1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data

Andrew M. Isserman and James Westervelt

International Regional Science Review 2006 29: 311
DOI: 10.1177/0160017606290359

INTERNATIONAL REGIONAL SCIENCE REVIEW 29, 3: 311-335 (July 2006)

1.5 MILLION MISSING NUMBERS: OVERCOMING EMPLOYMENT SUPPRESSION IN *COUNTY BUSINESS PATTERNS* DATA

ANDREW M. ISSERMAN
Departments of Agricultural and Consumer Economics and Urban and Regional Planning, University of Illinois, Urbana, isserman@uiuc.edu

JAMES WESTERVELT
Engineer Research and Development Center, Army Corps of Engineers, Champaign, IL, james.d.westervelt@ERDC.usace.army.mil

Missing data frustrate research and limit our understanding of regional economies. County Business Patterns *annually provides employment data for all U.S. counties and states at the most detailed industrial level, but two out of every three employment statistics are missing. In rural areas, this percentage is higher still. To protect the rights of employers to confidentiality, the U.S. Census Bureau has not disclosed the number of employees in 1.5 million cases in the 2002 data. Instead, it offers a suppression flag that represents an employment range. This article presents a two-stage method for replacing all the flags with employment estimates. Taking advantage of the hierarchical nature of the data both by industry and geography, the first stage identifies the smallest possible range for each suppressed number. Ensuring that employment adds up correctly up and down the industrial and geographical hierarchies, the second stage iteratively adjusts all the estimates until millions of constraints are met. The procedure simultaneously considers all industries in all counties, states, and the nation to produce a complete data set, which is available to the research community on the Internet.*

Keywords: *employment data; county data; data confidentiality; suppression; estimation; regional analysis*

This article presents the current status of efforts that began four decades ago with the development of an antisuppression algorithm in 1976 as part of a research contract between the Army Corps of Engineers and the University of Illinois. The authors first collaborated on overcoming suppression in 1980, teaching students how to do so with calculators for small counties in a regional analysis course. We thank our former colleagues, Wayne Hamilton, Oleg Smirnov, David Sorenson, and Ron Webster, who have contributed in major ways over the decades to development of successive software, always more comprehensive, creative, and computationally intensive than the previous version. They are not responsible for any shortcomings of this version.

DOI: 10.1177/0160017606290359

Data hunger rages on the frontiers of regional research. New research methods and technologies and the intellectual directions of several disciplines create lusty appetites for more data. Economic geographers and urban and regional economists study location patterns and growth with new fervor and computing power (Glaeser et al. 1992; Henderson, Kuncoro, and Turner 1995; Harrison, Kelley, and Gant 1996; Scott 1996; Ellison and Glaeser 1997; Holmes and Stevens 2004; Duranton and Overman 2005). The search for industrial and occupational clusters and regional competitiveness has researchers scouring data in search of recipes to colocate industries and create competitive advantage (Feser and Bergman 2000; Isserman 2001; Feser 2003; Porter 2003; Markusen 2004; Feser, Sweeney, and Renski 2005; Fingleton, Igliori, and Moore 2005). Policy analysts pick through spatial variation in search of detectable policy outcomes and implications (Henderson 1996; Isserman and Rephann 1995; Holmes 1998; Reed and Rogers 2003; Kahn 2004; Lacombe 2004; Partridge and Rickman 2005). New techniques of spatial and regional analysis devour vast quantities of data, the more location-specific the better (Anselin 1995; Goodchild et al. 2000).

This article can satisfy part of that hunger because it creates and documents the best possible publicly available county employment data for regional and spatial analysis. Its starting point is *County Business Patterns (CBP)*, a comprehensive annual compilation from the U.S. Census Bureau. Suppression rules riddle it with holes like a moth-eaten sweater and limit its usefulness. The techniques described in this article darn those holes, all 1.5 million of them. The antisuppressant strategy has two stages. The first stage—reducing uncertainty—mines all available information to identify the narrowest range within which each missing number must fall. The second stage—creating employment estimates—replaces each missing number with an estimate that is within its range and consistent with all the other actual and estimated numbers. The result is an employment database with consistent numbers from nation to state to county as well as up and down the industrial hierarchy for more than a thousand industries at the most detailed industrial level in existence.

There have been other attempts to overcome suppression, but the published scholarly literature is sparse. Porter (2003) uses the midpoint of the Census-provided range; Orr and Buongiorno (1989) described a procedure for reducing that range; Clapp, Pollakowski, and Lynford (1992) and Glaeser et al. (1992) made employment estimates using midpoints of establishment size groups; and Holmes and Stevens (2004) did the same using average employment size. None of these approaches follows the two-stage strategy, utilizes all the information embedded within *CBP*, and creates a complete data set that is industrially and geographically consistent. Gerking et al. (2000) presented such a method; it is a direct antecedent of this article and traces its ancestry to Isserman (1976).[1] More typical of the literature are journal articles that use estimates without documenting how they overcame suppression (e.g., Anselin, Varga, and Acs 1997; Gerking and Isserman 1981; Isserman 1977, 1980, 2001; Markusen 1994, 1996) or use less detailed data from other sources because of the suppression problem (Hammond and Thompson 2004). A noteworthy exception is Ellison and Glaeser (1997), who devoted more

than three pages to discussing their procedure for overcoming state-level suppression in the 1987 Census of Manufactures. Recently, Porter (2004, 30) designated access to unsuppressed data at the county level as a specific area for future research.

This article was written to focus attention on antisuppression methods as a research area, document the methods used to create a *CBP* data set with estimates replacing all suppressed numbers, and accompany the release of the data via the Internet to interested users. We hope to save other researchers the time and effort needed to craft their own solutions to the *CBP* suppression problem, enable wider use of these data, and encourage others to document and offer their enhanced data sets in a similarly public manner.

County Business Patterns: Characteristics and Suppression

CBP provides the most industrially detailed employment data publicly available for all states and counties of the United States. It has been available annually from the U.S. Bureau of the Census since 1964 and for some earlier years starting in 1946.[2] The data since 1988 can be downloaded from www.census.gov, and recent years can be purchased on CD-ROM for $50 (U.S. Department of Commerce [U.S. DOC] 2005). The years 1977 to 1987, plus more recent years, can be downloaded from the Inter-University Consortium for Political and Social Research at the University of Michigan, www.icpsr.umich.edu.

CBP has numerous desirable attributes. It draws on administrative records of the Internal Revenue Service, the Social Security Administration, and the Bureau of Labor Statistics, giving them a higher degree of reliability than voluntary, unchecked responses to census questions. Since 1974, *CBP* has been on an "establishment" basis, the physical location of the economic activity. Thus, it is more useful for regional and spatial analysis than "reporting unit" statistics, the pre-1974 *CBP* basis, which reflects the location of a parent company or payroll unit that can include numerous establishments in several counties or states.[3]

CBP has used the North American Industrial Classification System (NAICS) since 1998. Like its predecessor the Standard Industrial Classification System (SIC), NAICS is hierarchical (Office of Management and Budget [OMB] 1987, 1998, 2003). The most detailed level of the hierarchy, referred to as 6-digit, includes, for example, NAICS code 311513—cheese manufacturing. Each 6-digit industry is part of a 5-digit industry, a 4-digit industry, and so on. NAICS 311513—cheese manufacturing is part of 5-digit 31151—dairy product manufacturing except frozen dairy products, and part of 4-digit 3115—dairy product manufacturing, 3-digit 311—food manufacturing, 2-digit 31—manufacturing, and total employment.[4] There are twenty-one 2-digit industries, eighty-three 3-digit industries, two hundred ninety 4-digit industries, six hundred fifty 5-digit industries, and one thousand eighty-two 6-digit industries in the 2002 data.

The 2002 county data have 2,189,660 hierarchically arranged records. Each record is an industry in a county, and each includes industry employment, payroll,

number of establishments, and the size distribution of establishments by number of people it employees. For example, the industry, colleges, universities, and professional schools in Hampshire County, Massachusetts, has 7,380 employees with a first-quarter payroll of $41,804,000 and an annual payroll of $173,357,000; there are six establishments, one of which has 5 to 9 employees, one 50 to 99, one 250 to 499, one 500 to 999, one 1,000 to 1,499, one 1,500 to 2,499, and one 2,500 to 4,999. Another record is the industry, general medical and surgical hospitals, in the District of Columbia with 29,791 employees, a first-quarter payroll of $340,047,000, an annual payroll of $1,395,565,000 and fifteen establishments, of which three have more than 5,000 employees. Other records describe very small local industries, such as the 24 employees and three establishments in taxi service in Alameda County, California. These three examples are not quite what they seem, as is explained later.

CBP does not include all employment. It covers private, nonfarm employment but omits agricultural production employees, most government employees, self-employed individuals, employees of private households, and railroad employees. The included government employees work in NAICS 4248—wholesale liquor establishments, 44531—retail liquor stores, 522120—federally chartered savings institutions, 522130—federally chartered credit unions, and 622—hospitals. The employment statistics count full- and part-time employees, including salaried officers and executives of corporations, on the payroll in the pay period including March 12 (even if they are on paid sick leave, holidays, and vacations). Proprietors and partners of unincorporated businesses are not included. The March 12 pay period means users must be cautious when analyzing industries or economies with seasonal employment.

Nationally, private nonfarm employment was 68 percent of all employment in 2002, with nonfarm proprietors accounting for 16 percent, government 14 percent, and farm employment and proprietors 2 percent.[5] Returning to three previous examples, missing in the college and university data for Hampshire County are the state government employees at the University of Massachusetts at Amherst, and missing in the taxi data for Alameda County are all the drivers who rent cabs as independent contractors or drive cabs they own, but included in the hospital data for the District of Columbia are the federal government employees at the Walter Reed Army Medical Center and the Veteran Affairs Medical Center because hospitals are one of the five types of federal employees reported within *CBP*. These subtleties demonstrate that users are well advised to know the data definitions and be alert to their implications.

The time series of *CBP* is one of its appealing features, particularly with data since 1988 readily available at www.census.gov. The Census Bureau notes "analyzing economic changes over time" as one of *CBP*'s uses. Caution is necessary, however. Periodically changes occur in the industrial classification system, such as changes in the NAICS codes that make the 2003 data different from the 1998 to 2002 data, adoption of the NAICS codes in 1998 instead of the 1987 SIC code that had been used since 1988, and revisions within the NAICS and SIC codes that generally occur every five years for the economic censuses. The Census

Bureau points out that the change to establishments from reporting units did not occur until 1974 and the definition of "active" establishments changed in 1983.

Even for periods without such official changes in definitions and categories, care must be taken not to misinterpret year-to-year changes in the data. Recall that "the entire establishment is classified on the basis of its major activity and all data are included in that classification" (U.S. DOC 2005). Thus, an establishment can change classifications when its major activity changes. Simply making more men's shirts and fewer women's blouses one year can trigger a move from NAICS 315232—women's and girls' cut & sew blouse to NAICS 315223—men's and boys' cut & sew shirt. A researcher using the data to identify plant closings in the apparel industry might incorrectly identify the no-longer-existing blouse establishment as a plant closing and the shirt factory as a new plant.

Yet changes over time, the incomplete employment coverage, and seasonal employment are mere stumbling points compared to one staggering problem with the use of *CBP:* "In accordance with U.S. Code, Title 13, Section 9, no data are published that would disclose the operations of an individual employer" (U.S. DOC 2005). In 2002, employment and payroll was suppressed for two-thirds of the 2.19 million records. Thus, no payroll information and limited employment information exist for 1,461,702 county-industry records.

The procedures described in the next two sections mine available information within *CBP* to estimate employment in those almost 1.5 million records. There are rich veins to mine: "The number of establishments in an industry classification and the distribution of these establishments by employment-size class are not considered to be disclosures, so this information may be released even though other information is withheld from publication" (U.S. DOC 2005). Hence, all records contain the number of establishments and their distribution by size class. Furthermore, flags within the suppressed records indicate an employment range within which actual employment lies. Table 1 shows the flags, their corresponding minimum and maximum employment, and the employment size classes for the establishment counts.

The data's hierarchical nature, both industrial and geographical, contains additional useful information. All 6-digit industries must add up to their 5-digit classification, all 5-digit to their 4 digit classification, and so on. This property is illustrated in Table 2 for food manufacturing, one of the twenty-one 2-digit categories of manufacturing. In addition, employment in an industry summed across all counties in a state must equal the state's employment in that industry; employment summed across states and the District of Columbia must equal national employment; and all employment numbers within each state and the nation must sum up correctly within the NAICS hierarchy, as they must within each county.

Reducing the Uncertainty, Narrowing the Ranges

When opening *CBP*, a user is immediately confronted by an alphabetical blizzard where numbers are expected. There are 727,957 records with numbers and

TABLE 1. Minimum and Maximum Employment Corresponding to Each Suppression Flag for Data Withheld to Avoid Disclosure and the Employment Size Classes for the Establishment Counts

	Employment		*Number of Establishments*	
Flag	*Minimum*	*Maximum*	*Symbol*	*No. Employees*
A	0	19	N1_4	1 to 4
B	20	99	N5_9	5 to 9
C	100	249	N10_19	10 to 19
E	250	499	N20_49	20 to 49
F	500	999	N50_99	50 to 99
G	1,000	2,499	N100_249	100 to 249
H	2,500	4,999	N250_499	250 to 499
I	5,000	9,999	N500_999	500 to 999
J	10,000	24,999	N1000	1,000 or more
K	25,000	49,999	N1000_1	1,000 to 1,499
L	50,000	99,999	N1000_2	1,500 to 2,499
M	100,000		N1000_3	2,500 to 4,999
			N1000_4	5,000 or more

TABLE 2. The Hierarchical Properties of the North American Industrial Classification System (NAICS) Industrial Classifications: An Illustration, 2002

NAICS	*U.S. Employment*
3-digit industry	
311 Food manufacturing	1,443,766
4-digit industries of food manufacturing	
3111 Animal food	47,012
3112 Grain & oilseed milling	54,531
3113 Sugar & confectionery products	81,206
3114 Fruit & vegetable preserving & specialty food	158,905
3115 Dairy products	129,274
3116 Animal slaughtering & processing	495,730
3117 Seafood products preparation & packaging	38,663
3118 Bakeries & tortilla	288,577
3119 Other food	149,868
Sum	1,443,766
5-digit industries of dairy products	
31151 Dairy products (except frozen)	110,169
31152 Ice cream & frozen dessert	19,105
Sum	129,274
6-digit industries of dairy products	
311511 Fluid milk	56,617
311512 Creamery butter	1,604
311513 Cheese	37,636
311514 Dry, condensed, evaporated dairy products	14,312
311520 Ice cream & frozen dessert	19,105
Sum	129,274

1,461,702 with letters. The letters represent intervals as small as 19 in the case of A and as large as 24,999 in the case of K (L and M do not exist in the 2002 county, state, or national data).

This section discusses data mining techniques to reduce the intervals. The next section discusses the estimation methods for creating a consistent set of employment estimates. The end result is 1.46 million estimated numbers, each within its reduced interval, all of which sum correctly to meet industry totals from county to state to nation and up and down the NAICS hierarchy within county, state, and nation.

Suppression increases with industrial detail. There were 112.4 million private nonfarm jobs in 2002; only 77,331 were suppressed on the 1-digit level (total jobs), but 29.8 million, or one out of four, were suppressed on the 6-digit level (Table 3). Whereas total employment was suppressed for only eighteen counties, by the 3-digit level the number of records with suppressed employment almost equaled the number with actual employment. At the 6-digit level, there are almost three records with suppressed employment for each one with actual employment.

Fewer people are employed when there is suppression. For example, there are 1,095 jobs per record on the 3-digit level without suppression and 131 with suppression. The more detailed the industrial classification, the smaller the size gap between suppressed and actual employment.

The bounds of the flag codes (Table 1) are the starting point in narrowing the ranges before estimating suppressed employment. The flag code K on NAICS 336411—aircraft manufacturing in Snohomish County, Washington, means the industry there has a minimum of 25,000 employees and a maximum of 49,999. The sums of the minima and maxima of all 592,347 records on the 6-digit level with suppressed employment yields 17 to 53 million jobs (Table 3). This range, 36 million jobs or 32 percent of national private nonfarm employment, serves as the measure of initial 6-digit uncertainty. Digit level and uncertainty have a monotonic relationship; uncertainty is 10.5 million at the 3-digit level and 1.6 million, or 1.4 percent of employment, at the 2-digit level.

The data mining techniques to be described later in this section reduce the uncertainty by roughly three-quarters for all digit levels but one. The uncertainty regarding total employment in those eighteen suppressed counties is reduced from 65,122 to 1,922 jobs, or by 97 percent. For Snohomish aircraft manufacturing, uncertainty dwindles from 24,999 to 385 jobs.

Data Mining Procedures and Results

Language from family relationships—parents, children, and siblings—is helpful in describing the procedures. Referring again to Table 2, food manufacturing is the parent of nine 3-digit industries. They are its children, tagged genetically or numerically by their parent's entire NAICS code at the beginning of their own. To each other, the nine 3-digit industries are industrial siblings. One sibling, dairy products, has two children, of which one has four children of its own and the other has but one, ice cream and frozen dessert.[6]

TABLE 3. Summary of Actual and Suppressed Data by North American Industrial Classification System (NAICS) Digit Level, Counties, 2002

Digit	*Type*	*Records*	*Total Employment*	*Average Employment*	*Sum of Minima*	*Sum of Maxima*	*Uncertainty*	*Minimum Uncertainty*	*Percentage Reduced*
1	Actual	3,171	112,323,325	35,422					
1	Suppressed	18	77,331	4,296	49,390	114,512	65,122	1,922	97
2	Actual	43,626	110,919,737	2,543					
2	Suppressed	16,856	1,480,919	88	927,540	2,521,154	1,593,614	423,493	73
3	Actual	91,976	100,723,632	1,095					
3	Suppressed	88,968	11,677,024	131	6,879,640	17,397,342	10,517,702	2,305,311	78
4	Actual	162,013	92,311,292	570					
4	Suppressed	275,104	20,089,364	73	11,792,810	33,092,746	21,299,936	5,123,088	76
5	Actual	210,765	87,032,590	413					
5	Suppressed	488,409	25,368,066	52	14,577,370	44,620,271	30,042,901	7,707,651	74
6	Actual	216,406	82,566,165	382					
6	Suppressed	592,347	29,834,491	50	17,198,790	52,990,283	35,791,493	9,294,242	74

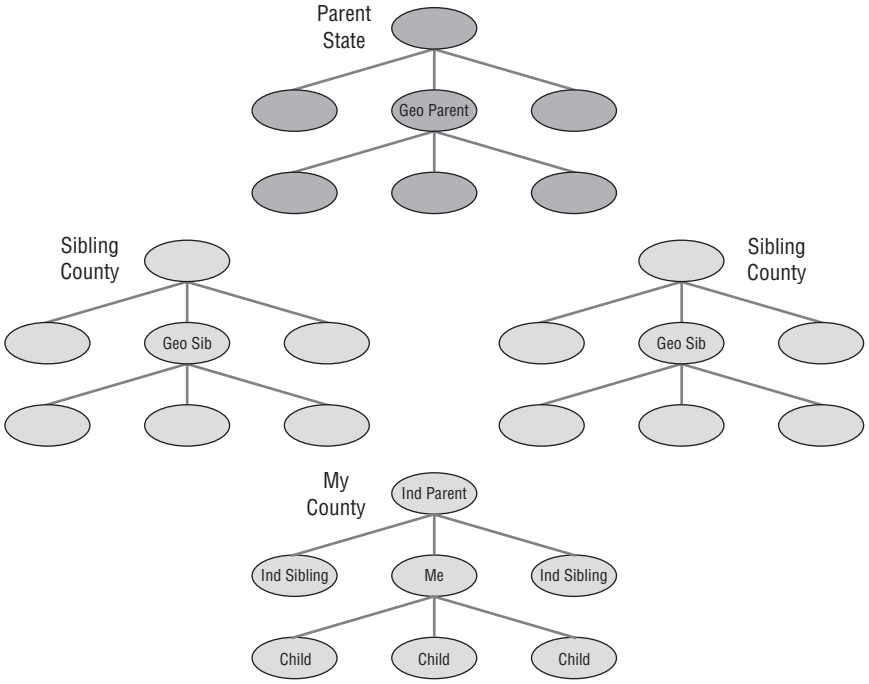

FIGURE 1. Relationships of a County Industry to Other County and State Industries
Note: Geo = geographical; Ind = industrial.

Places bring another set of relations into the family. Each record is an industry in a place. All the county records of a particular industry, for example, ice cream and frozen dessert, are geographical siblings if they are within the same state. Employment in the industry summed across counties must equal the state employment in that industry. States are the parents of counties, and the nation is the parent of states.

Each county industry record has unique parentage—an industry at the next level of aggregation within the county and its own industry at the state level. Figure 1 shows part of a family tree; it illustrates the situation from the perspective of one suppressed record, which is labeled "Me" in "My County." The genealogy is not traditional; Me's industrial parent is the parent of all Me's industrial siblings, and Me's geographical parent is the parent of all Me's geographical siblings.

Establishment size categories. The establishment size categories are finer than the flag categories (Table 1) and make possible the second set of bounds. Multiplying the minimum employment of each size category by the number of establishments in that category (except the redundant N1000) and then summing

over the categories yields an estimate of the industry minimum. Similarly, multiplying the maxima of the categories by the number of establishments leads to an industry maximum (except where there are establishments in the open-ended N1000_4, and no maximum can be calculated with this method).

To illustrate with a simple example, assume that an industry in a county has flag code G, which means 1,000 to 2,499 employees, and five establishments, all in the 100 to 249 size category. The employment bounds, according to the establishment size data, are 500 to 1,245. Together, the two systems yield the narrowest bounds yet; the actual number must be between 1,000 and 1,245. In the language used here, uncertainty has been reduced from 1,499 with the flag and 745 with the establishment size data to only 245 with the two methods combined.

Industry hierarchy. The hierarchical property of the NAICS code produces the third and fourth sets of bounds. Information about the industrial parent and siblings can be mined to create the third set of bounds. The minimum employment of a suppressed industry is (1) its parent industry's minimum (2) minus the summed maximum of all its siblings. Likewise, its maximum is (1) its parent's maximum (2) minus the summed minimum of all its siblings. These calculations are made with the largest minimum and the smallest maximum identified in all previous steps. For industries that are not suppressed, the minimum and maximum are the same (and equal to actual employment).

The industry hierarchy can be mined in the opposite direction to yield the fourth set of bounds. The minimum employment of a suppressed industry is the sum of its children's minima, and its maximum employment is the sum of its children's maxima. Again, these calculations use the maxmin and minmax identified thus far.

Geographical hierarchy. The geographical hierarchy produces additional bounds because for each industry employment in the counties must sum to employment in the state (and employment in all the states must sum to national employment). Mining this information produces minimum employment of a suppressed industry in a county equal to (1) its state's minimum employment (2) minus the summed maxima of its geographical siblings, that is, employment in all the other counties in that state. Likewise, the maximum county employment of the suppressed industry is (1) the state maximum (2) minus the summed minimum of all the other counties. The geographical hierarchy also can be mined in reverse fashion for states and the nation. For example, state employment in an industry can be no larger than the county maxima summed and no smaller than the county minima summed.

Iterations. The bounds from the flag codes and establishment distributions need only be calculated once, but the calculations for the industrial and geographical bounds are interdependent and must be repeated. Each time the bound of a suppressed number is narrowed, it creates the potential to produce narrower bounds for other suppressed numbers. The calculations always use the minmax and maxmin derived thus far, but additional iterations can produce still larger

minima and still smaller maxima. With reductions in uncertainty from one step providing revised information for subsequent steps, an iterative approach is necessary to mine all that can be learned. The entire set of industrial and geographical calculations is repeated for nation, states, and counties until no bound can be tightened any further. The end result is the greatest possible reduction in uncertainty, the narrowest possible range for the suppressed information, which can be derived from the known information.

Results. The narrowest bounds are facts, gleaned from mining the information contained in the data. Cumulatively, the mining process removed over three-quarters of the uncertainty regarding suppressed county, state, and national employment over all industries. Starting with the flag codes, the uncertainty was more than 110 million (the sum of the maxima minus the sum of the minima), but by the end of the process 84 million had been removed. Focusing only on the 6-digit level to avoid double counting, the uncertainty declined from 42 million to 10 million.

The steps vary greatly in what they contribute. Augmenting the flag code bounds with those from the establishment size data removed 40% of the uncertainty (Table 4). Adding the adjustments for parents and children, up and down the industrial hierarchy, removed an additional 34%. Mining the geographical hierarchy narrowed the bounds only another 2%.

Most of the work narrowing the ranges was complete by the first iteration, which reduced the uncertainty by 82,834,956 jobs. Note that the algorithm iterates within a step before moving to the next, for example, adjusting county bounds up and down the NAICS hierarchy repeatedly before turning to the geographical calculations (and then returning to the industrial calculations on the next iteration). The second iteration narrowed the ranges by 1,074,264, the third 41,248, the fourth 1,852, the fifth 23, and the sixth 0. In short, 98.67 percent was accomplished on the first round, 1.28 percent on the second, and 0.05 percent on the third.

Industry Size and Urban-Rural Character

Most suppression occurs in industries with relatively few employees in a county. Looking at the extremes by flag code and focusing on the 6-digit level, there are only 5 suppressed records in all counties with the K flag, meaning an actual employment between 25,000 and 49,999, and 50 suppressed records with a J flag, meaning 10,000 to 24,999, but more than 400,000 records with flag A (0 to 19) and more than 130,000 with flag B (20-99). Cumulatively, the initial uncertainty about county industries with fewer than 100 employees is greater than 18 million, far more than the 900,000 for county industries with 10,000 or more employees (see Table 5). Thus, the greatest effect of suppression is to protect the confidentiality of small businesses.

The dominant role of the small industries becomes even more evident after applying the entire data mining uncertainty process. The cumulative reduced or remaining uncertainty for all suppressed industries with 5,000 or more employees

TABLE 4. Effects of Each Step in Reducing Total Uncertainty, National, State, and County Data, 2002

	All Records	*1-Digit*	*2-Digit*	*3-Digit*	*4-Digit*	*5-Digit*	*6-Digit*
Uncertainty measured by flag code	110,650,008	65,122	1,617,035	10,968,529	23,047,087	33,332,392	41,619,843
Reduction in uncertainty by source:							
Establishment size data	44,318,993	18,470	606,111	3,795,065	8,914,487	13,901,015	17,083,845
Percentage reduction in uncertainty	40	28	37	35	39	42	41
Industrial hierarchy	37,478,220	30,670	547,998	4,753,215	8,519,925	10,563,895	13,062,517
Percentage reduction in uncertainty	34	47	34	43	37	32	31
Geographical hierarchy	2,155,130	14,060	36,762	79,911	288,731	629,549	1,106,117
Percentage reduction in uncertainty	2	22	2	1	1	2	3
Total reduction in uncertainty	83,952,344	63,201	1,190,872	8,628,192	17,723,144	25,094,460	31,252,480
Percentage reduction in uncertainty	76	97	74	79	77	75	75

TABLE 5. Reductions in Uncertainty by Employment Size Class, 6-Digit Level, All Counties, 2002

	Records	*Uncertainty*	*Reduced Uncertainty*	*% Reduction*	*Flag Code Uncertainty*	*Average Reduced Uncertainty*
R	216,406	0	0		0	0
K	5	124,995	1,498	99%	24,999	300
J	50	749,950	29,854	96%	14,999	597
I	169	844,831	66,452	92%	4,999	393
H	592	1,479,408	172,252	88%	2,499	291
G	2,562	3,840,438	462,431	88%	1,499	180
F	5,627	2,807,873	725,520	74%	499	129
E	12,355	3,076,395	1,060,296	66%	249	86
C	31,381	4,675,769	1,722,179	63%	149	55
B	132,322	10,453,438	2,790,952	73%	79	21
A	407,284	7,738,396	2,262,808	71%	19	6

(flag I or higher) is fewer than 100,000 jobs, whereas industries with fewer than 100 employees have a cumulative uncertainty of 5 million jobs. The data mining system has reduced uncertainty for the larger categories by 92 to 99 percent and for the smaller categories by 71 to 73 percent. The peskiest categories in terms of resisting reductions in uncertainty are C (100-249) and E (250-499). The five records whose initial uncertainty was 24,999 (the K flag) have impressive reductions, with the reduced uncertainty averaging only 300 jobs and amounting to at most 727 jobs. More detail is available in Table 5.

Examining the sources of reduced uncertainty for each size category helps explain the empirical findings. The reductions in categories A and B result primarily from substituting the establishment class minima and maxima (see Table 6); this finding is not surprising since there are three establishment size categories within flag A and two within B and the small industries are often aggregates of very small firms. The industrial parent-child calculations are particularly important in the larger flag categories; a disclosed parent category combined with some disclosed siblings allows substantial reductions in uncertainty. Categories C and E, those pesky laggards, and F also gain primarily through industrial parent-child adjustments in part because the flag ranges are identical to establishment size categories.

The relative importance of parent-children adjustments and the greater reductions in uncertainty for larger industries suggests a relationship between the size of an economy and the suppression problem. Indeed, the data show that suppression is a much bigger problem for rural economies than urban ones. In the most urban counties, only 13 percent of employment is suppressed on the 6-digit level, but in the most rural counties, 71 percent is suppressed. The percentage increases monotonically among the four categories of the rural-urban density classification (Isserman 2005), and metropolitan counties have less suppression than nonmetropolitan ones (Table 7). The reduction in uncertainty through narrowing the bounds is also somewhat less for rural, mixed rural, and nonmetropolitan counties.

TABLE 6. Reduction in Uncertainty by Source and Employment Size Class, All Records, National, State, and County (in percentages)

Step	*A*	*B*	*C*	*E*	*F*	*G*	*H*	*I*	*J*	*K*
Establishment size data	65.9	55.7	20.8	14.5	17.2	47.5	25.2	24.1	18.0	11.9
Industrial hierarchy	3.9	16.6	40.2	48.3	53.2	38.1	61.0	64.4	77.5	85.0
Geographical hierarchy	0.1	0.6	2.5	3.2	3.8	2.9	3.8	4.7	1.6	1.9
Total reduction in uncertainty	69.9	73.0	63.5	65.9	74.3	88.5	89.9	93.3	97.1	98.7

TABLE 7. Suppression by Type of County, 6-Digit Level, 2002

County Type	*Total Employment*	*Suppressed*	*Suppressed as Percentage*	*Uncertainty*	*Reduced Uncertainty*	*Percentage Reduction*
Rural	6,519,789	4,625,975	71	7,987,392	2,247,668	72
Mixed Rural	29,117,811	12,581,498	43	15,799,157	4,275,239	73
Mixed Urban	16,239,853	4,296,717	26	4,477,866	1,129,016	75
Urban	59,203,232	7,965,084	13	7,192,253	1,598,927	78
Nonmetro	14,159,347	8,802,430	62	13,326,990	3,754,460	72
Metro	96,921,338	20,666,844	21	22,129,678	5,496,390	75
All	111,080,685	29,469,274	27	35,456,668	9,250,850	74

Making Consistent Employment Estimates

Narrowing the ranges is just preliminary to the more demanding computational challenge: estimating the 1.46 million missing numbers so that each will be in its range and that all summations up and down the industrial and geographical hierarchies are correct. Referring again to the family tree in Figure 1, Me's employment and that of its two industrial siblings in My County must sum to the employment of their industry parent; Me's employment must equal the employment of its three children; and Me's employment and that of its geographical siblings in other counties in that state must sum to their industry's state employment, that is, their geographical parent.

The estimation procedure begins by assigning an initial employment estimate to each suppressed number equal to the midpoint of its narrowest possible range. Experiments using the minimum or maximum of the range as the initial estimate revealed that the starting point is unimportant for the result or the speed of computational convergence.

The initial estimates are then iteratively adjusted to increase the agreement of the employment estimates for industrially or geographically related industries. Within each iteration, adjustments are made to ensure that five sets of conditions are met:

1. national employment by NAICS code and associated state employments match,
2. state employment by NAICS code and associated county employments match,
3. all parent-children combinations at the national level match,
4. all parent-children combinations at the state level match, and
5. all parent-children combinations at the county level match.

In all cases, adjustments are made based on the difference between the elements (for example, the sum of the children and their parent), the amount of adjustment possible in the elements (an estimate further from its bound will change more than one closer to its bound), and the proportion of the change permitted in the iteration. A simple example of parent-children matches within a county will be helpful. Assume that a parent's employment estimate is 100 more than the sum of its children. Various adjustments could bring them into balance. Assume further that the parent estimate is 60 above its minimum, so the most it could contribute is a reduction of 60, and that its only two children are within 36 and 24 of their maxima, respectively. The parent has 60/120 of the total adjustment space, so it is reduced by 50, half the adjustment needed. The child with 36/120 of the adjustment space is increased by 30, and the child with 24/120 by 20. The three adjustments make the parent-children numbers match; numbers that can change more are adjusted more. The story cannot end there, however, because the parent and its own siblings also have to match their own industrial parent. Based on the same type of calculation, assume that the industry that had to decrease by 50 to match its children also has to increase by 15 to match its own parent. The two contradictory adjustments are both made, bringing the system closer to balance but requiring further adjustments and iterations.

The entire system moves toward balance with the iterations following a gradient-descent algorithm. The proportion of change permitted in an iteration, a simulated annealing term, declines from 100% in the first iteration to 50 percent in the final (1,000th) iteration. Hence, using the initial example, the parent would have been decreased by 50 in the first iteration but only by 25 if that imbalance had occurred in the final iteration. This process allows relative rapid adjustments at the outset but helps the process settle on a solution at the end. All estimates are adjusted using floating-point numbers. Experiments with full integer variants of the algorithm continue.

To meet conditions 1 and 2, the geographical constraints, each NAICS code is considered separately. The order of processing 2-digit, 3-digit, and so on is inconsequential, so the algorithm follows the NAICS order in the original *CBP* data file, that is, NAICS 11, 113, 1131, 11311, 113110, 1132, 11321, 113210, 1133, 11331, 113310, 114, 1141, 11411, 114111, 114112, 114119, 1142, and so on. Thus, the algorithm processes a parent, its first child, the child's first child, and that entire branch of the family tree before returning to the parent's second child. It follows a geographical hierarchy by adjusting the state numbers to match the national number and county numbers to match their state numbers. These adjustments are made even when the geographical parent number is an estimate, as was described in the numerical example. When the difference between the industry total at the state level is compared to the sum of the associated counties, the amount that the state estimate can shift without violating its bound is combined with the amount its respective counties can shift. The amount each estimate shifts is the ratio of the shift needed and the shift that can be accommodated. When a number is not an estimate, its minimum and maximum are identical to the number, so it can accommodate no change and remains unchanged.

For conditions 3, 4, and 5, the entire parent-child structure is looked at within each geographical entity and adjustments 2-digit parents and their 3-digit children are combined with adjustments to those 3-digit children and their own 4-digit children, and so on, before any adjustments are made. Thus, the –50 and the +15 in the numerical example are both scheduled and occur at the same time. All adjustments are based on the disparity between parents and children and the amount of adjustment possible within each; as in the numerical example, estimates that can accommodate more adjustment receive a greater share of the needed adjustments. After the complete industrial hierarchy in all geographical entities, nation, state, and county hierarchy is evaluated, all the scheduled adjustments are made. In short, the algorithm simultaneously adjusts the children to match their industrial parent, although the parent itself might be changing in the same round to make it and its siblings match their parent.

The process is repeated for one thousand iterations, at which point the solution has adequately stabilized. Table 8 shows the initial rapid drop in the imbalances and the slowly changing adjustment through one thousand iterations. The rows represent summation checks defined in the second column and related directly to the five conditions. The initial discrepancies, for example, the 1.65-million-job difference

TABLE 8. Cumulative Employment Adjustment by Iteration, Selected Results

			Imbalance after Iteration					
Related Condition	*Sum of Differences Between*	*Initial Imbalance*	*1*	*250*	*500*	*750*	*999*	*1,000*
1	National employment and employment in states for all industries	106,234	56,752.64	96.71	4.48	1.98	1.52	1.53
2	State employment and employment in state's counties for all industries	1,650,956	747,270.56	309.40	27.88	16.72	12.22	11.72
3&4	Parents and children in the nation and all states for all industries	46,089	4.93	8.45	1.11	.64	.70	.68
5	Parents and children in all counties for all industries	641,942	433.00	38.93	9.48	7.21	6.76	7.04

between the sum of counties and the state totals, reveal the importance of the consistency checks and the contribution of the balancing algorithm. Interpreted in another way, they also show that mining the whole system to reduce the ranges and selecting the midpoints of each range provide a shortcut approximation to a balanced system—within 1.8 million jobs of balancing geographically and 700,000 jobs of balancing industrially when 30 million jobs were suppressed on the 6-digit level alone.

The final result is an internally consistent set of floating-point employment numbers, with estimates replacing the 1,461,702 suppressed county numbers as well as the 32,611 suppressed state and 8 suppressed national numbers. Those employment numbers are only estimates, although the reduced bounds of each are fact and all the employment estimates are consistent with all the facts. They are not the only set of numbers that are consistent with all the facts.

How accurate are the estimates? There is a good reason to avoid this reasonable and obvious question. Imagine the logical consequences of documenting that the algorithm reproduces the suppressed employment data perfectly, almost perfectly, or very accurately. Faced with such evidence, the Census Bureau quite likely would have no choice but to suppress even more information. Nobody wins at that game. Accuracy could be measured by gaining access to the confidential data at a Census Bureau site and comparing the actual and estimated numbers, but it is a fool's game. Instead of adding pressure to suppress further, a better stopping point is to accept an unknown degree of accuracy within the narrowest bound for each number, leaving it up to the user whether to seek additional information regarding particular records.

Simulations can be designed, for example, by artificially injecting further suppression into the data and comparing the estimates for these "suppressed" numbers to their known actual numbers; pairing suppressed records with similar unsuppressed ones; or generating synthetic data, mimicking the suppression rules and then measuring accuracy. The first two research designs are flawed by the simple fact that if the other records were comparable to the suppressed ones, they themselves would have been suppressed, and the third must cope with the fact that the Census Bureau suppression algorithm is unknown and, therefore, difficult to mimic.[7]

Comparing differences between estimated and actual numbers confirms that the suppressed industries tend to have fewer employees and are not comparable to the disclosed ones. Within each industry employment size class, as delineated by the flags, average employment size is smaller for the suppressed county records. For example, within category A (0 to 19 employees), the mean for actual records is 11 employees and for suppressed records is 6 employees (Table 9). By category G (1,000 to 2,499 employees), the actual mean is 1,539 employees and the estimated suppressed mean is 1,476. An F-test, to determine whether the ratio of the variances of the actual and suppressed distributions is equal to one, shows that we can reject that hypothesis until flag H or for all size categories below 2,500 employees. Applying the appropriate t-test (equal or unequal variance), we can reject the hypothesis that the mean is the same for both distributions for all size classes until I, that is, all categories with fewer than 5,000 employees.

TABLE 9. Differences in Employment Means and Standard Deviations, Actual and Suppressed

Flag	*Suppressed Records*	*Actual Records*	*Mean Estimate of Suppressed*	SD *of S*	*Mean of Actual*	SD *of Actual*	F	*Prob.* F	t	*Prob.* t	*MAME*
A	407,284	37,277	6	5	11	5	0.97	0.00	–194.22	0.00	3
B	132,322	81,573	43	19	52	23	0.74	0.00	–92.94	0.00	13
C	31,381	44,019	155	38	159	42	0.81	0.00	–13.83	0.00	34
E	12,355	23,638	346	64	352	71	0.82	0.00	–8.53	0.00	52
F	5,627	14,588	683	131	699	141	0.86	0.00	–7.64	0.00	78
G	2,562	9,773	1,476	393	1,539	409	0.93	0.02	–7.08	0.00	107
H	592	3,265	3,396	710	3,474	695	1.04	0.49	–2.49	0.01	169
I	169	1,492	6,698	1,288	6,856	1,378	0.87	0.26	–1.42	0.15	225
J	50	638	14,419	3,439	14,514	3,755	0.84	0.45	–0.17	0.86	333
K	5	105	27,353	1,381	33,111	6,274	0.05	0.01	–6.62	0.00	166

There is wisdom in suppressing the results of accuracy tests to forestall any additional, more severe suppression, but a presumably harmless summary measure gives some sense of the properties of the estimates. The maximum error of an estimate is the absolute value of the difference between the estimate and its farthest bound; it is the most inaccurate an estimate can possibly be. The mean absolute maximum error (MAME), shown in Table 9, is small when compared to its flag code interval. For example, it is 3 jobs for all the suppressed records in flag category A (0-19 jobs), 13 jobs for B (20-99), 34 jobs for C (100-249), and, near the other extreme, 169 jobs for H (2,500-4,999), 225 jobs for I (5,000-9,999), and 333 jobs for J (10,000-24,999). The actual mean errors are smaller; how much so perhaps best remains unknown.

Completing an Employment Database

CBP with estimates for the 1.5 million suppressed numbers is valuable data for many research questions about the private nonfarm economy. It provides far more industrial detail than another valuable data source from the U.S. Department of Commerce, the Regional Economic Information System (REIS) of the Bureau of Economic Analysis (BEA). REIS, however, is more comprehensive, covering all employment including farming and government as well as proprietors within each industry. BEA's main source for wage and salary employment is the ES-202 series of the Bureau of Labor Statistics, which provides monthly employment and quarterly wages for each state and county in NAICS 5-digit detail. BEA estimates self-employment "based mainly on data tabulated from individual and partnership Federal individual income tax returns" (U.S. DOC 2006). The latter are only at the 2-digit level, so REIS, which combines wage and salary employment and self-employment by industry, only provides the 2-digit combined employment.

Even though *REIS* stops at the 2-digit level for employment, data are suppressed in the *REIS* system as well "to preclude the disclosure of information about individual employers." There are no flags indicating the range of employment and no establishment data to create bounds. There is only a symbol (L) indicating less than ten jobs and a symbol (D) indicating "not shown to avoid disclosure of confidential information, but the estimates for this item are included in the totals." Using methods like those described here, one also can create a set of consistent employment estimates to fill in the REIS data and match all totals, but that process will be not discussed here.

To gain more complete coverage of the economy without sacrificing industrial detail, farming and government employment from REIS can be added to *CBP*. The match is not perfect. Although *CBP* is only the pay period including March 12, the REIS employment is an annual average. REIS uses a slightly different geography than *CBP*, for example, combining the independent cities of Virginia with their surrounding counties, whereas the Census Bureau keeps them separate. Most important, adding the five available employment categories from REIS—farming, federal civilian, military, state government, and local government—completes all wage and

salary employment but still excludes proprietors. They can either be ignored or added as farm and nonfarm proprietors but otherwise undifferentiated as to industry. In either case, REIS provides a useful way to augment *CBP* data to incorporate farming and government.

CONCLUSION

The method for estimating suppressed employment numbers has two stages: identifying the narrowest possible range for every suppressed number and then making internally consistent estimates of employment. The first stage is totally factual; it is data mining to exploit known facts to identify the smallest actual range for each of the suppressed 1.5 million numbers. The second step produces estimates. The resulting numbers are internally consistent, but they are not the only possible internally consistent set of 1.5 million numbers that satisfy all the conditions the numbers must satisfy, including each being within its true range. Another set that meets all these constraints is the actual numbers. We are considering simulation experiments to increase our understanding of the range of possible solutions that meet all the conditions.

The best we can do is to make sure we have exploited all available information in identifying the narrowest ranges and that we have specified a reasonable procedure for creating the consistent estimates. There is one fragment of information whose use we have not explored fully. The *CBP* file for national employment includes the number of employees in each employment size grouping as well as the number of establishments (but subject to considerable suppression).

We are working on mathematical programming alternatives to the present algorithm for making employment estimates. The bounds for each number are constraints, as are the parent-child industrial and geographical consistency conditions. The arbitrary part is the choice of objective function. We are exploring minimization of the distances from a starting point, be that the midpoint of each suppressed number's narrowest range, the weighted sum found from using establishment counts and the national average employment by establishment size group and industry, or the average employment from selected unsuppressed records. Having tested the implications of starting the present algorithm with different estimates, we do not think the starting point is a major decision. Of more concern is that Ellison and Glaeser (1997, 922) have deemed a similar but far smaller optimization problem "intractable."

Development of algorithms to overcome suppression is perhaps the best that can be done within a federal data system that must keep employer's operational information confidential. One improvement over the status quo is to encourage the cooperative dissemination of cleaned up data and the open discussion of anti-suppression strategies, as is being done with this article. With today's ability to transfer data within the research community, we should be able to put behind us the days of lone researchers with undocumented handcrafted estimates that are not shared, leaving us each to invent our own crude wheels.

The federal data agencies could also improve the situation by greater cooperation within and across agencies. For example, CBP could be extended to include farming and government employment, sectors for which other branches of the Census Bureau collect data. A model in this regard is BEA's REIS, which combines information from several sources.

In the meantime, with its 1.5 million holes darned, the data created from *CBP* by these antisuppression methods might be the most detailed, consistent employment data ever publicly available. Useful for research, policy, and planning, they should satisfy some of the hunger at the frontiers of regional science.

NOTES

1. There also are a couple working papers (Gardocki and Baj 1985; Isserman and Sorenson 1987), brief accounts in manuals for practitioners (McLean and Voytek 1992; Redman 1994), and undocumented academic and commercial efforts. Gerking et al. (2000) had a longer discussion of the development of antisuppression methods.

2. The University of Illinois Library, a federal depository, lists 1946, 1947, 1948, 1949, 1951, 1953, 1956, 1959, and 1962 as the pre-1964 years.

3. Location and industrial classification are so important to regional research that the Census definition merits repeating: "An establishment is a single physical location at which business is conducted or services or industrial operations are performed. It is not necessarily identical with a company or enterprise, which may consist of one or more establishments. When two or more activities are carried on at a single location under a single ownership, all activities generally are grouped together as a single establishment. The entire establishment is classified on the basis of its major activity and all data are included in that classification" (U.S. Department of Commerce [U.S. DOC] 2005).

4. There are two exceptions to the hierarchy described here: North American Industrial Classification System (NAICS) 95—auxiliaries (except corporate, subsidiary, and regional management) and 99—unclassified establishments do not have 3-, 4-, 5-, or 6-digit components. Nationally in 2002, they accounted for 1,011,496 and 32,769 jobs, respectively. Effective with the 2003 data, NAICS 95 no longer exists in *County Business Patterns* (*CBP*); those auxiliary establishments are classified by the service they perform, for example, NAICS 518210—data processing, hosting, and related services or a 6-digit category of NAICS 484—truck transportation or NAICS 811—repair and maintenance.

5. These percentages are for 2002 and are the authors' calculations with Regional Economic Information System (REIS) data available online or CD-ROM from the Bureau of Economic Analysis (U.S. DOC 2006). REIS provides more comprehensive employment coverage than *CBP* but does not go below the 2-digit level, for example, manufacturing.

6. There are 444 instances within the 1997 NAICS of a classification that is not subdivided at the next finer levels of disaggregation, such as ice cream and frozen desert on the 5- and 6-digit levels. An industry is undivided over three levels in seventy cases and four levels in three cases. Logging is one of the former, and management of companies and corporations one of the latter. Sometimes an industry that is an only child has multiple children itself: for example, oil and gas extraction has siblings at the 3-digit level and is an only child at the 4- and 5-digit levels, before begetting NAICS 211111—crude petroleum & natural gas extraction and NAICS—211112 natural gas liquid extraction. The term *clone* will not be used to describe industries that are identical at two or more levels of the NAICS hierarchy.

7. After estimating suppressed Census of Manufactures numbers, Ellison and Glaeser (1997, 923) concluded, "the Census' withholding process is not sufficiently transparent that we felt confident that we could reasonably simulate it."

REFERENCES

Anselin, Luc. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27: 93-115.

Anselin, Luc, Attila Varga, and Zoltan Acs. 1997. Local geographic spillovers between university research and high technology innovations. *Journal of Urban Economics* 42: 422-48.

Clapp, John, Henry O. Pollakowski, and Lloyd Lynford. 1992. Intrametropolitan location and office market dynamics. *Journal of the American Real Estate and Urban Economics Association* 20: 229-57.

Duranton, Gilles, and Henry G. Overman. 2005. Testing for localization using micro-geographic data. *Review of Economic Studies* 72: 1077-1106.

Ellison, Glenn, and Edward L. Glaeser. 1997. Geographic concentration in U.S. manufacturing industries: A dartboard approach. *Journal of Political Economy* 105: 889-927.

Feser, Edward J. 2003. What regions do rather than make: A proposed set of knowledge-based occupation clusters. *Urban Studies* 40: 1937-58.

Feser, Edward J., and Edward M. Bergman. 2000. National industry cluster templates: A framework for applied regional cluster analysis. *Regional Studies* 34: 1-19.

Feser, Edward, Stuart Sweeney, and Henry Renski. 2005. A descriptive analysis of discrete U.S. industrial complexes. *Journal of Regional Science* 45: 395-419.

Fingleton, Bernard, Danilo Igliori, and Barry Moore. 2005. Cluster dynamics: New evidence and projections for computing services in Great Britain. *Journal of Regional Science* 45: 283-311.

Gardocki, B., and J. Baj. 1985. *Methodology for estimating non-disclosure in County Business Patterns*. De Kalb: Center for Governmental Studies, Northern Illinois University.

Gerking, Shelby D., and Andrew M. Isserman. 1981. Bifurcation and the time pattern of impacts in the economic base model. *Journal of Regional Science* 21: 451-67.

Gerking, Shelby, Andrew Isserman, Wayne Hamilton, Todd Pickton, Oleg Smirnov, and David Sorenson. 2000. Anti-suppressants and the creation and use of non-survey regional input-output models. In *Regional science perspectives in economic analysis: A festschrift in memory of Benjamin H. Stevens*, ed. Michael Lahr and Ronald Miller, 379-406. Amsterdam: Elsevier.

Glaeser, Edward L., Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer. 1992. Growth in cities. *Journal of Political Economy* 100: 1126-52.

Goodchild, Michael F., Luc Anselin, Richard Appelbaum, and Barbara Herr Harthorn. 2000. Toward spatially integrated social science. *International Regional Science Review* 23: 139-59.

Hammond, George, and Eric Thompson. 2004. Employment risk in U.S. metropolitan and nonmetropolitan regions: The influence of industrial specialization and population. *Journal of Regional Science* 44: 517-42.

Harrison, Bennett, Maryellen R. Kelley, and Jon Gant. 1996. Innovative firm behavior and local milieu: Exploring the intersection of agglomeration, firm effects, and technological change. *Economic Geography* 72: 233-58.

Henderson, J. Vernon. 1996. Effects of air quality regulation. *American Economic Review* 86: 789-813.

Henderson, Vernon, Ari Kuncoro, and Matt Turner. 1995. Industrial development in cities. *Journal of Political Economy* 103: 1067-90.

Holmes, Thomas J. 1998. The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy* 106: 667-705.

Holmes, Thomas J., and John J. Stevens. 2004. Spatial distribution of economic activities in North America. In *Handbook of regional and urban economics*, vol. 4, ed. J. Vernon Henderson and Jacques-Francois Thisse, 2797-2843. Amsterdam: Elsevier North Holland.

Isserman, Andrew, and Terance Rephann. 1995. The economic effects of the Appalachian Regional Commission: An empirical assessment of 26 years of regional development planning. *Journal of the American Planning Association* 61: 345-364.

Isserman, Andrew M. 1976. *A county-level, industry-specific employment data system.* Report to the Construction Engineering Research Laboratory, U.S. Army Corps of Engineers. Urbana: University of Illinois, Center for Advanced Computation.

———. 1977. The location quotient approach to estimating regional economic impacts, *Journal of the American Institute of Planners* 43: 33-41.

———. 1980. Estimating export activity in a regional economy: A theoretical and empirical analysis of alternative methods. *International Regional Science Review* 5: 155-84.

———. 2001. Competitive advantages of rural America in the next century. *International Regional Science Review* 24: 38-58.

———. 2005. In the national interest: Defining rural and urban correctly in research and public policy. *International Regional Science Review* 28: 465-99.

Isserman, Andrew M., and David J. Sorenson. 1987. *County employment: A description of federal data sources and the construction of a multi-source data base.* Research Report 8703. Morgantown: Regional Research Institute, West Virginia University.

Kahn, Matthew E. 2004. Domestic pollution havens: Evidence from cancer deaths in border counties. *Journal of Urban Economics* 56: 51-69.

Lacombe, Donald J. 2004. Does econometric methodology matter? An analysis of public policy using spatial econometric techniques. *Geographical Analysis* 36: 105-18.

Markusen, Ann. 1994. Studying regions by studying firms. *The Professional Geographer* 46: 477-90.

———. 1996. Sticky places in slippery space: A typology of industrial districts. *Economic Geography* 72: 293-313.

———. 2004. Targeting occupations in regional and community economic development. *Journal of the American Planning Association* 70: 253-68.

McLean, Mary L., and Kenneth P. Voytek. 1992. *Understanding your economy.* Chicago: American Planning Association.

Office of Management and Budget. 1987. *Standard industrial classification manual.* Springfield, VA: National Technical Information Service.

———. 1998. *North American Industrial Classification System: United States, 1997*. Springfield, VA: National Technical Information Service.

———. 2003. *North American Industrial Classification System: United States, 2002*. Springfield, VA: National Technical Information Service.

Orr, Blair, and Joseph Buongiorno. 1989. Improving estimates of employment in small geographical areas. *Journal of Economic and Social Measurement* 15: 225-35.

Partridge, Mark D., and Dan S. Rickman. 2005. High-poverty nonmetropolitan counties in America: Can economic development help? *International Regional Science Review* 28: 415-40.

Porter, Michael E. 2003. The economic performance of regions. *Regional Studies* 37: 549-78.

———. 2004. *Competitiveness in rural U.S. regions: Learning and research agenda.* Institute for Strategy and Competitiveness, Harvard Business School.

Redman, John M. 1994. *Understanding State Economies through industry studies.* Washington, DC: Council of Governors' Policy Advisors.

Reed, W. Robert, and Cynthia L. Rogers. 2003. A study of quasi-experimental control group methods for estimating policy impacts. *Regional Science and Urban Economics* 33: 3-25.

Scott, Allen J. 1996. The craft, fashion, and cultural-products industries of Los Angeles: Competitive dynamics and policy dilemmas in a multisectoral image-producing complex. *Annals of the Association of American Geographers* 86: 306-23.

U.S. Department of Commerce. 2005. *County Business Patterns CD-ROM: 2002 and 2003*. Washington, DC: U.S. Census Bureau.

———. 2006. *Regional Economic Information System (REIS) 1969-2004* (CD-ROM). Washington, DC: Bureau of Economic Analysis.