Methodological Appendix for FDI in U.S. Metro Areas: The Geography of Jobs in Foreign-Owned Establishments

Designed and Implemented by Nick Marchio[1]

June 2014

Method Overview:

The objective of this methodology is to create estimates of aggregate employment in majority foreign-owned establishments across all 50 U.S. states and D.C. and every metropolitan and micropolitan area, as well as all counties, during the period from 1991 to 2011. To arrive at the estimates, Brookings cleaned establishment level Dun & Bradstreet time-series data and adjusted it to Bureau of Economic Analysis-derived employment control totals at both the state broad-industry and national detailed-industry level. This approach adopts modest assumptions and attempts to stay within the constraints of the two best available data sources of foreign employment over time in the United States. The data are aggregated to the sub-national scale across a combination of data cuts including industry, mode-of-entry, establishment size, trading status, and country of origin. Interested researchers should feel free to contact the authors if they have additional questions about the methodology.

Description of Data Sources:

The first data source used is Walls and Associates' historical Dun and Bradstreet (D&B) database called the *National Establishment Time-Series* (NETS). Since the early nineties, D&B has sought to create a census of U.S. establishments and their employees, which Walls and Associates compiles into an archival database built from annual snapshots taken every January. Walls and Associates cleans, codes, and fits the raw D&B snapshots into a continuous longitudinal data series. D&B's collection technique primarily relies on telephonic surveys. According to Walls and Associates, D&B launches over 100 million calls per year from four call centers to maintain and update its database.[1] To identify new businesses, D&B examines multiple sources including the Yellow Pages, public records, legal and regulatory filings, government registries, third party vendors, corporate announcements, and news reports. All the information is subject to extensive automated quality checks through D&B's proprietary *Intelligence Engine*. NETS uniquely suits the requirements of this study as it is the only available source for time-series

[1] Please see the acknowledgment section of the report for a complete list of colleagues and partners who offered assistance in the development of this project. Along with the report's co-authors Devashree Saha and Kenan Fikri, special thanks go out to Howard Wial, Jonathan Rothwell, Alan Berube, Mark Muro, David Jackson, Don Walls, Sally Thompson, Thomas Anderson, Siddharth Kulkarni, Jesus Trujillo, Jeff Levy, Celine Hu, Junwei Chen, Rachel Jaffe, and Peter Blankenship. Any errors are the authors' alone.

establishment data with consistent industrial and geographic classifications at the sub-metropolitan scale.

The second data source used is the Financial and Operating Statistics data series collected by the Bureau of Economic Analysis (BEA). Though not as detailed as NETS, the BEA's estimates are considered more authoritative and reliable. The specific BEA data series in this current report belong to Table II.A.2 Selected Financial and Operating Data of Majority-Owned Affiliates and Table II.F.7 Employment of Majority-Owned Affiliates, State by Industry.[2] The BEA employment estimates reflect the number of full-time and part-time employees on the payroll at yearend and are seasonally adjusted to reflect normal levels of operation.

D&B / NETS cleansing process:

To ensure the validity of the underlying D&B data, Brookings embarked on a lengthy process to detect and, where possible correct, inaccurate records. This involved vetting and adjusting a number of data fields pertaining to industrial classification, foreign ownership status, geographic information, years registered as active, and reported employment.

Cleaning categorical data

NETS contains industry classifications at the six-digit North American Industrial Classification System (NAICS) and eight-digit Standard Industrial Classifications (SIC) levels and records these each year, providing time variant industry information for each establishment. While six-digit industry information is attractive for its specificity, Brookings analysts determined that because of the inherent subjectivity involved in selecting a single highly detailed industry to represent all firm activity it was both more conservative and meaningful to report industry information at the four-digit NAICS level. To ensure that changes in the establishments' reported industry over time were not driving volatility in the aggregate employment trends, each establishment's primary industry was reassigned based on their modal primary industry across time, with preference given to the more recent mode in cases of a ties.

Brookings made further adjustments to improve and update industry classifications when necessary. Establishments were coded to the nearest applicable four-digit NAICS industry if they belonged to an obsolete industry class, were too small to yield an uninterrupted trend, or fell into an industry outside the coverage of the BEA's FDI tables.[3] In addition, establishments belonging to industries not associated with private sector foreign investment activity were dropped from the database.[4]

Beyond industry revisions, Brookings identified and corrected inconsistencies in key fields representing the establishment's foreign ownership status, geographic code, and years registered as active.[5] In cases of duplicate records, analysts identified establishments that shared a common address and had similarities in their company name and employment size. Establishments that were on balance larger, older, and active during a more recent year were made the primary establishment for that address while the others were removed. This

duplicate identification criterion was designed to be strict enough to allow significant multi-establishment site activity, while still reducing double counting.

Cleaning employment data

Walls and Associates implements a number of steps to clean raw data from D&B, and Brookings implemented a number of additional protocols to improve the quality of the data underlying this analysis. To arrive at a continuous employment series, Walls and Associates, for its part, uses a basic time-series regression technique to re-estimate a portion of the D&B employment figures.[6]

Despite these precautions, as with any database created from millions of human responses, error is inevitable. One common issue in the NETS database is that establishments display a "ratcheting effect" in their reported employment that often leads to jagged movements and in some cases jarring spikes. There are a few explanations for this ratcheting effect. Firms have a tendency to report employment in numbers with factors of five or 10, which may contribute to a "lumpy" growth pattern. Among the largest firms, Walls and Associates attributes ratcheting to a tendency for human resource contacts to report whatever workforce information is on hand, which may not necessarily reflect the real-time employment level. And in other cases, it is possible that the contact may not recognize the distinction between establishment employment, campus-wide employment if it is a multi-establishment site, and enterprise-wide employment across all of the company's establishments.

Brookings analysts, therefore, identified establishments with irregular employment growth patterns or that fell on the upper-tail of the establishment size distributions. These establishments were then cross-checked against an array of online resources ranging from company announcements, news reports, to satellite imagery (i.e., using parking spaces to roughly approximate employment). Since it would be impossible to apply this tedious process to the entire database, an automated program was developed to simulate the human-guided "search and scrub" process and scan the database for outliers with a high probability of distorting the underlying employment distributions. A number of factors were considered in order to identify outliers including the variance of the annual rate of change from the establishment's average rate over time, the establishment's contribution to annual change at the aggregate state and industry level, and the standard error of establishment employment over time. After flagging the outliers, each was adjusted to more closely reflect observed growth patterns through the use of time-series smoothers whose weight and lead-lag count scaled with the magnitude of the observed error. This process was iterated five times. Roughly thirteen-thousand of the more than 3.7 million observations were altered, changing under 0.4 percent of the entire dataset.

Allocation of BEA control totals:

The lengthy effort outlined above to clean and cull the data culminated in an establishment database that was prepared for the dual allocation of BEA-derived employment control totals at

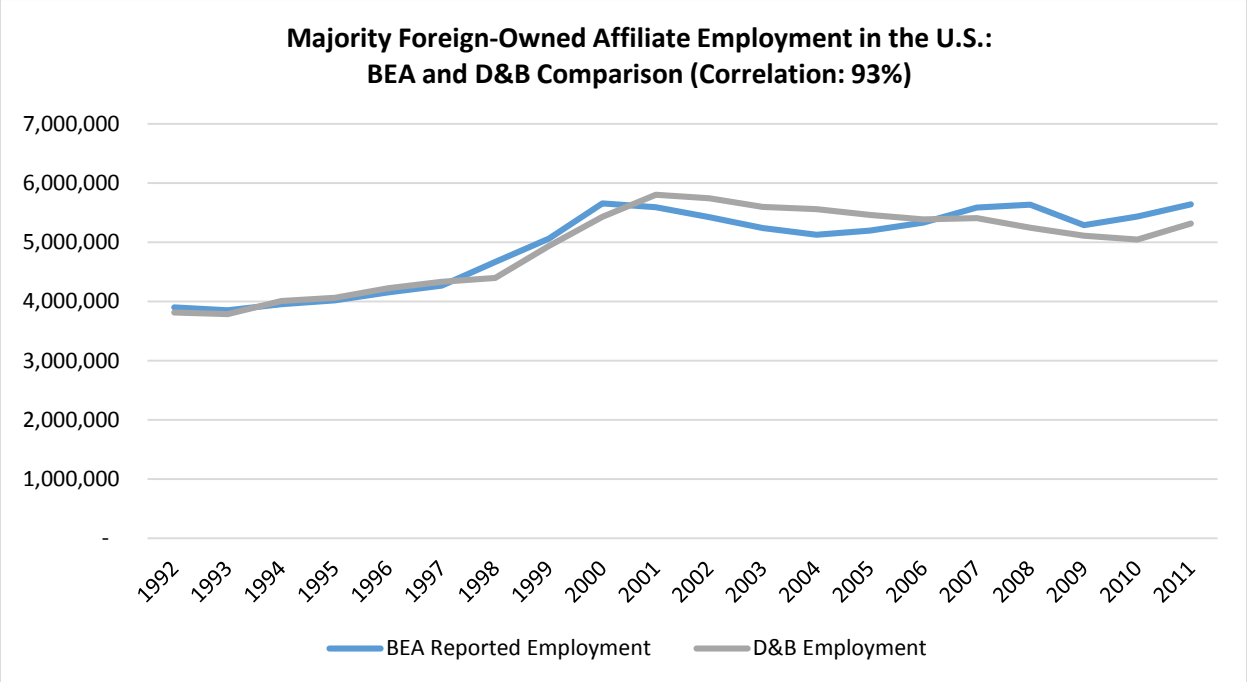
the national detailed and state broad-industry level. The objective for the allocation step was to minimize any spatial biases that may arise in D&B's outreach efforts, to reduce disturbances in reported employment from mistaken responses or typographical errors, and to increase the comparability between the Brookings data and official government statistics reported by the BEA.

BEA and D&B comparability

Before carrying out the allocation, a major concern was the comparability between the NETS and BEA data. While both data sources set out to measure ostensibly the same concept, significant differences in the approach, collection, and classification techniques make them not strictly equivalent.

The relevant BEA definition of foreign ownership is "ownership or control, directly or indirectly, by one foreign investor of 50 percent or more of the voting securities of an incorporated U.S. business enterprise or an equivalent interest in an unincorporated U.S. business enterprise, including a branch or partnership." According to Walls and Associates, D&B defines foreign ownership as an establishment having a headquarters or parent that is located in another country other than that of the establishment. Further, D&B's subsidiary Hoover's, which maintains its client-facing data service, defines a parent company as a corporation that owns more than 50 percent of another corporation's capital stock.

In an attempt to evaluate the definitional comparability, the raw BEA and D&B aggregated data display a high correlation over the last two decades. Based on this comparison, the BEA and D&B definitions appear to be compatible or at the very least sufficiently correlated to merit the use of NETS as a distributor of BEA control totals.



Another factor that potentially limits the comparability between the BEA and D&B data are the relative differences in the assignment of industry classifications. D&B bases the industry classification on the establishment's primary activity—that is, the primary source of revenue at the physical site of production or service operations. The BEA classifies most of its data by industry of affiliate, which is the industry that accounts for the largest portion of the sales across all of the affiliate's establishment sites. Yet in other data series, the BEA classifies data by the industry of sales, which generally approximates establishment level industry classifications, because an affiliate that has an establishment in a secondary industry usually also has sales in that industry unless the product is an input not sold on the open market. Given that many foreign-owned establishments may be part of an enterprise that controls a diversified chain of industrial activity, industry of sales is likely a more accurate means of classification. While not ideal for these reasons, this report still utilizes BEA data classified by industry of affiliate because of its availability at the state level. In cases where analysts observed a large difference between the NETS and BEA industry patterns, Brookings defaulted to a higher industry aggregate when allocating to reduce the effect of differences in coding criteria. Future updates will explore the possibility of using BEA employment coded by industry of sales, which should more closely align with the disaggregated establishment based industrial classifications available in NETS. Further comparisons are discussed later in the appendix.

BEA data preparation

Brookings undertook several steps to prepare the BEA control totals for the allocation process as a number of barriers complicate the utilization of BEA's tables. Detailed industry estimates frequently suffer from suppressions, the levels of aggregation are not separated by sortable industry codes, the surveys use International Survey Industry (ISI) codes which do not perfectly align with NAICS industries, and industries are not comparable over time due to periodic revisions to industry classifications and survey methodologies. As a result, Brookings was forced to implement a number of aggressive procedures to refine the BEA data so that it would fit within a continuous twenty year time-series across a consistent level of industry aggregation.

Analysts used an assortment of imputation techniques to resolve suppressions, including linear interpolation and extrapolation using NETS annual growth rates. Imputations were fitted within the upper and lower bound provided by the BEA, however in some cases violations were allowed. To account for discontinuities that coincided with industry revisions or periodic changes in survey methods, Brookings analysts took the earliest employment estimate available for the latest comparable data series reported by the BEA and extrapolated backwards using NETS annual growth rates or in some cases the annual growth rate within the series being extrapolated. In the case of bank affiliate estimates, which are unavailable for the years preceding 2007, analysts extrapolated employment for NAICS 5221 Depository Credit Intermediation using NETS growth rates moving backwards from the 2007 BEA estimates. Majority-owned state level data was also unavailable for years preceding 2002, so analysts squeezed observations in the affected years to national majority-owned employment totals.

Once imputed and harmonized across time, all employment estimates were adjusted to sum to their respective parent industry employment total. This process yielded two mutually consistent BEA-derived datasets covering state totals by broad ISI industry and national totals by detailed ISI industry. To ensure a consistent level of industry aggregation, the ISI industry aggregates were split into their two-digit and four-digit constituent parts, respectively, using the observed distribution from NETS. All BEA International Surveys Industry (ISI) codes were standardized in terms of NAICS 2012 classifications. Contact the authors for a complete NAICS to ISI correspondence for the 1992 to 1996 SIC-based ISI series and 1997 to 2001, 2002 to 2006, and 2007 to 2011 NAICS-based ISI series.

In cleaning the BEA data, Brookings adds value to existing federal data that up until this point was highly suppressed and lacked internal comparability over the last 20 years. As was stated earlier, at more detailed industry levels the Brookings estimates deviate from the BEA reported values due to differences in industry classification techniques. However, these differences lessen when aggregated to the state, national, and sector level.

Dual allocation procedure and robustness checks

The procedure used to adjust the NETS establishment employment estimates to the BEA-derived control totals is known as allocation, because control totals are "allocated" or shared out to more detailed industry or geographic sub-units based on a distributor. In the case of this study, the two control totals were the BEA-derived national four-digit and state two-digit NAICS employment estimates described in the previous section and the distributors were NETS establishment shares of each total. The control totals were iteratively allocated, first to the establishment shares of one total, and then to the shares of the other. This process was repeated, recalculating shares along the way, until the NETS employment estimates converged on a set of mutually consistent solutions, thus satisfying the constraints of the control totals.[7] While this technique increases confidence in the statistical aggregates, the central tradeoff is that any misestimate of employment at one establishment impacts employment estimates at all establishments within that specific industry and state.

Brookings implemented a series of robustness checks to assess the relative differences between the Brookings, BEA, and D&B / NETS employment series. The following table and charts summarize the key differences.

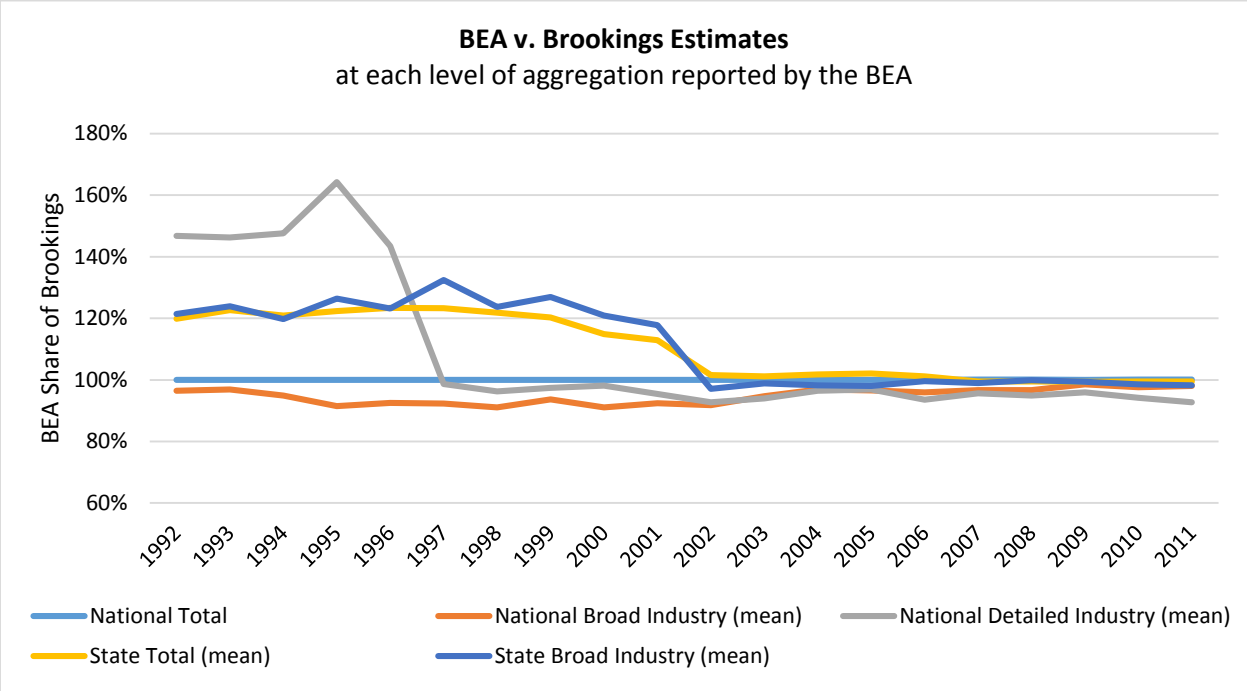
Employment (millions)

	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Brookings	3.90	3.85	3.95	4.02	4.16	4.27	4.67	5.06	5.66	5.59	5.43	5.24	5.13	5.20	5.33	5.58	5.63	5.29	5.43	5.63
BEA	3.90	3.85	3.95	4.02	4.16	4.27	4.67	5.06	5.66	5.59	5.43	5.24	5.13	5.20	5.33	5.59	5.64	5.29	5.44	5.64
D&B	3.82	3.79	4.01	4.06	4.23	4.33	4.40	4.94	5.43	5.80	5.74	5.60	5.56	5.46	5.39	5.41	5.25	5.11	5.05	5.32

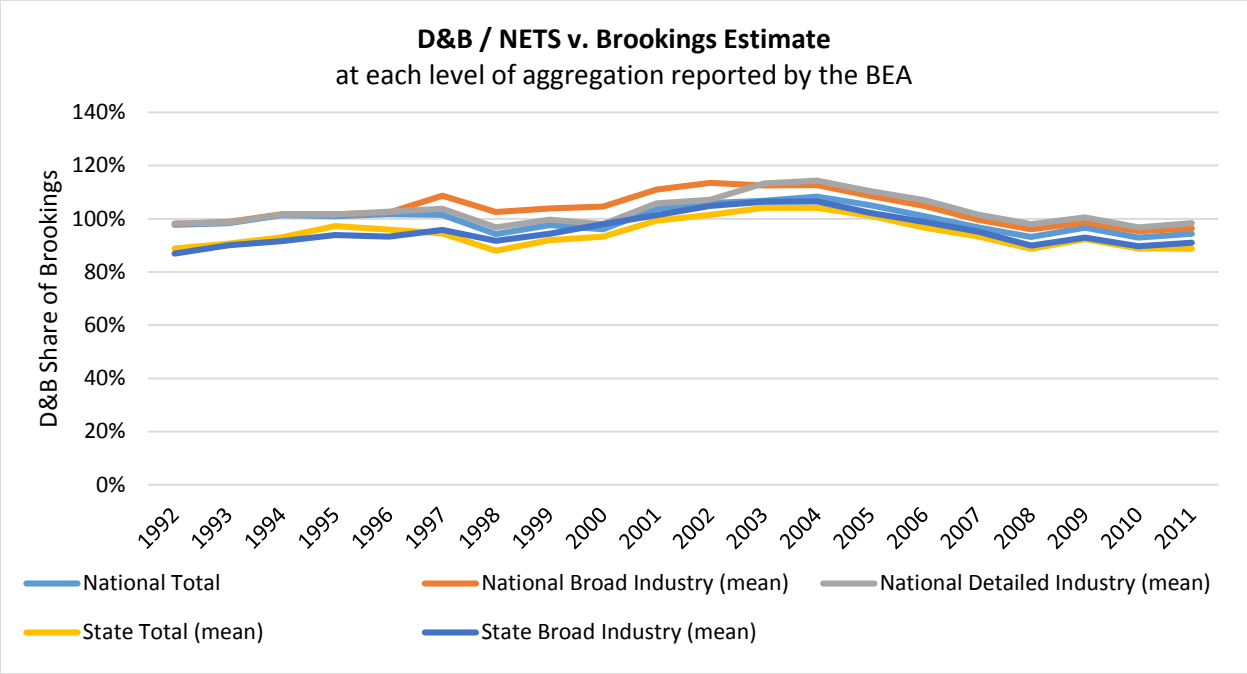
	Correlation with Brookings	Average Share of Brookings
BEA	93.4%	100.0%
D&B	93.4%	99.7%

This employment table includes the estimates reported by Brookings, the BEA, and NETS at the national level. In the aggregate, both the BEA and NETS series show a high correlation and tight fit to the Brookings estimates. Although the BEA data is a perfect match by design, the D&B data also exhibit a good fit, albeit with slightly lower average share of the Brookings employment level over time. Note, for the sake of comparability the bank estimates for the years preceding 2007 were excluded from each series due to their absence from the BEA data.

The next two charts compare BEA and NETS data against the Brookings estimates across the different levels of aggregation reported by the BEA. These levels represent the national totals, national broad-industry, national detailed-industry, state totals, and state broad-industry. The industry aggregates are in terms of the BEA's ISI-based industry classification, which is the most detailed industry grouping allowable across all three datasets.



As mentioned earlier, the further back in the past and the deeper the industry or geographic detail, the higher the expected deviation between the BEA and Brookings data. This mainly results from periodic revisions to the BEA's survey methodologies over the past two decades that make the BEA data not entirely comparable across time. For instance, deviations at the state level preceding 2002 reflect an accounting of employment in minority-owned in addition to majority-owned affiliates and deviations in detailed national industry data preceding 1997 reflects the difference between Brookings NAICS based estimates and the now defunct SIC data reported by the BEA during that period. Since the Brookings estimates were adjusted to account for these changes, these kinds of deviations are a natural and expected part of the methodology. What is more, the near perfect fit between the BEA and Brookings estimates after 2002 and after 1997 at the national level (the years when the data should be compatible), confirms the expected and desired result of the allocation method.



This next chart compares NETS to the Brookings estimates using the same industry aggregation groups as the previous chart. The central point here is to show the relatively consistent relationship between the Brookings and NETS data over time. Since NETS is a more methodologically consistent database, the fact that Brookings estimates holds such a stable relationship across time suggests that in the areas where Brookings was forced to deviate from the BEA trends and derive its own estimates it still adhered to NETS at more or less the same rate as it always has. Again, although this is an expected result, this chart confirms that the Brookings estimates are relatively close to the D&B estimates at all points in time and thus likely provide a conservative estimate of employment in foreign-owned establishments throughout the entire series at these specific levels of industry and geographic aggregation.

Relationship with Brookings Estimate

Levels of Aggregation	**Residual Sum of Squares**		**Coefficients**	
Dependent Variable	BEA	NETS	BEA	NETS
National Total	100.0%	87.3%	1.00	0.95
National Broad ISI*	91.9%	76.4%	1.22	0.86
National Detailed ISI*	53.6%	81.0%	1.28	0.86
State Total**	98.5%	75.5%	0.99	0.99
State Broad ISI**	63.2%	56.3%	1.03	0.85

All coefficients have p-values under 1%

**BEA regression covers period following the NAICS revision, 1997-2011*
***BEA regression covers period following the majority-owned revision, 2002-2011*

As a final diagnostic to summarize the discrepancies between the Brookings estimates and the BEA and NETS data, analysts calculated the residual sum of squares and coefficient estimates using a basic fixed effects panel regression. The results in the above table provide further detail confirming the same conclusions described in the previous robustness checks. As the level of industry detail increases at either the national or state level, the ability of the Brookings estimates to explain the variation within the BEA data declines, however its tightness to fit of the NETS holds relatively better, especially at the national level. As shown in the regression coefficients, the Brookings estimates tend to fall somewhere between the BEA and NETS data, with Brookings consistently falling below BEA, but above NETS. The average coefficient across the selected levels of industry aggregations was 1.10 for the BEA regression and 0.90 for NETS, a result indicating that the Brookings estimates tend to fall directly at the mid-point of both datasets when averaged across the various levels of industry and geography. In conclusion, these robustness checks demonstrate how the dual allocation methodology manages to produce estimates that fit within the constraints of the two best available subnational data resources covering foreign employment over time.

Other methodological notes:

Geographical definitions

This study estimates foreign employment for each of the 3,113 counties in the 50 states plus the District of Columbia, which allows the simultaneous calculation of values at the metropolitan level, at the state level, and for the metropolitan portions in each state in the case of metro areas that cross state borders. This report uses the BEA definition of a county.

This study assesses foreign employment across the largest 100 metropolitan areas in 2010, using metropolitan statistical areas (MSAs) as defined by the Office of Management and Budget (OMB) in 2009. There were 366 metropolitan areas in 2009 and this report focuses on the largest 100 metro areas by population: those with 500,000 or more residents in 2010. This report concentrates on this group to narrow the scope analysis, increase the meaning behind metro area to metro area comparisons, and because they collectively contain two-thirds of the nation's jobs and generate three-quarters of GDP.

Country of origin

The country of origin data, acquired directly from D&B, includes all available location information on foreign headquarters of NETS establishments for the last year (2011) in the data series. Since D&B does not archive historical data, this universe of foreign headquarters consists of establishments reporting as active in 2013, the time when the data was acquired. Using this technique, over 94 percent of NETS establishments in 2011 had a matching foreign headquarters. The remaining subset failed to turn up any available foreign headquarters records despite repeated searches. A number of factors contribute to missing records when data are directly extracted from D&B. These include suppressions from non-responses to survey inquiries, privacy requests from firms, or data quality concerns. Additionally, the missing

records could result from changes in the establishment's ownership structure that occurred between the last available NETS year and current D&B year. In any event, it is unlikely that the lack of coverage disproportionately affects one country more than another, so the relative composition of foreign ownership by country of origin should remain unaffected.

Readers may notice that this study reports sub-national foreign geographies at a level of aggregation called "city-region." Since there exists no standardized definition of global metropolitan areas, Brookings utilized administrative divisions (such as Canadian provinces, German states, or Japanese prefectures) to delineate each country's subnational regions, and then attached the names of the largest cities in that region to better orient readers. For example, the name "Toronto-Ottawa-Ontario, Canada" represents the province of Ontario, but details Toronto and Ottawa as the cities with the largest concentration of jobs in FOEs within that province. Country-specific deviations to the naming convention were adopted in some cases to improve readability or better approximate true economic regions. Future updates will explore the use of geocoding programs to assign longitudes and latitudes to the global parent addresses to delineate consistent boundaries that better approximate metropolitan areas.

Foreign share of domestic employment

When Brookings reports the foreign share of domestic employment for a given industry or place, otherwise known as "foreign intensity," the baseline employment value was derived from Moody's Analytics data. This measure of domestic employment only includes private sector employment. In rare cases when the Brookings foreign employment estimate exceeded the Moody's estimate, the domestic employment estimate was revised upward so the foreign share of domestic employment would not exceed 100 percent. This adjustment occurred after the data was aggregated and was only applied to individual four-digit industries at the metropolitan and sub-metropolitan level. As a result, the all-industry and sector totals were unaffected by the adjustment.

Advanced industries

The R&D and STEM-intensive industries defined as Advanced Industries include: Oil & gas extraction (2111); Electric power generation, transmission & distribution (2211); Basic chemical mfg. (3251); Resin, synthetic rubber, & artificial synthetic fibers & filaments mfg. (3252); Pesticide, fertilizer, & other agricultural chemical mfg. (3253); Pharmaceutical & medicine mfg. (3254); Other chemical product & preparation mfg. (3259); Iron & steel mills & ferroalloy mfg. (3311); Alumina & aluminum production & processing (3313); Agriculture, construction, & mining machinery mfg. (3331); Industrial machinery mfg. (3332); Commercial & service industry machinery mfg. (3333); Engine, turbine, & power transmission equipment mfg. (3336); Other general purpose machinery mfg. (3339); Computer & peripheral equipment mfg. (3341); Communications equipment mfg. (3342); Audio & video equipment mfg. (3343); Semiconductor & other electronic component mfg. (3344); Navigational, measuring, electromedical, & control instruments mfg. (3345); Electric lighting equipment mfg. (3351); Household appliance mfg. (3352); Electrical equipment mfg. (3353); Other electrical equipment & component mfg. (3359);

Motor vehicle mfg. (3361); Motor vehicle parts mfg. (3363); Aerospace product & parts mfg. (3364); Railroad rolling stock mfg. (3365); Ship & boat building (3366); Other transportation equipment mfg. (3369); Medical equipment & supplies mfg. (3391); Software publishers (5112); Other telecommunications (5179); Satellite telecommunications (5174); Other information services (5191); Computer systems design & related services (5415); Scientific research & development services (5417).

[1] While firms may not always report accurately, they have an incentive to do so because the information informs their credit score, which D&B sells to lenders. If a firm provides false information it could have an adverse effect on their ability to raise capital in the future

[2] To obtain employment estimates, the BEA mandates all foreign-owned U.S. affiliates pursuant to the International Investment and Trade in Services Survey Act to complete an annual survey of FDI. Affiliates that fail to comply face up to $25 thousand in civil penalties and potential prison time for willfully failing to complete the survey. Affiliates with total assets, sales, or net-income that fall below an absolute value of $40 million may exempt themselves from the longer annual survey, but are still required to file an exemption form containing pertinent details on its level of operations. Estimates of the exempt and non-sampled affiliates that existed prior to the current year were derived by extrapolating forward their data from earlier years, while estimates of new exempt or non-reporting affiliates were derived using data they reported in BEA's quarterly surveys of FDI.

[3] These included: Forest Nurseries (113210); Hunting (114210); Pipelines (486910, 486990, 486110, and 486210); Scenic and Sightseeing Tours (487990, 487210, and 487110); Internet Publishing (516110); Telecommunications Resellers (517310); ISP Services (518111 and 518112); Dentists (621210); Specialized Residential Care Facilities (6222, 6232, and 6239); Specialty Hospitals (622310); Boarding Houses (721310); Specialized Restaurants (722110, 722211, 722212, and 722213).

[4] These included: Postal Service (491110); Central Banks (5211); Non-Profits (813); Private Households (814); Public Administration (92); Non-Classified (99). Analysts also dropped all establishments whose company names included keywords associated with government, educational, or religious institutions.

[5] The most significant of these corrections included: Removing industries that were present only one year and that year was not the current or establishment start year, removing firms with geographic codes that failed to match standardized crosswalks, removing establishments never coded as foreign-owned, recoding an establishment as foreign-owned if its headquarters was foreign for another establishment during the same year, recoding an establishment as foreign-owned if it was coded as not foreign-owned between two foreign-owned years, resurrecting all firm with post-recession closures in 2010 and 2011 that were active in 2013, and dropping all 2009, 2010, 2011 openings of firms with fewer than 50 employees that were inactive in 2013. The decision to drop presently inactive greenfield establishments was to yield a more conservative estimate of the post-recession surge in greenfield employment.

[6] In cases when a new establishment had no reported data, then the median size of the parent firm's branches or the median size of firms in the establishments corresponding eight-digit SIC were utilized. For existing establishments with missing employment in between reported years, Walls and Associates used a straight line linear interpolation technique.

[7] For more information on dual allocation procedures, see the "Gross Domestic Product by State Estimation Methodology" developed by the BEA.