INTEGRATING SAMPLE DESIGNS FOR ENVIRONMENTAL INDUSTRY AND OCCUPATIONAL EMPLOYMENT STATISTICS SURVEYS October 2010

Shail Butani, Dave Piccone, and Edwin Robison, U.S. Bureau of Labor Statistics

Abstract

Keywords: Integrating sample designs, sample rotation, sample overlap within and between surveys, stratification

In fiscal year 2010 budget, the U.S. Congress funded U.S. Bureau of Labor Statistics to develop measures of jobs associated with environmental activity, also known as "green jobs." Two separate measures of environmental activity are desired—employment by industry; and employment and wages by occupations. In this paper, we discuss the challenges associated with integrating the sample design for a new environmental industry of employment called the Green Goods and Services (GGS) Survey with the existing Occupational Employment Statistics (OES) Survey. Both the GGS and OES Surveys call for producing very detailed estimates at various levels of industry and geography that are vastly different for each survey. Statistical issues such as level of stratification, sample rotation, amount of sample overlap within each survey and between the two surveys are discussed.

Introduction

The 2010 Congressional Appropriation tasks the U.S. Bureau of Labor Statistics (BLS) with producing occupational employment and wage data on "green jobs". This initiative is for producing information on: (1) the number of and trend over time in green jobs, (2) the industrial, occupational, and geographic distribution of green jobs, and (3) the wages of the workers in these jobs (Federal Register, 2010).

The survey managers began the survey process with the daunting tasks of what concepts are to be measured to define green jobs, how are they to be measured, and what should be the scope of the survey. The general plan for conducting the GGS is given. Preliminary and subsequent modifications to the sample allocation procedures are outlined. Alternative sample selection procedures for GGS and a description of the OES sample design are given. Issues related to alignment of the GGS and OES samples are highlighted. Sample rotation options for the GGS sample are given. Some future research options are suggested.

Measurement Issues

Green jobs definition

After reviewing the extensive literature on green jobs, the survey managers realized there is no widely accepted available definition. The criteria they set for the definition are it should: be objective and empirically measureable; and use standard industrial and occupational classifications to provide comparability to other data.

BLS broadly defined green jobs as jobs involved in economic activities that help protect or restore the environment or conserve natural resources. The initial seven categories of green economic activity are: 1) renewable energy; 2) energy efficiency; 3) greenhouse gas reduction; 4) pollution reduction and clean-up; 5) recycling and waste reduction; 6) agricultural and natural resources conservation; and 7) education, compliance, public awareness, and training (Federal Register, March 2010).

BLS's definition of green jobs (Federal Register, September 2010) includes "jobs in businesses that produce goods or provide services that benefit the environment or conserve natural resources" as well as "jobs in which workers' duties involve making their establishment's production processes more environmentally friendly or use fewer natural resources."

Concepts

A review of the literature and talking to staff at Statistics Canada indicated that measurement of number of people employed in green jobs would at best be problematic since most workers perform both green and non-green activities. Thus, a decision was made that proportion of revenues from green activity would serve as a proxy for proportion of employment in green activity (Statistics Canada, 2004 and 2000).

Scope of the Survey

The GGS survey covers the private sector, local government, state government and the federal government in all 50 states and the District of Columbia. After extensive discussions with the users and other statistical agencies, a set of industries at the 6-digit North American Industrial Classification System (NAICS) level was determined to be within the scope of the GGS Survey. It is worth noting that NAICS is assigned according to the primary activity that generates the most revenue for an establishment. The industrial scope was an ever revolving process that took up a considerable amount of time. Initially about half (556) of the 6-digit NAICS industries covering about 45% of all employment were identified as potentially having green activity. Results presented in this paper were based on those 556 industries. This has since been reduced to 333 6-digit NAICS industries covering about 20% of all employment.

BLS presented its approach to measuring green jobs and the proposed definition of green jobs and the scope, including the list of 6-digit industries for the GGS Survey, in the March 16, 2010, Federal Register. The measurement approach includes two types of surveys: one on jobs related to producing green goods and services (GGS), and one on jobs related to using environmentally friendly production processes and practices. This paper is about integrating the samples for GGS with the existing OES sample.

Data Collection Issues

BLS initiated a research project to understand the collection environment and learn what information establishments have available that would help BLS collect data on green goods and services industry employment. The plan included: 1) conducting cognitive interviews to better understand the collection environment; 2) testing multiple variations on a form during this research project; 3) testing the form variations on panels of respondents; 4) testing non-response prompting and edit reconciliation processes; and

also 5) conducting follow-up interviews to contact establishments that responded and establishments that did not respond to the form during panel testing to ask about the form, difficulty in completion, respondents' understanding of the questions to assess response error, and reasons for non-response.

The primary purpose of this research is neither to finalize a definition of the green goods and services sector nor to determine what defines the green goods and services sector. Rather, the focus is on learning what collectable information firms have available about their products, services, and other items that might be used to collect data on this sector.

Sample Design Requirements

Develop a new GGS Survey that would measure employment by industry; the measurement of green activity would be based on receipts or revenues. For example, if an establishment has 10 percent of its revenue coming from green activity, then the assumption is 10 percent of the establishment employment is related to green activity; this is not a measure of specific jobs or persons involved in green activity.

The sample size for GGS is about 120,000 establishments drawn annually from the Bureau's Quarterly Census of Employment and Wages file.

Estimates of employment by occupation would be derived by linking data from GGS to the existing OES Survey. This is the reason to integrate the GGS and OES samples.

Estimates are desired at the: State/ 2-digit NAICS level; top-side State level (i.e., across all industries); National/4-digit NAICS; and top-side National level. Requirements evolved to include 1) some minimal publishablility for all 6-digit industries at the national level and 2) a breakout of some 4-digit NAICS into 6-digit NAICS for selected industries of particular interest. A special ANAICS ("allocation" NAICS) code was created mostly of 4-digit NAICS codes but with expansion to up to the 6^{th} digit for the particular industries of interest; also limited collapsing to 3-digit NAICS. Initial reliability criteria called for the same level of reliability for all states, but national data needs made that impractical.

Although not strictly a design requirement, research assumed a general scheme of sampling Probability Proportional to Estimated Size (PPES). An establishment's size was defined as the maximum employment on the frame over the last 12 months of available data. If noncertainty establishment A has twice the employment of noncertainty establishment B then it will be assigned twice the probability of selection, provided it is in the same stratum. The largest units in this type of sampling will necessarily be sampled with probability 1.000, or with certainty.

GGS Allocation Procedures

About 100,000 establishments are to be allocated to the private sector. We started the sample design process by developing GGS as an independent sample without any constraints from the OES Survey. There are many sampling unit and data collection issues pertaining to the government sample. It was therefore decided, at first, we'd concentrate on allocation and selection of establishments in the private sector. In the preliminary allocation, we kept aside a sample of about 20,000 units for certainty and local, state, and Federal Government samples.

In a first test allocation we set a state minimum sample of 1,500 proportionally allocated by employment to 2-digit NAICS. Nationally, a minimum of 40 establishments for each 6-digit NAICS and the remainder allocated proportionally to employment; the sample size of 40 with a response rate of 75 percent yields an effective sample size of 30 establishments. This simple preliminary allocation overemphasized industries of limited interest that had large employment; for example, restaurants. It also gave about equal sample to all states because of the 1,500 minimum. As a result, the reliability of the national estimates was comprised.

In the second test allocation, industries were grouped at differing NAICS levels of detail to provide proper balance needed for analysis (ANAICS).

- Kept national allocation of certain 6-digit NAICS for selected industries known to have strong green activity; the most common NAICS industry level was 4-digit or 5-digit; and three 3-digit NAICS for industries with large employment and of limited interest (e.g., food services and drinking places, NAICS 722).
- Reduced the state minimum allocation to 1,000 establishments.
- Used power allocation (square root of employment) at state level to temper the emphasis on larger 2-digit NAICS; similarly used power allocation at the national level to temper emphasis on larger defined industries (Lawley 2007).
- Limited the sample size for any defined industry to 1% of the private sample when summed across states then reallocated to obtain state minimum allocations of 1,000.

Several allocations were made for the second allocation, and then reconciled to a single allocation of about 100,000 private establishments: state minimum power allocations of 1000 establishments; minimum national 6-digit NAICS allocations; and a national power allocation of about 80,000 establishments. A given sample size would have been assigned by several allocations in the process. The allocations were "reconciled" by choosing the largest probability of selection for each establishment. Procedures were also needed to prevent the sample size for industries from exceeding the allowed 1% of the total sample.

The changes for the second test allocation yielded an allocation that was acceptable to the program managers. It allocated more sample to the larger states than to smaller states. At the same time, it limited the sample size for some of the industries with large employment but of limited interests; for example, restaurants, and colleges and universities. Attached is a table of sample size by 2-digit industries for the second allocation.

A test GGS sample was drawn using second allocation criteria, and the selection was independent of OES sampling. The size of establishment i on the frame (the maximum employment value from the last 12 reported months) and its state and industry were used in the reconciliation process to assign the establishment a probability of selection p_i. A fairly straightforward unequal probability sample was selected by generating a random numbers rn_i between 0 and 1 and selecting establishment i when rn_i was less than or equal to p_i. It is interesting to note that strata per se were not formed. The sampling was not controlled to exactly equal the desired state and national allocations, but it was verified that the results of the sampling process were within statistical tolerance.

Integration of GGS with OES Sample

OES Sample Design

One of the major requirements for GGS is to link to the existing OES Survey sample because the intention is to profile occupational staffing patterns and their associated wages for establishments with green activity to those with non-green activity. Thus, the goal is to maximize the sample overlap between the two surveys.

OES surveys about 1.2 million establishments over a 3-year period with six semi-annual samples of about 200,000 establishments. A major constraint is that no establishment (including certainty units) is surveyed more than once during a 3-year period. The estimates are produced by combining 3 years of data. Stratification is State/metro area (multi-state metro areas are split to the various states) by 4-5 digit NAICS (different than GGS definition). A Neyman power allocation method is used (Lawley 2007). Each sample is selected using probability proportional to modified employment size. That is, all non-certainty units within each state and size class are assigned the mean employment value of all units in that state and size class in order to add some stability to employment data. The wage data are updated using the data from the National Compensation Survey which is also conducted by BLS.

Alternative Sample Selection Procedures

The independent GGS sample that was selected following the second test allocation was matched to the OES sample for the six panels in 2007-2009. Overall there was an overlap of about 50 percent for the number of establishments and 80 percent for employment since establishments with large employment are selected with higher probabilities. A major drawback to independent sample selection is that linking the probabilities of selection between two surveys becomes complicated. The linking of probabilities of selection is important to enable valid analysis.

The second test GGS allocations were independently derived without reference to the OES survey. The allocations for GGS were closely compared to existing OES samples and comparisons made to determine 1) natural overlap and 2) the potential for subsampling GGS from OES. Since many states/industries would have GGS sample needs exceeding what is available in OES, we wanted to determine the extent that the OES sample would need to be expanded or augmented to allow GGS subsampling from OES.

For each sampled 2007-2009 OES establishment i, the OES probability p_{iOES} and the desired GGS probability P_{iGGS} were known. A random process was used to subsample units for GGS when P_{iGGS} was less than p_{iOES}. (subsampling probability P_{iGGS} / p_{iOES}.) All OES units were selected where P_{iGGS} was greater than or equal to than p_{iOES}. The overall shortfall of this subsample when compared to GGS allocations is an approximation of the minimum number of extra establishments that would needed to be added to OES to enable GGS subsampling.

The results of comparing the second test allocation to OES are also shown in the attached table. The OES sample could fulfill 80 percent of the GGS sample allocation needs in terms of establishments and 90 percent in terms of employment. There were two industries where the OES sample was particularly weak; that is, it lacked a sufficient

number of establishments for GGS. These two industries are: Agriculture, Forestry, Fishing and Hunting where only 19 percent of GGS establishments could be covered; and Finance and Insurance where only 37 percent of GGS establishments could be covered. The results across states were very similar to the national level.

2-Digit NAICS	Sector Name	GGS Alloc	GGS Sub-Sample of OES	Diff	Pct
11	Agriculture, Forestry, Fishing and Hunting	5,810	1,080	4,730	19%
22	Mining, Quarrying, and Oil and Gas Extraction	2,327	1,927	400	83%
23	Construction	9,231	8,298	933	90%
31	Manufacturing	6,465	5,541	924	86%
32	Manufacturing	8,819	7,286	1,533	83%
33	Manufacturing	10,333	8,583	1,750	83%
42	Wholesale Trade	11,126	10,171	955	91%
44	Retail Trade	8,151	7,389	762	91%
45	Retail Trade	6,009	5,495	514	91%
48	Transportation and Warehousing	5,502	4,654	848	85%
49	Transportation and Warehousing	3,025	2,691	334	89%
51	Information	3,675	3,340	335	91%
52	Finance and Insurance	851	315	536	37%
53	Real Estate and Rental and Leasing	1,394	1,020	374	73%
54	Professional, Scientific, and Technical Services	6,958	6,152	806	88%
55	Management of Companies and Enterprises	1,894	1,801	93	95%
56	Administrative and Support and Waste Management and Remediation Services	3,796	3,222	574	85%
61	Educational Services	1,745	1,610	135	92%
71	Arts, Entertainment, and Recreation	1,149	882	267	77%
72	Accommodation and Food Services	1,638	1,532	106	94%
81	Other Services (except Federal, State, and Local Government)	4,097	3,707	390	90%
	Totals:	103,993	86,696	17,297	83%

Panel Rotation

Several alternative panel rotation schemes for GGS were considered. The one thought most likely to be implemented is the one with three panels of about 40,000 establishments each (including both private and government units). Certainty in-scope establishments will be included in every panel. Each year 3 panels will be surveyed. After start-up of the GGS, one panel will be dropped and a new panel added each year. Noncertainty establishments will generally be surveyed three years in a row, and then will be dropped.

Plans for Additional Research

Additional GGS allocations are being tested with the reduced list of 332 industries. Overlap with and the potential for subsampling from OES will be re-analyzed. After this comparison, simplifications of the design will be considered. For example, it may be possible to simplify the design and explicitly stratify the frame. Then research will concentrate on adding sample to the OES survey and recomputing OES probabilities of selection so that GGS can be selected as a subsample from OES.

Test allocations used an establishment's maximum reported number of employees over the last 12 available months as the measure of size for power allocation and selection of samples. We are studying the impact of size changes over time in establishments, and how that affects variance. Indications are that we should increase the probabilities of selection of the smallest establishments, and future research on that issue is proposed. It is unknown to what extent newly formed establishments (births) will differ from existing establishments in terms of "greenness." Various options for sampling births can be considered for improving coverage and reducing bias in GGS.

The latest BLS information on "Green Jobs" can be found at http://www.bls.gov/green/ .

Acknowledgements

The authors thank Rick Clayton, George Stamas, and Marie Stetser, for their insightful comments leading to many refinements to the Green Goods and Services Survey sample design.

The views expressed are those of the authors and do not represent official positions of BLS.

References

http://www.bls.gov/green/ .

Federal Register, Vol. 75, No. 50, March 16, 2010, pp. 12571-12573.

Federal Register, Vo. 75, No. 182, September 21, 2010, pp. 57506-57514.

Lawley, Ernest, Stetser, Marie, and Valaitis, Eduardas. (2007) *Alternative Allocation Designs for a Highly Stratified Establishment Survey*. 2007 Joint Statistical Meetings.

Statistics Canada. *Environment and Industry Sector: 2002 revised and 2004.* http://www.statcan.gc.ca/pub/16f0008x/16f0008x2007001-eng.pdf

Statistics Canada. *Measuring employment in the environmental industry: 1998 and 2000.* http://www.statcan.gc.ca/pub/16-001-m/16-001-m2004001-eng.htm