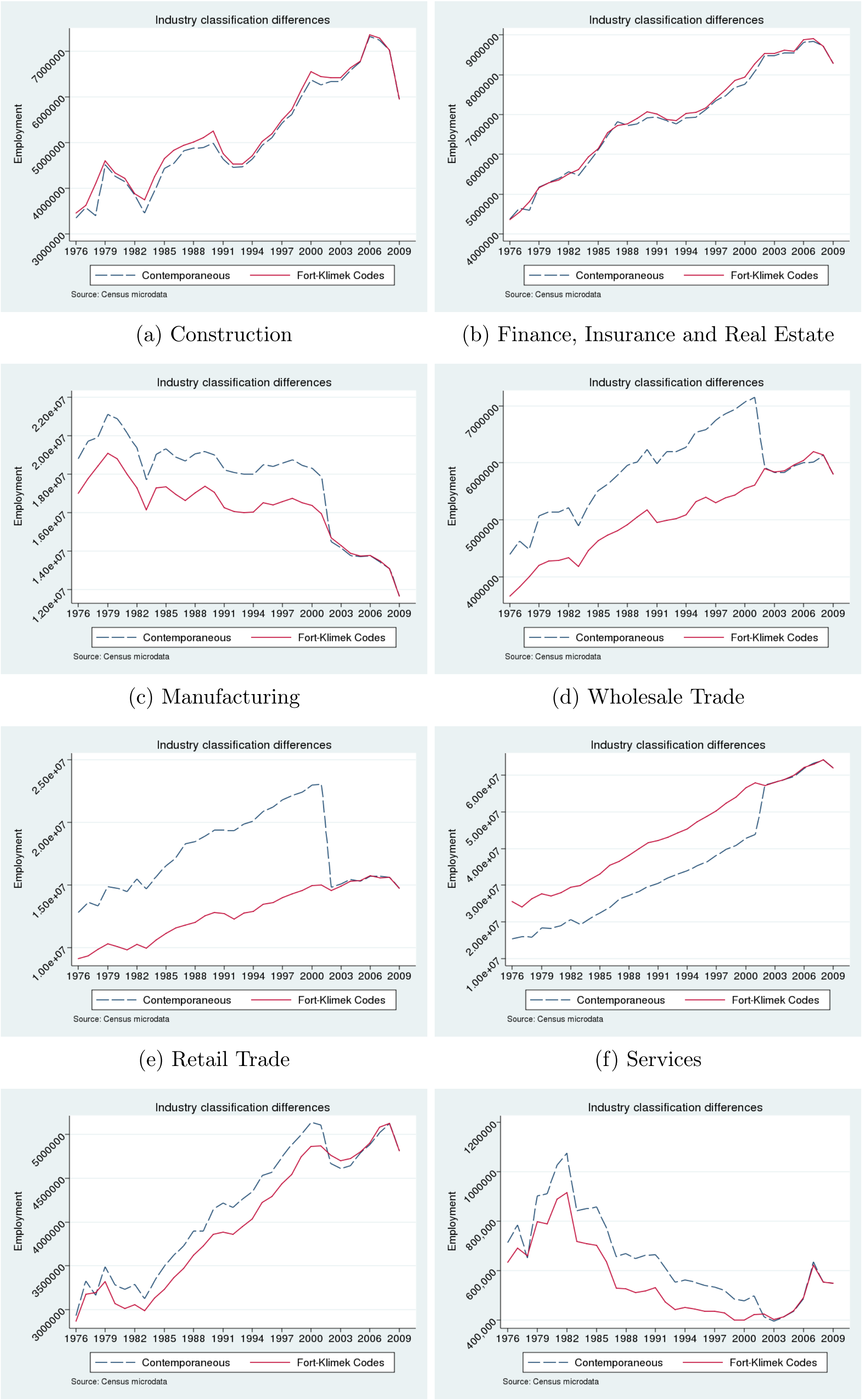
Pierce, Justin and Peter K. Schott, “The Surprisingly Swift Decline of U.S. Manufacturing Employment,” Working Paper 18655, NBER 2015. 1, 1, 6

Rajan, Raghuram G. and Luigi Zingales, “Financial Dependence and Growth,” *The American Economic Review*, June 1998, *88* (3), 559–586. 1

Triplett, Jack E., D. Mark Kennet, Ron Jarmin, and Frank M. Gollop, “Do Industrial Classifications Need Re-Inventing? An Analysis of the Relevance of the U.S. SIC System for Productivity Research,” in “Proceedings of the 6th ASIS SIG/CR Classification Research Workshop” 1995, pp. 145–172. 1, 2.2

Yuskavage, Robert, "Converting historical industry time series data from SIC to NAICS," in "Federal Committee on Statistical Methodology 2007 Research Conference" 2007. 10

Figure 1: Employment Differences Between SIC and NAICS



(a) Construction

(b) Finance, Insurance and Real Estate

(c) Manufacturing

(d) Wholesale Trade

(e) Retail Trade

(f) Services

(g) Transportation, Warehousing, Utilities

(h) Mining

Notes: Figures show aggregate employment at the sectoral level using contemporaneous industry codes in the LBD (SIC prior to 2002 and NAICS starting in 2002) and Fort-Klimek industry codes.

Appendix

A Assignment Methodology

In this section, we provide an overview of the methodology used to assign a unique NAICS 2002 industry code to every establishment in the LBD for 1976 to 2007. These codes are designed to: 1) improve the accuracy of the industry codes in the LBD; 2) replace missing and partial industry codes; 3) provide a continuous industry code basis for the entire LBD and; 4) to minimize "industry switching" that might be induced by random assignments of NAICS codes that do not map uniquely from SIC codes. The newly assigned NAICS 2002 codes will be useful for projects that require consistent establishment industry codes across multiple years of data.

This is done first by using longitudinal information to fill missing or incomplete information, and second by using cross-walks and random assignments when longitudinal information is unavailable or insufficient. Note that our goal IS NOT to identify cases that might be considered errors (e.g. establishment A is classified in activity 1 then 2 then back to 1 where all are legitimate industries, though assignment to 2 is possibly an error). In that sense industry codes are allowed to change over time for reasons that include changes in activity, as well as errors in the underlying data. The remainder of this section describes the distinct steps involved in assigning the FK codes.

A.1 Concordance construction

We construct a number of concordances between different vintages of SIC and NAICS codes using: a) the 1997 and 2002 Census data (for example, the 1997 census includes a SIC87 code as well as a NAICS97 for each establishment); b) official concordances posted on the Census Bureau website; c) implemented concordances from the Census Bureau website; d) detailed concordances obtained from S. Klimek; e) a County Business Patterns (CBP) concordance; f) a concordance documenting implementation problems between NAICS97 and NAICS02 codes; g) a scope concordance that documents all NAICS97 AND NAICS02 codes that are out of scope for the 2007 Economic Census; and h) official concordances published in the SIC87 and SIC77 manuals.

We use these concordances to create a master concordance for the following classification changes: a. SIC72 to SIC77 b. SIC77 to SIC87 c. SIC87 to SIC92 d. SIC92 to NAICS97 e. NAICS97 to NAICS02 The master concordances reconcile discrepancies between the various concordances listed above. For example, in some instances the official concordance mapped SIC industry A to NAICS industries 1, 2 and 3, while in practice SIC A only ever maps to NAICS 1 and 2. Additional details are provided in the NASS memo.

A.2 Code assignments

We assign a six digit NAICS02 code to all establishments in the LBD except for: a. Deaths b. Flaga= I c. Flaga=s & pay= . d. Flaga=6 & pay= . These exceptions mean that FK codes are assigned to records that are payroll active. We assign codes whenever the LBD code: a) another vintage code (e.g., SIC or NAICS97); b) is not a valid NAICS02 code; c) is missing; d) is a partial/incomplete code. Columns 2-4 in Table XXXX list the yearly industry classification vintage in the original LBD data, as well as the codes used from the BR and the Economic Censuses. The final column shows the intermediate code assignments necessary to assign a NAICS02 code in each year.

To assign NAICS02 codes, we divide the LBD data into four time periods that span at least one industry code change and begin/end with census years. These periods are: 1) 2002-2009, 2) 1997-2002, 3)1987-1997, and 4) 1976-1987. We begin assignments with the 2002-2009 data and work backwards. This methodology has two benefits: 1) The most recent data already have NAICS02 codes and therefore provide valuable information for assigning NAICS02 codes to establishments that also existed prior to 2002 and were assigned earlier vintage codes in those earlier years; 2) More recent data require fewer concordances to assign a NAICS02 code. Although each of these periods has its own unique issues, I follow the same general assignment steps for all of them.

The codes we assign are called fk_indcode (where FK stands for Fort-Klimek code). We document all of the assignments using a source, method, and year flag. The source flag describes where the industry code is from (see Table 2). I use information from the Longitudinal Business Database (LBD), Business Register (BR), and the Economic Censuses. When the LBD and Censuses have the same establishment assigned to a different industry, the fk_indcode corresponds to the industry code in the Economic Census data.

The method flag describes how the code was assigned (see Table 3), and the year flag describes the year from which the assigned code originated. Note that when multiple steps are involved, these flags only capture the last step.

All non-original codes have a source flag value of "FK". These FK assignments may be the result of: a) a unique mapping in the concordances—method flag="B"; b) use of the panel nature of the data to "roll" an industry code assigned to a particular establishment in one year to another year in which that code is missing—method flag values C through I, M or N; c) out of scope code (see section D.1 below); or d) random assignment—method flag R.

Random assignments are only done when it is not possible to use a unique concordance or the panel nature of the data. In these cases, an industry code in year t maps to multiple codes in year t+1. The random assignment concordances (details in the NASS Methodology memo) include the share of establishments by size class (total of 5 size classes) that map to all the potential new vintage codes for one old vintage code that splits. To assign a new vintage code to an establishment with an old vintage code that splits, I draw a random share value from the uniform distribution. I assign an establishment the new vintage code with a calculated share that corresponds to the establishment's randomly drawn share. The variable fk_indcode_splits denotes the potential number of new vintage codes to which the old vintage code could have been assigned. After randomly assigning a code to one establishment, I apply the assigned code to all years for which that establishment existed and for which there is no other conflicting industry information.

The random flag indicates whether a particular lbdnum had any splitting codes. Table 4 shows the share of lbdnums in each year that have some element of randomness, as well as those records that were completely randomly assigned (very_random_share). The latter denotes records in which there was no industry information available for the establishment in any year from any of the data sources used (LBD, BR, and Economic Censuses). The number of splits for these establishments is generally over 900, and I do NOT recommend using these codes for any research or analysis.

A.3 Auxiliary establishments

Under the SIC classification scheme, auxiliary establishments were classified under the industry that they served. For example, a headquarter establishment serving manufacturing was assigned a partial SIC manufacturing code. Under NAICS, an establishment is assigned to the industry that best categorizes what the establishment does. All headquarter establishments are therefore classified under 551114.

This issue has been partially fixed in the current FK codes. In particular, a) all establishments with an FK NAICS02 in 2002 that starts with 55 have their full FK NAICS02 code rolled back to all years (methodology S); b) all establishments with an FK NAICS02 code in an auxiliary industry (auxiliary industries are identified from the 2002 BR aux_flag) have their full FK NAICS02 2002 code rolled back for all years if one of the following is true: i) the establishment is one of the Census of Auxiliaries, ii) the establishment has TOC code that starts with 8, or iii) the establishment has a SIC auxiliary bridge code between 1997 and 2001 (methodology T); c) all establishments in the Census of Auxiliaries (did not exist in 2002 or have an FK NAICS02 code that is not in a potential auxiliary industry) are re-assigned to a partial NAICS02 code that corresponds to the auxiliary type information in the most recent Census of Auxiliaries; d) all establishments that have an auxiliary SIC bridge code between 1997 and 2001 are assigned their full FK NAICS02 code, or if they did not exist in 2002, they are assigned the partial NAICS02 code that corresponds to the auxiliary type denoted by the bridge code (methodology V); e) all establishments that did not exist in 2002 and have TOC=81 in the majority of their years of existence are assigned to 551114 (methodology W).

Ultimately, we plan to randomly assign full codes to the partially assigned NAICS02 codes described above. However, the current data now have the following partial codes:

Auxiliary Partials to expect in pre-2002 years: 1. 551110 2. 811000 3. 484000 4. 493000

Researchers must address this issue since these will always be full codes, mapping to multiple different industries, in post-2001 data. There will also be full codes for these partials in the pre-2002 data. One possible fix is to reassign all full codes to these partial industries. Another option is for researchers to randomly assign these partials to full codes.

A.4 Notes to users

In addition to addressing the existence of partial auxiliary codes described in the previous subsection, there are several important issues about which data users should be aware.

First, when a new version of the LBD is created, lbdnums are re-generated, making it necessary to re-run all of the NASS programs. When these programs are re-run, the randomly assigned codes in the data will change. (Ideally we would do a large number of implicate assignments and choose the most stable codes, but this is not possible at present.) As a result, research results that rely too heavily on random assignments (and especially the very random codes), could well change. See 2-4 below for input on dealing with the randomness.

Second, the random flag indicates whether a particular lbdnum had any splitting codes. Table 4 shows the share of lbdnums in each year that have some element of randomness, as well as those records that were completely randomly assigned (very_random_share). The latter denotes records in which there was no industry information available for the establishment in any year from any of the data sources used (LBD, BR, and Economic Censuses). The number of splits for these establishments is generally over 900, and their industry codes should not be used for analyses.

Third, it is important to note that the count of industry splits only provides one measure of the "randomness" of an assignment. Most random assignments are made using the census data concordances in which each potential code has its own share (denoting the share of establishments with industry code A that map to industry code B in the Census year). It is possible for a code A to map to both codes B and C 50% of the time. However, it is also possible for code A to map to code B 90% of the time, while it maps to code C only 10% of the time. In both cases, the count of splits will equal 2, however, the former might be considered more random than the latter. This distinction is not captured by the current flag system in the data.

Fourth, the split counts variables may overstate the number splits. This occurs when a sic code splits to a naics97 code but then all possible naics97 codes map to a single naics02. For example, I know this happens with the out of scope 9xxxxx sic codes.

Fifth, all industry codes that are out of scope in the 2007 Economic Census are assigned partial "scope" codes. (Table to be added to this documentation.) These codes range from 2 to 5 digits with zeros appended to create a six digit code. For example, an establishment in agriculture is coded as 110000. This methodology was adopted because we cannot follow these establishments consistently after 2007 (or earlier, since most out of scope codes in 2007 were also out of scope in

2002 and 1997). In addition, we rely on the 2007 Economic Census to construct random assignment concordances so our methodology cannot be applied to codes that go out of scope.

Finally, there are some industry codes that come in to scope between 1976 and 2007. While we maintain these codes with full detail, we note that these out of scope codes will be under-represented in the years for which they were out of scope. This is because are random assignment methodology relies on the Economic Census data to construct concordances. As a result, the out of scope codes will not be in these concordances, so any establishment that should have been assigned one of these codes will definitely have an incorrect code.

A.4.1 Additional notes on industry codes in the LBD

The LBD has a "BEST" SIC and a "BEST" NAICS code variable. These variables are based either on the modal industry code, or on the code in the closest economic census. As a result, an establishment can only ever have at most one BEST SIC and one BEST NAICS code. This means that an change in the establishment's activity over time that changes its industry classification will not be captured by this variable. In addition, the BEST NAICS variable will only be populated for establishments that exist post 2001, while the BEST SIC variable will only be populated for establishments that exist prior to 2002. Researchers should bear these facts in mind when deciding on the appropriate industry classification variable to employ in any particular study.

B Public Concordances

We provide a concordance between 1987 SIC codes and 1997 NAICS codes based on publicly available data from the 1997 Economic Census. This census collected industry information for the same establishments on both an SIC and a NAICS basis, so it is possible to construct a construct a concordance between the two classification systems that measures the fraction of activity assigned to different industries. The data are available for download here `http://www.census.gov/epcd/ec97brdg/index.html`, but the downloaded files are not in a user-friendly format. We provide the code and final dataset necessary to convert the public data into a usable format.

It is important to note that the public data does have a certain number of suppressed cells. While the number of establishments is always available, the amount of employment, payroll, and value of shipments are sometimes unavailable at detailed levels. We impute these values by assigning the average amount of activity per establishment at higher levels of aggregation to the number of establishments at the more detailed levels. This approach means that summing the detailed industry amounts we report to a higher level of aggregation may lead to totals that are higher or lower than the reported aggregate amounts. In principle, one should be able to calculate the average amount of activity only for the suppressed establishments. Unfortunately, the detailed activity amounts in the public tables do not sum up to the aggregated totals. These discrepancies mean that one cannot reliably use the published totals to infer the total amount of aggregate activity that corresponds to suppressed cells.

We also provide concordances between SIC 1972 and SIC 1977 and between SIC 1977 and SIC 1987 concordances. These files were manually typed in from published manuals kept in the Census Bureau.