Employment in Nonprofit Entities: Coverage, Bias, and Measurement Errors in QCEW and public IRS Information, 1999-2003

Martin H David
Associate Scholar, Urban Institute, Washington DC 20036

Abstract

Accurate and timely estimates of employment in nonprofit organizations are essential to understanding the performance of the US economy. In the last decade nonprofit employment at national, state, and MSA levels has been estimated by matching the IRS *registry* of exempt-organization identifiers (EIN's) to employment reported to administrators of the Unemployment Compensation Insurance System, UC. (See Salamon-Sokolowski, 2005). EIN errors corrupted these estimates. Failed matches (false negative) reduced coverage and biased estimates. False positive matches added employment that is outside the nonprofit sector and distorted covariances between employment and attributes of the nonprofit entity.

This paper applies methods that reduce bias from false negative matches and discard likely false positive matches. False positive matches were removed by editing. [1] Some failed matches were identified by invalid EIN's in BLS *Quarterly Census of Employment and Wages*, QCEW. The probability of invalid EIN's was modeled using level of employment, state, and year. Probability of an invalid EIN was predicted for each establishment in each year. Weights were calculated from predicted probabilities that recover information in the universe from the subset of establishments with valid EIN's. Weights were applied to IRS records on nonprofit organizations that match the QCEW.

Keywords: Matching error, Model-based weighting, Nonprofit employment

1. Problem

This paper reports part of an effort to improve estimates of nonprofit employment by combining measures in IRS public data files and the high quality measures available for employers liable to pay Unemployment Compensation benefits. Nonprofit employment is growing faster than the private workforce (Salamon-Solokowski, 2005). Matching administrative records by Federal Employer Identification numbers (EIN's) has produced numerous estimates of nonprofit employment in the last decade. The IRS public *registry* of exempt organizations and the *QCEW*[2] constitute the most current and accurate sources of identifiers of nonprofits and employment counts, respectively. IRS Form 990/990EZ provide additional information on compensation and employment that complements the QCEW.

However, even the best sources of information contain errors in the EIN and produce matching errors. Invalid EIN's lead to failed matches (false negatives) and result in underestimating employment. Errors in reporting and processing also lead to false positive matches. False positive matches distort covariances and may cause overstatement of employment. This paper establishes the extent and distribution of invalid EIN's in the QCEW. A model of the probability of invalid EIN's is used to construct "nonresponse" weights. Weighted estimates from matched data offset bias and incomplete coverage resulting from failed matches. False positive matches are reduced by editing matched records. Estimates for 2003 excluded false positive matches. All Forms 990/990-EZ with expenses less than 5 times the first quarter payroll on QCEW and all nonprofits linked to NAICS sector 52 were delinked.

2. Conceptual structure

The notation used in studies of measurement error follows: $M_{ist} = 1$ indicates an IRS-QCEW match for the i^{th} EIN in state *s,* year *t*; otherwise $M_{ist} = 0$. M_{ist}* indicates true matches. $\theta \equiv \Pr[M=0 \mid M^*=1]$, $\eta \equiv \Pr[M=1 \mid M^*=0]$ define the probability of false negatives and false positives, respectively. The expectation of M* can be estimated as $EM^* = (EM - \eta) / (1- \theta - \eta)$. The model that estimates θ, conditions on state, employment, and year. As η is probably less than 0.06, we use the approximation, $EM^* \approx EM / (1- \theta)$.

Weights are controlled to the observed number of establishments, n_{st}, by state and year. $k_{st} \equiv n_{st} / \sum_i \theta_{ist}$. The establishment weight is $w_{ist} = 1/(1 - k_{st} \theta_{ist})$. (Alternatively, weights could be controlled to employment totals within each state.)

3. Data and modeling problems

The IRS public information was matched to the QCEW using EIN's. The IRS *Registry* used contains all EIN's that have historically been approved as exempt organizations. It is a "gold standard" of correctness.[3] A few EIN's on IRS Form 990/990-EZ do not match the *Registry.* These matches are typically newly-formed organizations whose approval is pending.

The QCEW for the first quarter of the years 1999 to 2003 is the universe for our model of θ. The universe includes 47 states, Puerto Rico, and the Virgin Islands.[4] QCEW was scanned for invalid EIN's. Invalid EIN's have one of three patterns: less than nine characters, uninformative character combinations (all '9', all '0'), or leading two characters followed by '0' or '9' ciphers).[5]

Numerous mistakes lead to invalid EIN's. These include reporting errors by the organization, change in legal form of the enterprise, errors in state processing of employers' unemployment insurance accounts, and errors in administration of the Federal Unemployment Tax Act (FUTA). In addition, some new enterprises do not have an EIN at the time they become liable for unemployment insurance payroll taxes. Some establishments lose an EIN as they are absorbed in mergers.

Invalid EIN's can also occur on FUTA records that BLS receives from the Employment and Training Administration of the Department of Labor. These EIN's are merged into the QCEW. In summary, invalid EIN's emanate from compliance failures by business, from idiosyncrasies of state administration, and the administration of FUTA.[6]

State, industry, and number of employees are likely explanatory variables for θ. Each of these variables is problematic. A combination of state and NAICS subsector would create 5000 cells. Many would be empty, precluding use of NAICS in this model. (NAICS requires specifications that smooth industry effects over states.) The functional relationship between number of employees and θ is ambiguous. Zero may mean a new employer, a failed business, or a seasonal business. We conjecture that new employers are more prone to transmitting invalid EIN's than others. However, seasonal employers or liquidating enterprises may also report no employees. Employers of 1-3 workers may have incomplete records and poor accounting methods that lead to inadvertent omissions of EIN. Employers of more workers may be more prone to change legal ownership (triggering a new EIN) than employers of less than 4 workers.

Year is used as a surrogate for changes in technology or processing of UC databases. We hypothesize that administrative capacity to detect and eliminate EIN errors increases over time. Temporal variation is modeled by a trend centered on the first quarter of 2001. More complex temporal patterns in the changing prevalence of invalid EIN's can not be estimated reliably within the 5-year universe for each state. However, nonlinear time trends in the prevalence of invalid EIN's are accommodated by the proportional adjustment, k_{st} for each state and year. The model prediction of number of worksites for each state is increased (or decreased) to the state universe of worksites by that proportional adjustment.

Two models were estimated to explore variation in the probability that worksites have an invalid EIN. The first pools all states and conditions on number of employees at each worksite and year. The employee effect is modeled as a broken line. Effects of number of employees increase (decrease) linearly over the intervals: 1-3, 4-7, 8-10, and 11-1,010 employees. Zero employment is reflected in the constant term of the model; employee effects are capped at 1,010. The second model fits worksite employee effects and time coefficients separately for each state. Had state effects proved insignificant the simpler model would have sufficed. See section 5.

4. Invalid EIN's in the QCEW, 1999-2003

Invalid EIN's occurred for 1.7% of all establishments over the five-year period and generally declined from 1999 to 2003. The rate of invalid EIN's varied tenfold across states. The vertical axis of Figure 1 displays the statewide average rate of invalid EIN's in the first quarter of 2001. That quarter is the center of the five years being analyzed. States in the highest quartile of valid EIN's are shown with diamonds; states in the lowest quartile are shown with rectangles. The horizontal line is at the median rate of invalid EIN's. The vertical line signifies the absence of trend in invalid EIN levels.

Over the 5-year period, the rate of invalid EIN's declined in 36 of the 49 jurisdictions. The x-axis (Figure 1) displays the model coefficient for the trend. (The x-axis does not display percentage points.) Negative trends dominate for all quartiles, showing that the rate of invalid EIN's generally declined from 1999-2003. Among states with the highest probability of invalid EIN's (the rectangles) only two show a positive trend. The trend coefficient shown will be multiplied by 1 in 2002 and 2 in 2003 and added to the employee effects for worksites in computing a predicted probability of invalid EIN for those years.

Displaying the impact of employees on the probability of invalid EIN's is difficult. Although employee effects are correlated across states, the 49 jurisdictions vary widely. Most show decreasing rates as number of employees increases; a few show a u-shaped pattern. The predicted probability associated with ten levels of employees is shown in Appendix Table A1. (The levels chosen correspond to break-points in the employee coefficients.) The coefficients for the model are shown for each jurisdiction in Appendix Table A2.[7]

5. Evaluating the model

Preliminary examination of a single year (2000) revealed that state effects could not be ignored. Adding fixed-effects for states modulated the number-of-employee effects (constrained to be identical in all states). High covariances of state effects with number of employee parameters indicate that identical number of employee effects across all states is not tenable. The 5-year data give more information about each state.

The model is estimated on 35 million establishments in the database. This large universe assures that extremely small variations in the level of invalid errors can be detected as "significant" effects in a single year or a single state. However, descriptive statistics show that the probability of invalid EIN's is not stationary over time. Nor are trends identical among states.

Modeling demonstrates that invalid EIN's are not missing completely at random. The 49-state, 5-year average rate of 1.7% is not uniformly distributed within states nor is it on average the same across states. Conditioning on year, state, and employee is a first step towards understanding how invalid EIN's bias outcomes of matching data to the QCEW across record systems (or matching QCEW across states). Weights constructed from the model estimates reduced bias from failed matches.

Differences in trends across states may indicate degrees of effort applied to eliminating invalid EIN's. However, the variability in trends for the 12 states with the lowest rates of invalid EIN's suggest that year-to-year shocks may push rates away from an irreducible minimum level. Extrapolating trends to out-of-universe years is unwise.

Theory and practical proxies for poor record-keeping or lax administration could improve the model. Estimates are calculated as if observations were independently distributed. That is inappropriate. Some organizations have multiple branches. One EIN corresponds to each organization's branches in a particular state. Modeling each organization in each state as a single observation would provide more appropriate estimates.

Rates of invalid EIN's vary widely across NAICS subsectors (Section 6). This indicates that industry class should be incorporated into the model.

4. Matching QCEW to the IRS EIN's

Registry matches. Before removing false positive matches, 3.7 percent of QCEW establishments match to a nonprofit organization in the *Registry* (Table 1, last row). Establishments in 4 states were excluded from this tally. Coverage for our comparison extends to nearly 90 percent of establishments.[8] Matches to the registry increase systematically over the 5-year period. All 501(c) organizations are included in the match. Charitable organizations, IRC §501(c)(3), dominate the matched cases (Table 2).

The registry can be compared to private industry rather than the universe of all UC employers. Private nonprofit organizations may employ workers within government entities. For example, organizations may be co-located in governments. A parent-teacher organization or booster club for the home team may be associated with a public school district; foundations supporting charity to patients may operate on the premises of municipal hospitals.

Registry and Form 990/990-EZ matches. Weighting increases the count of employees in §501(c)(3), charitable organizations by an average of 111,000 in each of the five years studied. The upward adjustment derives from 2,300 establishments predicted to have failed matches in each year. The mean probability for invalid EIN's is 0.0135, less than the 0.017 for the QCEW universe. Predicted probabilities are lower for nonprofit establishments than for the QCEW universe because the conditioning variables have a different distribution in the nonprofit sector.

Table 2 reports weighted employment and establishments for 2003, including subtotals for charitable organizations, §501(c)(3), and all other exempted organizations. The total row includes organizations exempted under subsections 11 or higher. The 8.7 million charitable organizations needs to be extrapolated to the US as a whole. It compares to a national estimate of 8.8 million for 2002 estimated by the Salamon-Solokowski (2005) without weights or eliminating false positives.

6. Nonprofit prevalence by NAICS

Table 3 summarizes the prevalence of invalid EIN's in the QCEW for 2002 by NAICS subsectors (Appendix Table A3). These rates average across all states and employment levels. Several features of the distribution are notable. The median is 0.9%; the Q1-Q3 range is 0.7%; and the 90th percentile is 1.8%. Six of the subsectors with large concentrations of nonprofit organizations are above Q3. Clearly further work on modeling industry impacts on the probability of invalid ein is in order.

Conclusions

Three important conclusions emerge from this work. Matching QCEW to IRS information on nonprofits can not proceed without discarding false positive matches. Weighting QCEW data for invalid EIN reduces bias in estimating counts of worksites and employment obtained from EIN matches. A more satisfactory

solution to both problems is to verify EIN numbers as they are processed by BLS and IRS.

Nonprofits are dispersed into many industries and have a substantial share of employment in some. Exempt status needs to be accurately measured across a broad spectrum of the economy. Regularly linking exempt status to the QCEW will provide a significant statistical product.

Acknowledgements[9]

The Bureau of Labor Statistics kindly provided access to the QCEW in its research enclave. Rick Clayton, David Talan, Amy Knaup, and Merissa Piazza provided guidance and helpful comments.

The Center For Nonprofits, Urban Institute, supplied its database and training in its use. Tom Pollak and Linda Lampkin provided technical information about IRS information returns. Jen Auer and Kendall Golladay assisted and provided useful checking on inconsistencies in the match.

Errors are my own.

References

David, Martin H, Thomas Pollak and Paul Arnsberger. 2006 Compliance with information reporting: Exempt Organizations. *IRS Research Bulletin: Recent Research on Tax Administration and Compliance* (Publication 1500) 231-245.

Fellegi, I. P., and A. B. Sunter. 1969. A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183--1210.

Foster, Lucia, Joel Elvery, C. J. Krizan, David Talan. 2007. Preliminary Micro Data Results from the Business List Comparison Project. *Proceedings of the American Statistical Association*, 2006 (forthcoming).

Gronbjerg, Kirsten A. and Erich T. Eschmannn. May 2005. Indiana nonprofit employment. Ctr. on Philanthropy, School of Public and Environmental Affairs, Indiana University and the Johns Hopkins Nonprofit employment data project.

Independent Sector. 2003. *The Nonprofit Almanac: 2002*.

Michigan Nonprofit research program. 2005. Economic benefits of Michigan's nonprofit sector: 2004.

Okolie, Cordelia. 2004 July. Why size class methodology matters in analyses of net and gross job flows. Monthly Labor Review, 3–12.

Salamon, Lester M. and S. Wojciech Sokolowski. 2005 Sept. Nonprofit organizations: new insights from QCEW data. Monthly Labor Review, 19–26.

Spletzer, James and Joel Elvery. 21 October 2005. Presentation to CNSTAT workshop "Benefits of Interagency Business Data Sharing".

Vilhuber, Lars, Bryce E. Stephens, John M. Abowd, Fredrik Andersson, Kevin L. McKinney, Marc Roemer and Simon Woodcock (forthcoming, draft 2005 CRIW conference at NBER.org) The LEHD Infrastructure files and the creation of the Quarterly Workforce Indicators.

Winkler, William E. 2004. Methods for evaluating and creating data quality. Information Systems, 29: 531-550.

---

[1] When wages reported to UC exceed expenses reported to the IRS, the match is not credible.

[2] Some estimates have used UC data obtained from state administrators. QCEW enhances state data by breaking out worksites and regularly updating NAICS.

[3] Only 500 out of 1.9 million records in the cumulative *Registry* maintained by the Urban Institute are invalid, according to the criteria described.

[4] Massachusetts, Michigan, New York, and Wyoming refused to release their data for investigation at the BLS secure research site.

[5] Ein's were also checked against a list of "legal" combinations for the first-two digits (originally referring to IRS districts). Less than 0.002 of establishments have ein's with invalid first-two digits. Some that are geographically concentrated may be legitimate. We did not count any of these cases as invalid.

[6] States have unique identifiers for each business in the UC system and do not use the Federal EIN. States have little need to detect and eliminate incorrect EIN's. Similarly, the IRS has little incentive to correct EIN's used to file Federal Unemployment Tax payments.

[7] Available from the author on request.

[8] Non-cooperating states are headquarters for many well-known nonprofits. If they have branches in included states, employees in those branches are included in the QCEW matched employee count. Employees of organizations that operate *only* in excluded states are not counted.

[9] Empirical work is based on the authors' calculations and is exploratory research. Views expressed are those of the author and do not reflect policies and estimates of the Department of Labor or the Bureau of Labor Statistics.

Table 1. Universe and match to Registry, 1999-2003 (including false positive matches)

Source	entity	yr1999	yr2000	yr2001	yr2002	yr2003	Total
Universe	estab	7,820,860	7,879,116	7,984,529	8,101,872	8,228,840	
Private industry	Estab	7,560,567	7,622,274	7,724,965	7,839,903	7963,340	
49-state,q1	Estab	6,821,207	6,895,095	6,999,209	7,142,948	7253,238	
Coverage	Estab	0.8722	0.8751	0.8766	0.8816	0.8814	
Masterldb	Estab.	246,086	253,704	259,136	264,889	269,959	1293774
prevalence, matches		0.0361	0.0368	0.0370	0.0371	0.0372	

Table 2. Employment, establishments by subsection 501(c), 2003q1 (excluding false positive matches) in thousands

Subsection 501(c)	501(c)	Organ-izations	Establishments		QCEW employment	
			Count	Wtd.	Count	Wtd.
Charities	**03***	277	173	*175*	8,514	*8,717*
Select	**01-03****	2	6	*6*	521	*535*
others	**04**	7	9	*9*	128	*130*
	05	14	17	*17*	171	*172*
	06	15	17	*17*	158	*164*
	07	8	8	*9*	149	*150*
	08-10	12	15	*16*	246	*250*
Total Others		57	73	*74*	1,372	*1,402*
All		334	246	*249*	9,886	*10,118*

* Includes 2,000 NA subsection

**Includes 500 NA subsection, delinked from a financial institution.

Table 3. Invalid EIN rate by NAICS

NAICS	Quantile	Rate
335	10	0.45
324	Q1	0.64
337	50	0.90
113	Q3	1.34
924	90	1.80
	Industries with large nonprofit concentrations	
611	99	7.27
621	34	7.24
622	88	1.74
623	79	1.47
624	86	1.66
711	81	1.55
712	80	1.53

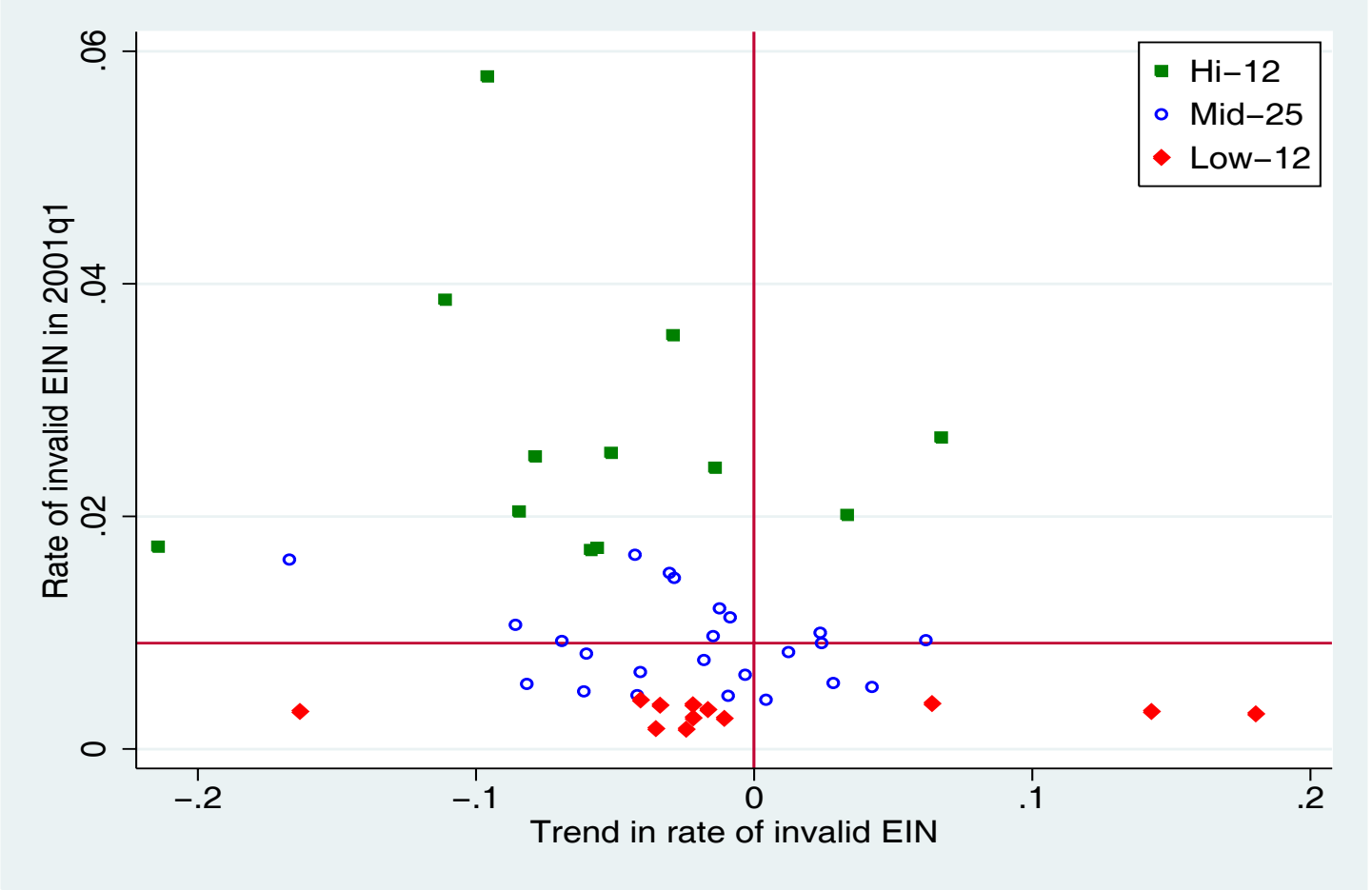

Fig.1 Modeled trend in 49 jurisdictions by rate of invalid EIN in 2001