WHY EVALUATE BUSINESS DATA COLLECTIONS?

Martin H. David
Joint Program in Survey Methodology, University of Maryland – College Park

Abstract

More than25 years have passed since Bailar and Kallek, Wolter and Monsour (1986) evaluated the *Economic Census*. Twelve years have passed since the Internal Revenue Service conducted its 1988 *Taxpayer Compliance Monitoring Program*. In that time much has been learned about data capture, improving business lists, imputation, and editing. However, findings about measurement error in business statistics leave questions about the quality of data that are the foundation for national account estimates. Those questions raise further questions about the quality of models of the economy and forecasts. Each of these facts motivate more frequent, comprehensive, and cumulative evaluations of data collected from business. Understanding the associated variance and biases of those measures assists in improving data quality and creating more valuable official statistics.

Keywords: Survey methodology, experiments, administrative records, business statistics

1 Introduction – Global Positioning Sensors[1]

Users of Federal statistics on business know too little about the quality of the data they devour. Producers of those statistics, particularly producers of monthly series and leading economic indicators, have too little time, too many obligations, and too few resources to evaluate their methods and to publish reports on data quality. How do we go beyond this vicious cycle?

I think I see a way. It rests on three assumptions. First, enormous support (financial and otherwise) will come from users who are convinced of the quality of information. Second, the route to improved quality is experimentation in design and careful evaluation of the results. Third, experimentation and evaluation need to be continuous to keep the statistical system tracking the changing economy. Let me begin with an example of how an improvement in the quality of measurement has changed the number of people who use measurement for making money.

This talk will translate lessons from the physical sciences into lessons for measuring the economy from surveys and censuses on organizations. I will use global positioning sensors (GPS) to illustrate the physical sciences.[2] The value of precision in measurement of position on the globe and the importance of evaluating observations to produce that precision will be explained. What is the connection between GPS and measuring the economy? Both depend on evaluation to assess the quality of information provided to users. Both incorporate redundancy to compare alternative

[1]Assistance and inspiration from John Gates, Judy Dodds, Patrick Cantwell, and Carol Caldwell is appreciated. The errors are mine.

[2]How do GPS locate where you are? Past measurement tells us how fast satellite signals travel to earth. (The speed of light [electro-magnetic waves] is the most precisely measured physical constant and governs satellite transmission speeds.) The signals broadcast time measured by a fiendishly accurate atomic clock. Thus the signal tells when it was transmitted. Comparing that time to the equally accurate clock on the ground gives a derived measure of distance. Now we use Euclid's geometry – The locus of distances from three satellite signals is a point on the ground. Adding more signals provides redundancy through which three or more estimates of ground position can be compared, and the best estimate derived. (GPS will use as many as seven satellite signals to refine the estimate.)

sources of information. Both depend on theory to structure measurement. Both entail recovering best estimates from noisy data.

GPS are handheld units that interpret signals from satellites to give position on the ground to the nearest square yard. They also provide altitude and links to computerized maps. For a modest cost, they give global coordinates of position – what Columbus would have given his right arm to obtain in 1492. Of course, they are playthings for the techies, a consumer toy available for about $400. But they also have restructured industries. Two examples make this clear.

Twenty years ago the trucking industry was deregulated. Many willing individuals entered the business, but found it difficult to locate business to fill the truck when shippers provided partial loads. They also often were empty on the backhaul. Dispatchers for the company were sometimes uncertain where trucks were; they could not contact the driver until he called. Now we have a totally different situation. GPS on the truck broadcast position to home base over cell phones once or more an hour. Dispatchers divert trucks with partial loads and empty backhauls to pick up goods. These diversions reduce the elapsed time between ordering and receiving trucking services. Because trucks carry larger loads, because drivers can be advised of road delays, and because problems with pickup and delivery can be detected quickly, goods can be carried more cheaply and in less time. The combination of cell phone and GPS technologies ignited this revolution in the trucking industry.

GPS created "precision" farming. Farmers now equip tractors and combines with on-board computers and positioning equipment. The equipment senses location in the field, links location to soil maps, and causes seeding and fertilizer application rates to vary across the field. Later the same equipment records the yield for every square yard of the field. Seed, feed, and soil inputs for each square yard are correlated to yield in the same area. Over several seasons farmers can use these correlations to "tune" application rates for every square yard to the most profitable outcome.

Understanding that business gains profit from relatively inexpensive measurements helps us to understand why "real-time" measurement of the economy can also improve the efficiency of business. Were economic measurements as precise and timely as global position readouts, demand for economic data would be astronomical.

What does this tell us about evaluating measurements of business activity? Four elements are crucial to GPS measurements:

> Theory: Without deductive reasoning from spherical geometry and the logical relationship between speed and distance, satellite signals would be useless.
> Redundancy: Data sufficient to generate one estimate of position contain measurement error. Additional contemporaneous estimates give the power of larger samples and the ability to use statistics to create a best estimate.
> Knowledge about error: Years of observation of satellite signals has generated precise information about the "wobble" in orbits that produces changing bias in the signals.
> Modeling data to produce "best estimates": As distributions of error in the signals are known, efficient and consistent statistics can be calculated from small samples.

How do business measurements compare? Theory is not robust. For example, we have difficulty

in defining a unit of output in service industries. Few measurement systems incorporate significant redundancy. Little is known about measurement error associated with particular sources of information (the enterprise, company, plant, or process being measured). Little is known about retrieving information within the business and transmission of information through respondents. Lacking theory and knowledge, "best estimates" often are viewed with unwarranted assumptions about what published estimates and variances connote.

This negative picture is not cause for gloom. Columbus discovered the new world without a GPS. Instead, the similarity between the problem of locating position on the globe and understanding where the economy is, calls for strengthening the four dimensions of activity that created GPS: theory, redundancy, knowledge about error, and modeling data to produce "best estimates".

2 The value and failures of theory

2.1 Anecdote: A phantom tax shelter boom?

A year ago the Congress held hearings on the precipitous decline in corporate income tax levies and raised the specter that tax shelters are eroding the basic accounting framework of tax law. The Treasury reported that discrepancies between book income reported to the SEC and income appearing on corporate returns increased from less than $20m in 1991-1992 to over $120m in 1997 for corporations with more than $1b of assets (Manzon-Plesko 2001, Figure 1). Manzon and Plesko (2001) simulated the effect of existing tax rules on income reported to shareholders. Taxable income simulated from the SEC data shows about $60b difference for a sample of 365 large corporations. A more tightly reasoned reconstruction of taxable income shows that nearly 70 percent of variance in the difference between book values and values adjusted for existing tax provisions can be explained by a dozen characteristics of the firm and its accounts in the current year.[3]

2.2 Why is this anecdote important?

The logic of accounting (theory) tells us that income is the difference between revenue and cost; furthermore accounting tells us that equity plus liabilities equal the value of assets. The difference in tax accounting and accounting permitted for the presentation of corporate information to shareholders lies in differences between advisory rules issued by the Financial Accounting Standards Board (FASB, GAAP) and formulae created by statute and administrative rules under the *Internal Revenue Code.* These differences matter! Two examples make the point.

First, the difference between value of shipments plus change in inventory and costs of goods shipped

---

[3]Petrick (2001) compares NIPA corporate profits to the earnings of S&P corporations, a comparison that gives insights to the problems of differences in accounting system as well as differences in coverage. Manzon and Plesko draw from a larger universe than the S&P500 as they obtain their panel from all corporations in the COMPUSTAT database, which represents all publicly traded companies, and therefore has less turnover than the companies in the S&P index.

The *Washington Post,* 22 July 2001 Business, p.1, reports on additional creative accounting practices that may enter respondents' reports to the government – giving impressions that incomes are larger or costs less than generally accepted accounting practices.

(a major concern in analyzing the performance of establishments in manufacturing), will depend on the accounting system in use. Every measurement needs to be related to the rule system that generated the accounts.

Second, when tax records are imputed in place of direct collection of data from small business firms, it is necessary to know that no systematic differences between the tax accounting and generally accepted accounting practices bias the procedure. (Moreover, the variance of the difference between survey measures and tax data must be known. See below on error.)

These examples illustrate that our system of economic measurement does not take proper account of the accounting systems used by responding entities. Two hundred years ago a similar problem aggravated measurement of location on the ground . Explorers and sailors knew little about the difference between the location of the north pole and the magnetic north that was measured by their compasses. They wandered about northern Canada and its polar waters with absurd discrepancies between their sextant readings and their compasses. Evaluation and measurement resolved the problem. The same can be done for economic measurements. For a start, design of surveys must use CPA's and tax accountants to help us collect enough information to interpret the data we analyze.

3 What kind of redundancy do we need?

The satellites that provide data to GPS were placed in orbit by design. A satellite is not poetically positioned, "I shot an arrow in the air, it came to earth, I know not where ...". A number of satellites had to be visible from any point on the globe. Equally important, the collection of satellites has to transmit appropriate signals even though particular satellites fail.

Two examples help to understand why we need redundancy in the measurement of economic data.

4.1 Retail sales

Retail sales are one of the leading indicators of the economy. Estimates of level and change are devoured by the media. Quarterly GDP estimates rely on change in retail trade as an indicator that can be used to estimate components of personal consumption expenditures. The advance report of quarterly GDP appears less than one month following the end of the quarter. Thus the advance quarterly GDP contains one month where advance estimates of retail trade drive estimates of "most goods" in personal consumption expenditures (just under a quarter of GDP). Thus downstream consequences of the quality of the advance retail sales estimate strongly influence GDP.

Understandably, estimates of retail sales contain several important redundancies: Advance, preliminary, and final estimates of monthly sales are released approximately 15, 45, and 75 days following the month of sales activity. Data are collected through the sample of the *Monthly Retail Trade Survey* (MRTS). The sample for the *Annual Retail Trade Survey* (ARTS) includes the MRTS sample and additional entities sampled from the noncertainty stratum. ARTS yields estimates of annual sales that are used to benchmark the monthly estimates. Errors in the advance MRTS are reduced by subsequent preliminary and final MRTS estimates, subsequent ARTS, and *Economic*

Census estimates.[4] The retail trade measurement programs appear to have an appropriate level of redundancy.

However, one sample drives the all the monthly estimates. A select subsample is asked to supply preliminary data within the first five business days of the month. Several problems arise because of the short time to respond. (1) Substantial censoring occurs as many entities are unable or unwilling to respond within the reporting period. It is clear that reporting units do not represent non-reporters and the non-response is non-ignorable.[5] (2) Some of the advanced reports are judgements rather than readout from auditable accounting systems. (3) Attrition is probably exacerbated by the pressure to provide advance data. Lastly, significant differences in responses made within 30 days and responses made for the second prior month up to 60 days after the month being reported were detected by Cantwell and Caldwell (1998) under a different sample design.[6]

Cantwell and Black (1998) discuss alternative sample designs. One design includes two independent samples: one to provide advance and the other to provide preliminary estimates of monthly sales. Independent samples can provide a basis for evaluating several current procedures. Currently some companies report for their past month twice: first an advanced, then a preliminary sales figure. The remaining companies report only once. The advanced figure is judged as accurate as reports made by units that have a full month to prepare their response. (Effectively, response to the advance response is substituted for a response that could have been made later in the month.)

Using two samples allows us to learn more about the quality of preliminary estimates. What questions can we answer?

A. Does the knowledge that the unit has two reporting opportunities decrease the quality of the advance response? (Conversely, does the knowledge the unit has exactly one opportunity to report increase motivation for an accurate response?)
B. Does asking some respondents for two estimates increase attrition from the MRTS panel? and
C. Is selective transfer of advance estimates to the preliminary database unbiased?

Answers to these questions provide important information about data quality. The cost of these answers is modest. Additional observations in two independent samples are costly. Additional sample costs will be offset in two ways. Follow-up for second reports will be eliminated. The cost of transferring some (but not all) advance estimates to the database for preliminary estimates will be eliminated. Lastly, reducing the burden of two reports may increase response rates and the care with which responses about advance sales are prepared.

---

[4] Corresponding adjustments revise consumer expenditures when *Economic Census* data for benchmark years triggers revision of the GDP product series.

[5] Modeling, extrapolation, and forecasting are required to arrive at the advanced estimate.

[6] Part of the difference in that design stems from smaller sample size and an unbalanced rotating panel design, neither of which are pertinent in this context. Bias due to the *differences* in reports within the first 30 days and reports made subsequently is essential.

4.2 Anecdote: Redundancy in business lists, the BEL and SSEL

Both the Bureau of the Census and the Bureau of Labor statistics compile a list of businesses. Most of you know them as the SSEL and the BEL. For twenty years or more, advocates of statistical quality have advocated pooling information from the two agencies to create a unified list to be used by the entire statistical system for sampling. I plead the fifth amendment on the question of a unified list. Deep problems of confidentiality and protecting data suppliers lurk in the proposal. I prefer to talk about the value of differences in methodology for the two lists and how we can learn from them.

Both lists use tax records as a principal source of information about businesses. BLS concentrates on employers whose places of work are identified through filings for state unemployment insurance taxes, and a supplemental questionnaire that "breaks out" significant workforces located in different counties. The Census Bureau also uses tax records related to employment (at the Federal level) but supplements those records with information from Social Security and income tax records to identify non-employer businesses, among other things (Walker 1997). BLS does not cover non-employers. Census updates its information on physical location of business activities more or less continuously, with an annual survey to establish multi-unit businesses. BLS updates its information on a rotating triennial cycle.

Why spend a lot to maintain two business registries? Redundancy. A correct description of business activities is the basis for industrial classification. For smaller single unit-establishments the Census updates industry classification with every economic census, every 5 years (Census 2000). In years ending in 2,7, and 12 information on one-third of the business population is collected through two different channels. Over the last five years the two agencies have been comparing industry codes to identify and understand discrepancies. (This process has not yet been completed, and we should learn a good deal about the quality of both lists from this evaluative activity.)

Even more interesting is the fact that Census bases its disaggregation of company activity on physical location. Every establishment has a permanent plant number. BLS bases its disaggregation of company activity on the basis of workforce. If the company changes legal form, but continues to employ the same workforce, the identity of the workforce will transcend the legal organization. This leads to the fascinating possibility of following work groups through different legal structures and possibly different locations in a panel of administrative records that BLS has compiled and will continue to develop (Spletzer 2000) .

Redundancy can reduce some errors. It also facilitates answering questions related to differences in the methodology used in the two business registers. Where there is overlap, as in classification, the redundancy can validate common data. Where there are differences we can gain insight into the complex process of evolution of business activity.

4 What do we know about survey errors?

4.1 Errors in orbit – Changing measurements for a changing economy

The greatest challenge to economic statistics comes from continuing change in the anatomy of the economy. The entities that make decisions change; products in the market change; markets change as regulations, communication, and institutions change. Measuring the economy is like hitting mutating birds in flight.

While the economy changes, the science of measurement also changes. Computer-assisted interviewing was the revolution of the 1980's. Cognitive methods applied to large surveys was a new force in the 1990's (Ware-Martin 1999). Now methods of data collection are challenged by web-based collection, a limited ability to access administrative records, and the possibility of dispensing with fixed forms in favor of data collection instruments that are tailored to each responding unit. How can we apply new technology to measuring the evolving economy? We already do so.

Many forward-looking innovations in measurement have been adopted by Federal Statistical agencies in the past decade. New indices quantify the national product; government investments in the economy have been given prominence in national accounts; new understanding of the performance of computers and software has been quantified. A stellar effort has been made to capture the importance of e-commerce to the economy. The innovation that I will focus on here is measuring how companies use computer networks for electronic contracting, control, purchasing, and sales.

The Bureau of the Census supplemented the *Annual Survey of Manufactures (ASM)* with a useful request for information about the use of computer networking in business in 1999. Just under fifty questions on two pages elicited a snapshot of many uses of networking software in internal and external business activities (Form *MA-1000(EC)*).

The supplement created design problems for the Bureau. The sample was set by the ASM, other difficult aspects of the design were resolved in an innovative and scientific way. The Bureau had already experimented with electronic submission of responses and wished to test that mode in a production survey. To do this the Bureau used the networking supplement to ASM. Establishments in the ASM sample were randomly assigned to two treatments: (I:CONTROL) a mail-out form and a URL address for voluntary electronic submission of responses and (II:EXP'T) a mail out letter with a URL address both for viewing the form and electronic submission of responses.[7] Field procedure called for 30-day follow up of non-responding units. All non-respondents received the paper form. Thus non-responding entities default to treatment (I) after 30 days.

A preliminary report on the experiment reveals two facts:
• The response rate is roughly five percent higher for group (I), the control group who could use paper and pencil for response.
• Most establishments in both groups responded on paper forms – 51% of group (II) and 89% of group (I). (See Dodds 2001.)

A superficial view of these findings is "If it ain't broke, don't fix it". That is, electronic reporting poses staggering difficulties for establishments. This view would be wrong on three counts:
• A trial lasting only 30 days increased electronic response more than four times, from 11% to 49%.
• The experiment withheld important information from part of the experimental group (II) for the first 30 days. Until they accessed the web, group (II) could not assess the burden of answering supplemental questions. This could clearly lead some to reject the questionnaire altogether, increasing non-response. Willingness of non-responding establishments to use electronic reporting

[7] The letter indicated that after 30 days a paper form would be mailed.

can not be inferred from the experiment.
• Randomized experiments are the most efficient tool for learning about the efficacy of survey methodologies.

In the long run, experimentation and evaluation are essential to scientific knowledge (Box and Tiao 1999). Experimentation exhorts us to: Ask new questions, implement untried modes for response, and investigate alternative channels to supply information required by respondents. When methodological initiatives take place in an experiment, we can be certain of one of three outcomes:

No difference in treatments at conventional levels for tests. The new methodology doesn't look different than the status quo. This "failed" experiment means we can pool all the data into one estimate without misgivings.

Experimental treatment less satisfactory than status quo. The second outcome, which is certain to happen often, is that a new technique is less effective, in terms of total survey error, than the status quo. This negative outcome yields vital information on how the production of surveys affects bias and variance of surveys. We do *learn* from negative findings.

Experimental treatment more satisfactory than status quo. The hoped for outcome of every experiment is that the technique on trial overwhelms status quo methods. Then we have the "winner's curse" – a random part of our data, less than the full sample, gives us estimates that have more variance and a different level than the status quo. Temporarily, we are more uncertain about the position of the economy. The compensation for this uncertainty is perspective on the meaning of past estimates, and knowledge that we can use to create future estimates with less mean square error, or equal error at less cost.

What is the value of the Bureau's experiment with electronic response over the web? The Bureau of the Census initiative on use of networks is forward-looking and methodologically appropriate. It simultaneously offers new information needed to understand our changing economy (http://www.census.gov/eos/www/ebusiness614.htm) while it tests survey methodology. Ultimately knowledge from tests of methodology increases the accuracy and timeliness of data capture and reduces respondent burdens. The bottom line is clear: planning, fielding, and interpreting experiments should be a day-to-day activity in all statistical agencies.

4.2 Outsourcing: Instructive anecdotes

Much has been made of businesses "outsourcing" some part of their operations. The term is shop-worn, but the impacts on measurement are barely understood. Firms contract to have part, even all, of their workforce employed by other entities. Firms contract to have essential software applications written by contractors. Firms even contract to have other firms manage their inventory. The results make odd statistics:

• Entities in the transportation industry manage inventories of inputs and product for manufacturing firms. They initiate orders, legally hold the inventory, and use their logistical expertise to minimize costs of holding inventory.
• Construction entities erect facilities worth hundreds of millions of dollars without employees. The

employees are all hired by their other business activities.
• Headquarters produce no sales, yet employ labor, commit the company to large electronic networks, and mastermind change in the structure and investments of the enterprise.

Why do we care?
A. We can not trace producing entities by their employees.
B. We can not assume that service industries hold no inventories.
C. We can not rely on complete reporting from a single form designed to fit a specific industry class.
If these conclusions are correct, major redesigns will be needed in data collection on businesses.

Do we ask the right questions?

When the Bureau of the Census extended measurement of the use of computer networks to service industries, it discovered that some companies providing transportation services were managing inventories for their customers. The transport company had better understanding of the logistics required to prevent stock-outs than the establishment producing the goods. By empowering the transport company to hold inventory the contractee was relieved of a major headache – but ownership of inventory was shifted from goods-producers to service industries. The problem facing the Bureau in its surveys and *2002 Economic Censuses* is how to assure that significant inventories in service industries are measured, without burdening small establishments (copy shops) whose inventory (paper) is neither particularly variable nor important, and may not need to be counted.

In the *1997 Economic Censuses* nearly 500 different forms were prepared and made available over the web. Key differences in forms related to thousands of products and inputs, many of which are idiosyncratic to a particular industry class or product group. Responding establishments were instructed to request forms that covered activities outside of their pre-coded NAICS classifications. This "demand feeding" for appropriate forms helped highly motivated firms supply complete information. However, I am not aware of an evaluation. Do some, or most establishments, with activities in several industries complete all the forms implied? What happens to data quality when relevant forms are not filled out? These are questions that may already be answered – If so, the conclusions need to be published. If not, evaluation is in order.

5 Putting it all together

Theory, redundancy, evaluation, and knowledge of error are required to produce high quality estimates. More conceptual support for measurement, including accounting and legal knowledge, will increase accuracy of responses.

" Measure twice, cut once" applies to economic statistics – More redundancy will lead to a better understanding of non-sampling errors and the ability to incorporate that understanding in economic models. Although policy-makers detest two estimates of the same phenomenon, statisticians need to enhance redundancy and distill knowledge about noise in the measurement system from differences in estimates.

Experiments play a key role in the design of data collection procedures. Use them frequently in most data collections (even in administrative record systems). No status quo methodology for collecting

information is so accurate to be exempt from experiments . Alternative methods may be superior. In many cases, forms and questions, mechanisms for selecting respondents, instructions to assist respondents, and records of the processes that generate responses need intensive evaluation.

Most importantly, evaluation needs to communicated to professionals outside of statistical agencies. This is tricky, because confidentiality is assured to data suppliers. Nonetheless, publication of findings is essential. As Churchill commented: "Those who don't read history are doomed to repeat it." The future statistics and economics profession needs have access to published history of evaluating collection of data from businesses and organizations.

References

Box, George E. P. and George C. Tiao. 1972. *An Introduction to Bayesian Statistics*. New York: Wiley, Ch. 1.

Bureau of the Census. 2000. *History of the 1997 Economic Censuses*. (POL00-HEC).

Cantwell, Patrick and Jock Black. 1998. Redesigning the monthly surveys of retail and wholesale trade for the Year 2000. *Proceedings of the Survey Research Section, American Statistical Association*, 481-486.

Cantwell, Patrick and Carol Caldwell. 1998. Examining the revisions in *Monthly Retail* and *Wholesale Trade Surveys* under a rotating panel design. *J. of Official Statistics,* 14(1) 47-59..

Dodds, Judy. 2001, draft. Response to the 1999 ASM computer network use supplement. Washington DC: Bureau of the Census.

Manzon, Gil B. and George Plesko. 2001, draft. The relation between financial and tax reporting measures of income. Cambridge MA: Sloan School of Management. 49 pp.

Petrick, Kenneth A. April 2001. Comparing NIPA profits with S&P 500 profits. *Survey of Current Business*, 16-20.

Shearson, Michael A. and Tracy Farmer-Farmer. 1998. Quality of the BLS establishment list as a frame. *Proceedings of the Survey Research Section, American Statistical Association* (Anaheim).

Spletzer, James. 2000. The contribution of establishment births and deaths to employment growth. *J. Business and Economic Statistics* 18:113-26.

Wolter, Kirk and Nash Monsour. 1986. Conclusions from Economic Censuses Evaluation Studies. *Annual Research Conference, 1986*. Bureau of the Census. 41-53

Walker, Edward. 1997, draft. The Census Bureau's business register: Basic features and quality issues.

Ware-Martin, A. (1999): "Introducing and Implementing cognitive Techniques at the Energy Information Administration," in *Statistical Policy Working Policy No. 28: A seminar on Interagency Coordination and Cooperation*, ed. by Office of Management and Budget/ Office of Information and Regulatory Affairs/ Statistical Policy Office, Washington DC: Executive Office of the President of the Unite States, 66-82 (Part 1 of 2).