# The Feasibility of State Corporate Data

*Brian D. Francis, Internal Revenue Service*

ompiling State level corporate financial statistics from tax returns can be risky. One reason is because a company is free to file a tax return in a State where it has no operations. The Internal Revenue Service's (IRS) Statistics of Income Division (SOI) has not produced these data since 1962 for precisely this reason. We posit that if a strong relationship exists between gross receipts reported on tax returns and employment levels by State, then meaningful corporate statistics can be produced at the State level. This paper tests this theory by matching receipts and employment by State and industry and computing correlation coefficients.

Gross receipts for industries by State were obtained from the SOI Corporation Statistics Branch; the Bureau of Labor Statistics (BLS) provided employment data by State and industry. The Small Business Administration (SBA) compiles these data and publishes them through its Office of Advocacy. Common elements in all three of these sources are State identification and Standard Industrial Classification.

## ■ Statistics of Income/Bureau of Labor Statistics

SOI creates population estimates from a sample of corporation income tax returns (Form 1120 series) filed with the IRS.[1] This study uses data from corporate fiscal year 1995. For the period July 1, 1995 to June 30, 1996, the total number of corporate returns filed was 4,852,186. This population was stratified by asset class. Sampling rates ranged from .25 percent to 100 percent, generating a total sample of 97,605 returns. The breakpoint for 100 percent sampling was for corporations reporting total assets of $50 million or more.

BLS publishes employment data under the Covered Employment and Wages or ES-202 program.[2] The data pertain to workers covered under State unemployment insurance laws and to Federal civilian workers covered by the Unemployment Compensation for Federal Employees program. Descriptive statistics for the SOI and BLS data are given in Table 1.

## Table 1

| Industry | SOI Receipts | | BLS Employment | | |
| --- | --- | --- | --- | --- | --- |
| | Mean | Standard Deviation | Mean | Standard Deviation | Observations |
| Agriculture | 140,681 | 492,066 | 32,949 | 70,126 | 24 |
| Mining | 3,023,715 | 4,472,343 | 11,560 | 22,870 | 15 |
| Construction | 1,192,711 | 2,571,821 | 102,057 | 100,627 | 28 |
| Manufacturing | 55,301,530 | 83,580,386 | 361,761 | 373,158 | 47 |
| Transportation and Public Utilities | 18,501,799 | 27,922,157 | 115,963 | 123,692 | 39 |
| Retail Trade | 19,630,682 | 34,289,291 | 126,094 | 140,858 | 34 |
| Wholesale Trade | 16,591,961 | 22,796,358 | 419,105 | 426,247 | 44 |
| Finance, Insurance and Real Estate | 15,914,724 | 28,651,401 | 130,798 | 156,700 | 46 |
| Services | 4,361,607 | 28,651,401 | 618,753 | 701,493 | 47 |

As mentioned above, the dominance of large companies may thwart attempts to produce meaningful receipt data by State. A solution explored below is to break down receipt data into several classes, which allows correlation at lower levels to be revealed. Aggregate correlation analysis shows that most industrial groups exhibit strong relationships *without* a further breakdown. These statistics are presented in Table 2.

SAS statistical software was used to produce Pearson correlation coefficients between receipt and employment variables (see Appendix for formula). The initial analysis tested the relationship between SOI business receipts and average employment from BLS by industrial division, with the 50 States and the District of Columbia as observations.

Further analysis by receipt class (RC) was performed on the four weakest relationships in Table 1. Figure A displays the correlation coefficients for Mining, Manufacturing, Transportation and Public Utilities, and Finance, Insurance, and Real Estate.

**Table 2–Aggregate Level Correlation of SOI/BLS Data**

| Industry Classification | Pearson Correlation Coefficient |
|---|---|
| Agriculture, forestry, and fishing | 0.92220 |
| Mining | 0.83553 |
| Construction | 0.97572 |
| Manufacturing | 0.85752 |
| Transportation and public utilities | 0.82502 |
| Wholesale trade | 0.91171 |
| Retail trade | 0.91260 |
| Finance, insurance, and real estate | 0.86182 |
| Services | 0.97006 |



Figure A--Selected Correlations of SOI/BLS Data
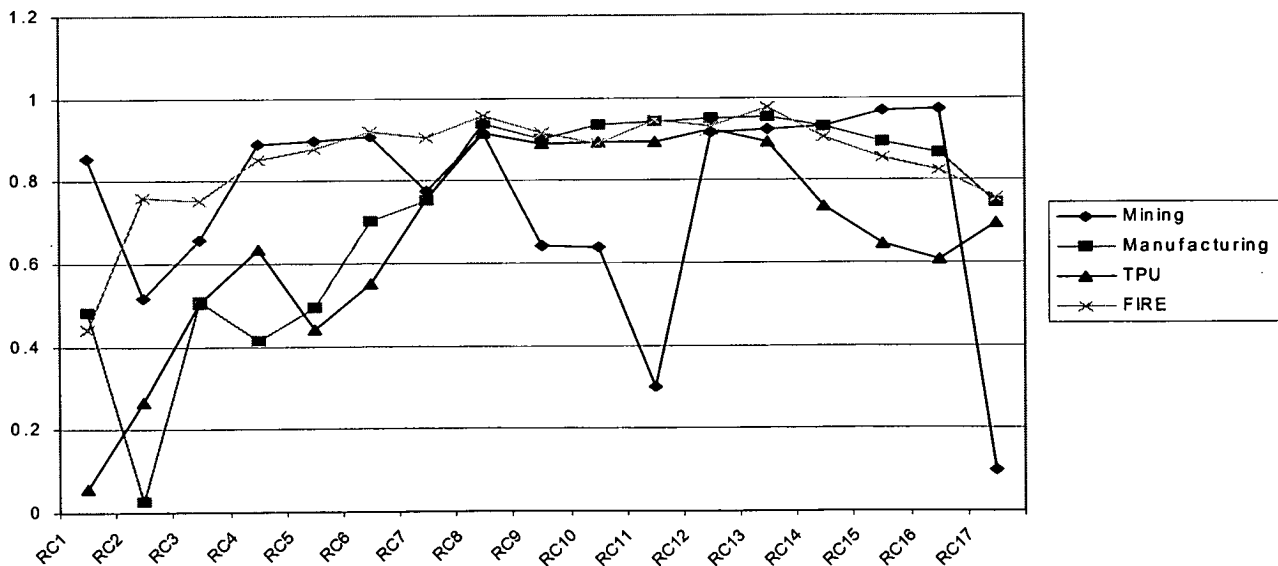
Notes: Table 3 on the following page indicates the receipt size classes designated by the labels on the horizontal axis, i.e., RC1-RC17. Note the strong relationships, a Pearson coefficient of 0.9 or more, for several receipt classes. The correlation appears weakest for the lowest and highest receipt class.

**Table 3–Receipt Class Designations**

| Gross Receipts | Receipt Class |
|---|---|
| 1$ under $5,000 | 1 |
| $5,000 under $10,000 | 2 |
| $10,000 under $25,000 | 3 |
| $25,000 under $50,000 | 4 |
| $50,000 under $100,000 | 5 |
| $100,000 under $250,000 | 6 |
| $250,000 under $500,000 | 7 |
| $500,000 under $1,000,000 | 8 |
| $1,000,000 under $2,500,000 | 9 |
| $2,500,000 under $5,000,000 | 10 |
| $5,000,000 under $10,000,000 | 11 |
| $10,000,000 under $50,000,000 | 12 |
| $50,000,000 under $100,000,000 | 13 |
| $100,000,000 under $250,000,000 | 14 |
| $250,000,000 under $500,000,000 | 15 |
| $500,000,000 under $1,000,000,000 | 16 |
| $1,000,000,000 or more | 17 |

## ■ Small Business Administration

The Small Business Administration (SBA) publishes data on firms showing number of establishments and employment, as well as business receipts.[3] The SBA Office of Advocacy contracts with the U.S. Bureau of the Census to provide firm size data estimated from their County Business Patterns program. Only firms with employees are included. Data on self-employed individuals are not considered. The Office of Advocacy obtains business receipts from the Internal Revenue Service and merges these data with the firm information from the Census Bureau by State and industry division. Descriptive statistics for the SBA data are given in Table 4.

We can then perform the same experiment as above. That is, does the relationship between estimated receipts and employment hold up if the data are stratified by State and industry? Table 5 shows the Pearson coefficients correlating employment and estimated receipts from the SBA data file.

**Table 4**

| Industry | SBA Estimated Receipts | | SBA Employment | | Observations |
|---|---|---|---|---|---|
| | Mean | Standard Deviation | Mean | Standard Deviation | |
| Agriculture | 675,615 | 968,487 | 11,453 | 16,551 | 51 |
| Mining | 2,988,671 | 7,930,521 | 10,526 | 26,537 | 48 |
| Construction | 13,032,196 | 13,218,202 | 98,552 | 98,208 | 51 |
| Manufacturing | 67,012,999 | 72,199,451 | 364,924 | 378,155 | 51 |
| Transportation and Public Utilities | 20,041,801 | 22,746,825 | 115,414 | 127,243 | 51 |
| Retail Trade | 75,470,640 | 92,477,039 | 129,540 | 152,033 | 51 |
| Wholesale Trade | 44,955,522 | 48,502,040 | 413,463 | 424,293 | 51 |
| Finance, Insurance, and Real Estate | 40,973,954 | 59,118,974 | 136,639 | 167,868 | 51 |
| Services | 42,917,884 | 54,580,609 | 680,581 | 768,146 | 51 |

**Table 5–Pearson Coefficients of Employment and Business Receipts by Firm Size**

| Industry | Employees per establishment | | | | | | |
|---|---|---|---|---|---|---|---|
| | Overall | 1-4 | 5-9 | 10-19 | 20-99 | 100-499 | 500+ |
| Agricultural Services | 0.98909 | 0.97542 | 0.99156 | 0.99461 | 0.98525 | 0.95276 | 0.96112 |
| Mining | 0.99842 | 0.99107 | 0.99482 | 0.99303 | 0.98422 | 0.97930 | 0.98718 |
| Construction | 0.97835 | 0.98342 | 0.97947 | 0.97267 | 0.97490 | 0.96295 | 0.98171 |
| Manufacturing | 0.99299 | 0.99208 | 0.99428 | 0.99507 | 0.99603 | 0.99338 | 0.97376 |
| Transportation and Public Utilities | 0.99326 | 0.98705 | 0.98963 | 0.99232 | 0.99400 | 0.98702 | 0.99000 |
| Wholesale Trade | 0.99491 | 0.98994 | 0.99041 | 0.99258 | 0.99513 | 0.99281 | 0.99405 |
| Retail Trade | 0.99616 | 0.99153 | 0.99476 | 0.99598 | 0.99242 | 0.98560 | 0.98561 |
| Finance and Insurance | 0.95548 | 0.98239 | 0.97757 | 0.96836 | 0.92938 | 0.90166 | 0.97602 |
| Services | 0.99060 | 0.98863 | 0.99356 | 0.99230 | 0.99142 | 0.98648 | 0.98839 |

Note: Table 5 shows strong relationships across the board. The probability of observing a larger coefficient is .0001 for all cells in the table.

# Conclusion

The evidence provided here gives strong support to the feasibility of producing meaningful State corporate data. Certainly, the aforementioned caveats regarding the mobility of firms' reporting exists. It just does not appear to be strong enough to dilute the relationships of available State data. Spillover firms are undoubtedly responsible for a portion of the high correlation. Say a large corporation files its tax return in a State different from where it operates. In the latter State, there will be smaller firms that are classified under the same industry division and do file where they operate. A linked data base (similar to the Worker-Establishment Characteristic Data Base [4]) would be needed to ascertain the true nature of the relationship of business receipts to employment.

# Notes and References

[1] See Internal Revenue Service, Statistics of Income–1995 *Corporation Income Tax Returns*, Washington, DC, 1998.

[2] U.S. Department of Labor, Bureau of Labor Statistics, *Employment and Wages Annual Averages*, 1995.

[3] See the SBA web site on the Internet at: http://www.sba.gov/ADVO/stats/int_data/html.

[4] The Worker-Establishment Characteristic Data Base (WECD) attempts to combine information on worker characteristics obtained from the Longitudinal Research Data Base. The WECD is based on 1990 Census data.

# Appendix

Pearson's correlation coefficient, r, can be expressed as:

$$r = \left[ \frac{\sum xy - 1/n\left(\sum x\right)\left(\sum y\right)}{\sqrt{\sum x^2 - 1/n\left(\sum x\right)^2 - \sum y^2 - 1/n\left(\sum y\right)^2}} \right]$$

where  n = number of observations,
x = employment, and
y = business receipts.