Independent Statistics & Analysis
U.S. Energy Information
Administration

Assessment of consumption and expenditure data collected from energy suppliers against bill data obtained from interviewed households: Case study with 2009 RECS

February 2013

Independent Statistics & Analysis
www.eia.gov

U.S. Department of Energy
Washington, DC 20585

This report was prepared by the U.S. Energy Information Administration (EIA), the statistical and analytical agency within the U.S. Department of Energy. By law, EIA's data, analyses, and forecasts are independent of approval by any other officer or employee of the United States Government. The views in this report therefore should not be construed as representing those of the Department of Energy or other Federal agencies.

Table of Contents

Tables

Figures

Introduction

The Residential Energy Consumption Survey (RECS) is a national area-probability sample survey that statistically selects housing units across the U.S. to collect energy-related data for occupied primary housing units. This survey mainly consists of two parts: the RECS Household Survey (RECS-HS) and the Energy Supplier Survey (ESS). In RECS-HS, the respondents are to provide information about their household characteristics, energy-using equipments, fuels used, and other information related to energy use. In ESS, the energy suppliers are required to provide twenty months of energy consumption and expenditure data for the RECS-HS sample households.

In RECS-HS, the households are asked to submit their energy supplier information via utility billing statements. Beginning with the 2005 RECS data collection cycle, the interviewers used portable devices to scan the respondents' utility bills in order to gather the supplier names and account numbers more easily and accurately. The scanned bills actually contain more information than supplier names and account numbers. They contain the energy consumption and expenditures of at least the current service period, and we utilize this information to examine the ESS collected data in terms of numerical accuracy. By comparing these sources, we hope to learn more about any limitations in the data that we collect, which we can then attempt to address. As such, this limited empirical study is an example of the research that EIA conducts to evaluate and subsequently improve on the quality of data that EIA collects.

Research design

Population of interest

The 2009 RECS-HS has 12,083 cases. In this analysis, we compare the ESS collected data to the RECS-HS scanned electricity bill data, as electricity is the most commonly and widely used energy source. The population of interest is the 2009 RECS-HS interviewed households with scanned electricity bills, for which energy suppliers provided consumption and expenditure data. The size of this analysis population is 6,150. Note that the RECS-HS sampling weights have no bearing in this analysis and that we are not inferring any findings to a wider population.

Data

Since the analysis population is too large to examine in its entirety, we limited ourselves to selecting one household per supplier. There are 408 unique electricity suppliers found in the 6,150 scanned bills, and we randomly selected one household from each supplier. This approach is based on our assumption that the data quality does not vary much within each supplier, while the variation could be substantial across suppliers. Although this assumption may not actually be the case, we decided to use this simple approach for our initial study. There are other approaches that could have been used, such as acceptance sampling or probability proportional to size sub-sampling, but those would have required considerably more resources since many more cases would be needed.

In the analysis population, the maximum number of households linked to any one supplier is 325. The minimum was one, and there were 46 such suppliers. The mean number of households per supplier is about 15; the median is 5.

ESS coverage period

The ESS data we examined contained raw data, i.e., data submitted directly from the supplier, on household electricity consumption and expenditures from September 2008 to April 2010. The reference period of the 2009 RECS-HS is from January to December 2009.

Data extraction from scanned electricity bills

An electricity bill for each selected sample case was extracted from our image data files and its content was manually processed to produce consumption and expenditure values consistent with the ESS data requirements. This was a time-consuming process because of the irregularity of the bills and the specificity of numerical information we needed to collect (i.e., electricity consumption in kWh and only its cost and tax).

Matching ESS data and scanned bill data at one billing month

Since the energy bill was scanned by the interviewer at the time of the RECS-HS interview, the bill usually contains only the consumption and expenditure information, if any, of some billing period preceding the interview time, which fell sometime between February and August 2010. The ESS collected data, which are expected to span the period from September 2008 to April 2010, and the RECS-HS scanned bill data can be matched only at one particular billing period within the overlapping months (February – April 2010). We might expect that if any systematic data problems exist in the ESS collected data they are likely to prevail in all months within a given supplier. This comparison will give some, if not complete, insight into the quality of ESS data relative to bill data.

Problems comparing ESS collected data with scanned bill data

Some ESS collected data and RECS-HS scanned bill data could not be numerically compared. Here is a list of problems encountered in the RECS-HS scanned bill data.

Problems in RECS-HS scanned bill data:

- Partial or insufficient information – some bills did not have all of the needed data
- Scanning problems (bad, incomplete, or incorrect scanning)
- Not an electricity bill (e.g. a gas bill)
- No breakdowns of charges
- Wrong scanned bill for the interviewed household (e.g. wrong address, wrong account, etc.)
- Bill information outside the 20-month ESS data coverage period (unable to match with ESS collected data)

And, the following non-numerical problems were found in the ESS collected data.

Problems in ESS collected data:

- Not correctly reporting the supply or/and delivery information (B = Both supply and delivery, S = Supply only, and D = Delivery only)
- Not covering the months when RECS-HS was administered (unable to match with scanned bill data)
- Clerical errors on data submissions – typo, incorrect unit of measurement, decimal place problem, and so on
- Non-matching addresses with the interviewed household

Due to these data problems, of the 408 sampled cases, 91 cases could not be numerically compared with respect to the consumption values, while 65 cases could not be numerically compared with respect to the expenditure values.

Description of numerical differences in ESS collected data from scanned bill data

Percent with zero difference

Consumption

Examining the ESS collected data and the RECS-HS scanned bill data with respect to electricity consumption; we found that there were 317 (78%) numerically comparable cases and 91 (22%) non-comparable cases out of the 408 sampled cases. (See Table 1.)

Table 1. Number of cases with numerically comparable consumption values

	Count (N)	Percent (%)
Total Sampled Cases	408	100
Comparable Cases	317	78
Non-Comparable Cases	91	22

The measure of difference is defined as: Difference = ESS value – bill value.

Of the 317 comparable cases, there are 305 cases (96%) with zero difference (or perfect agreement) and 12 cases (4%) with some difference in the electricity consumption values. (See Table 2.)

Table 2. Number of cases with zero difference in consumption values

	Count (N)	Percent (%)
Comparable Cases	317	100
With Zero Difference	305	96
With Some Difference	12	4

Expenditures

With respect to the electricity expenditures, there are 343 cases out of the 408 sampled cases (84%) that were numerically comparable between the ESS collected data and the RECS-HS scanned bill data; and 65 (16%) numerically non-comparable cases. (See Table 3.)

Table 3. Number of cases with numerically comparable expenditure values

	Count (N)	Percent (%)
Total Sampled Cases	408	100
Comparable Cases	343	84
Non-Comparable Cases	65	16

Of the 343 comparable cases, there are 272 cases (79%) with zero difference (or perfect agreement) and 71 cases (21%) with some difference in the electricity expenditure values. (See Table 4.)

Table 4. Number of cases with zero difference in expenditure values

	Count (N)	Percent (%)
Compared Cases	343	100
With Zero Difference	272	79
With Some Difference	71	21

The distribution of difference

Consumption

Table 5 shows some distribution statistics of consumption difference. We have computed the mean, minimum, median, and maximum. Since most of the cases have zero difference, the mean is close to zero at a positive value of 0.129. The difference is the ESS value minus the bill value so the positive mean value suggests the ESS value is larger than the bill value on average.

Table 5. Distribution statistics of consumption difference

	kWh
N	317
Mean	0.129
Minimum	-488.000
Median	0
Maximum	663.000

Expenditures

Table 6 shows some distribution statistics of expenditure difference; we have computed the mean, minimum, median, and maximum. Since most of the cases have zero difference, the mean is close to zero at a negative value of -1.599. Since the difference is the ESS value minus the bill value, the negative mean value suggests the ESS value is smaller than the bill value on average.

Table 6. Distribution statistics of expenditure difference

	$
N	343
Mean	-1.599
Minimum	-189.950
Median	0
Maximum	85.180

The distribution of non-zero difference

Consumption

In this section, we exclude the cases with zero difference in consumption values and repeat the analysis in the previous section. The goal is to examine non-zero differences in consumption values more closely in order to find any patterns in the consumption value differences.

First, note that there are only twelve non-zero-difference cases. As expected, the mean difference becomes much larger. (See Table 7.)

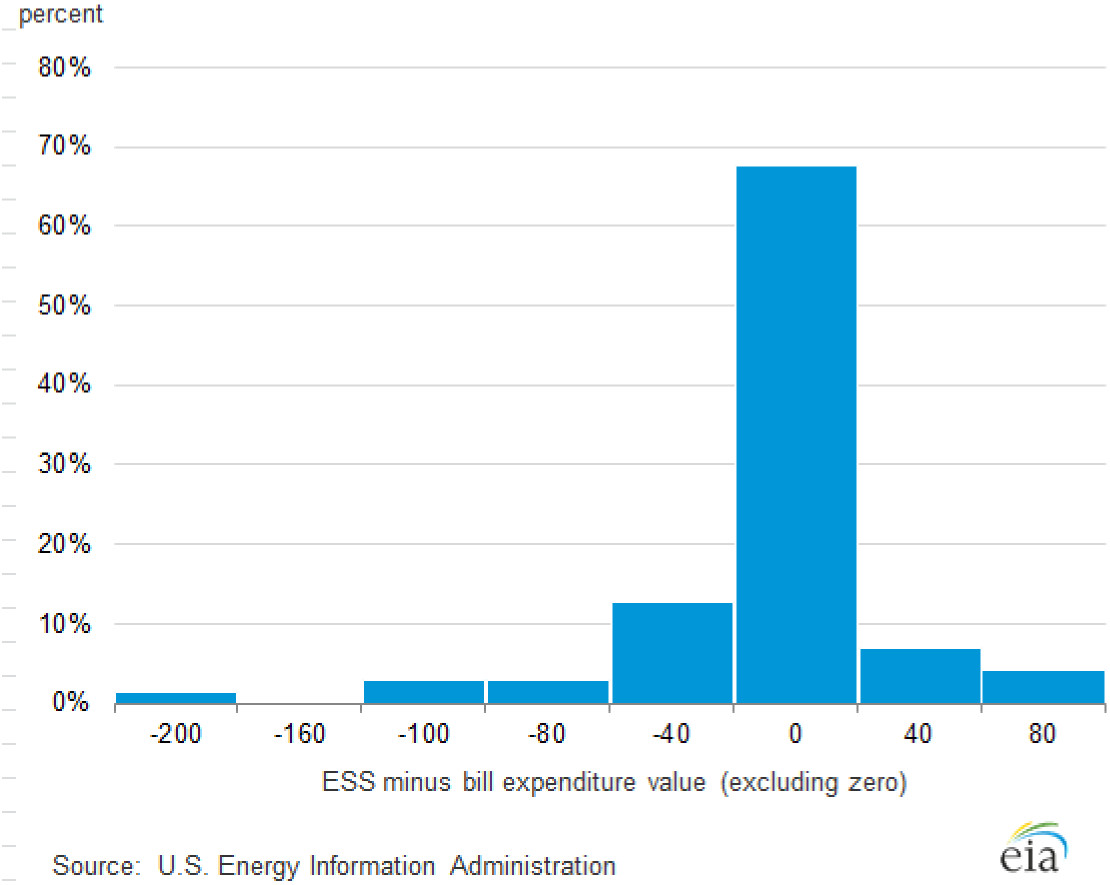
Table 7. Distribution statistics of consumption difference without zero-difference cases

	kWh
N	12
Mean	3.416
Minimum	-488.000
Median	4.500
Maximum	663.000

Expenditures

There are 71 cases that produce non-zero differences in expenditures between the ESS collected data and the RECS-HS bill data. The number is larger than for consumption (12 cases), which may not be surprising because expenditure is a less straightforward quantity than consumption. The distribution is skewed to the negative direction. (See Figure 1 and Table 8.)

Figure 1. Histogram of the non-zero difference between the ESS expenditure and the bill expenditure



Source: U.S. Energy Information Administration

Table 8. Distribution statistics of expenditure difference without zero-difference cases

	$
N	71
Mean	-7.725
Minimum	-189.950
Median	-0.840
Maximum	85.180

Characteristics of the cases with non-zero differences

Consumption

The differences in consumption ranges from the minimum value of -488 to the maximum value of 663, as shown in Table 9. In terms of their relationship to the service state (just one of account characteristics), we point out only that the non-zero differences did not concentrate in one or a few service states. In fact, the service states turned out to be all unique except for California, which appear twice. (Our sample includes all suppliers in the analysis population but each supplier is represented by only one account. A larger state or a state with more suppliers has a higher chance of being in our sample.)

Table 9. List of the non-zero differences in consumption and their service state

Difference in Consumption Values (kWh)	Service State
-488	NY
-387	SC
-212	SD
-200	TN
-14	VA
-13	CO
4	CA
50	MA
51	AL
106	KY
481	CA
663	NC

Expenditures

Next, we examine the 71 non-zero differences in expenditures and their service states. The count shows the frequency of non-zero differences in each service state. (See Table 10.)

Table 10. Frequency of the non-zero difference in expenditures in service state

Service State	Count
AL	2
AR	2
CA	5
CO	1
FL	5
GA	3
IA	3
IN	1
KS	2
KY	2
LA	2
MA	1
MN	4
MO	7
MT	1
NC	1
ND	1
NJ	1
NM	1
NV	1
NY	1
OR	1
PA	3
SC	1
SD	2
TN	2
TX	8
UT	2
VA	1
WA	3
WI	1
Total	71

Texas has the most cases (eight), followed by Missouri with seven. Table 11 shows the non-zero differences in those service states. Note again that each difference comes from a unique supplier.

Table 11. Non-zero differences in expenditure in Texas and Missouri

Service State	Difference in Expenditure Values ($)
TX	-18.93
	-15.17
	-9.65
	-2.58
	-0.95
	-0.68
	-0.65
	29.12
MO	-80.63
	-7.57
	-2.62
	-0.10
	1.33
	2.21
	7.00

Relationship between the consumption difference and the expenditure difference

Since the consumption data and the expenditure data were collected together (for the ESS collected data) or reported together (for the RECS-HS scanned bill data), occurrences or absences of any problems in the data may not be independent. Discrepancies can exist in both consumption and expenditures. Table 12 shows the pairs of non-zero consumption difference and non-zero expenditure difference. There are ten pairs. Recall that there were 12 non-zero expenditure differences and 71 non-zero consumption differences. When both consumption and expenditure ESS data depart from the RECS-HS bill data, the magnitude and sign of consumption difference appear to be related to those of expenditure differences. This may suggest that there exists a common mechanism producing both the non-zero consumption and expenditure differences.

Table 12. Pairs of non-zero consumption difference and non-zero expenditure difference

Difference in Consumption (kWh)	Difference in Expenditures ($)
-488	40.29
-387	-39.95
-212	-21.37
-14	5.77
4	0.23
50	10.41
51	12.20
106	11.44
481	85.18
663	78.63

Final remarks

As mentioned, extracting the right data from scanned bills was a time-consuming manual task, because different suppliers use different forms for their billing statements. Further, many expenditure values required derivation or interpretation. For example, a total utility charge may include expenditures that are not directly related to the consumption for the current month (e.g., a late fee).

In a sense, ESS requires of suppliers a similar burden of extracting and submitting only applicable data. However, the burden is distributed among suppliers, and each supplier is expected to have a common billing statement form or system for all its customers. It is clear that ESS is operationally a more reasonable method to collect consumption and expenditure data than manually extracting the data from bills scanned at interviewed households.

Given the current method of ESS, our research purpose was numerically to understand the ESS data quality, comparing the electricity consumption and expenditure data from ESS against those in the RECS-HS scanned bills. With the 2009 RECS, our analysis population consisted of 6,150 household electricity accounts, which provided electricity bill statements in the 2009 RECS-HS and for which ESS collected consumption and expenditure data. We randomly selected one account from each supplier or stratum. There were 408 electricity suppliers in the analysis population; thus, our sample size was 408.

We examined the proportion of accounts that we could numerically compare between the ESS collected data and the RECS-HS bill data. It turned out that the proportions were higher for expenditures than for consumption. However, of the numerically comparable accounts, the proportion of zero-numerical-difference cases was lower for expenditures than for consumption.

Examining the non-zero-difference cases, we observed no systematic patterns in consumption or expenditures, although the mean differences were not exactly zeros. Furthermore, since our comparison was possible only for one given month, the observed difference may not represent other months in the same account or in the same supplier. These two limitations (one month from one household per supplier) imply that we do not have a statistically valid sample from which we can generalize our results. Nevertheless, these findings provide important insight into the quality of the data that EIA collects.