Considering Alternative Outlier Detection Methods to Improve the RECS Square Footage Data Editing Process

Chrishelle Lawrence
U.S. Energy Information Administration, 1000 Independence Avenue SW, Washington, DC 20585

Abstract

Square footage is a critical variable in modeling the amount of end-use energy consumption of individual homes. The Residential Energy Consumption Survey (RECS) has traditionally used trained interviewers to measure homes and ask respondents to also provide a square footage estimate. In recent data collection cycles conducted via the web or mail, data editing for square footage has proven to be a challenging task. The current data editing process resolves to verify the accuracy of each respondent reported square footage that failed any outlier detection checks, which can be arduous if the number of edit failures is large. This research will consider alternative outlier detection schemes to determine the best approach for future RECS cycles.

Key Words: Data editing, outlier detection

1. Introduction

The Residential Energy Consumption Survey (RECS) has been conducted periodically since 1978, and it provides a wealth of information about energy usage in American households. A major component of the RECS is obtaining an accurate measure of the square footage of homes. Square footage is used in tabulations, public microdata files, and as input for energy end-use models. Traditionally, Computer-Assisted Personal Interviewing (CAPI) was used to conduct the survey, and interviewers were trained to properly measure square footage. The most recent RECS, conducted in 2015, used three survey modes: CAPI, web, and mail. In each mode, respondents are asked to provide an estimate of the square footage of their homes, but only CAPI respondents had interviewers to measure their homes. The interviewer measured square footage is considered the "gold-standard". Previous research has shown that respondents tend to provide smaller estimates than the interviewer measurements. As the RECS moves toward alternative data collection modes, the respondent reported square footage becomes vitally important, and this information needs to be of comparable quality to the interviewer measured estimate.

2. The Current Editing Process

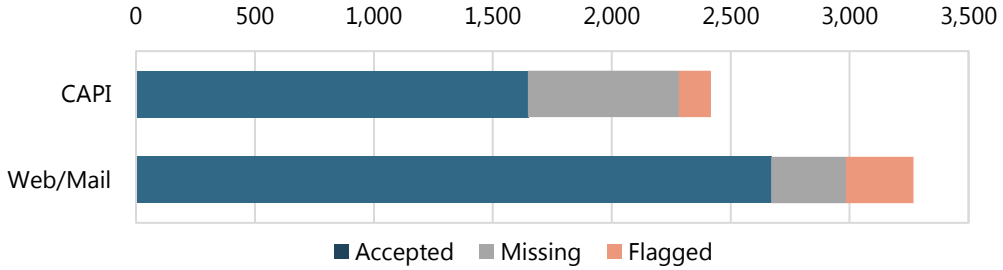
In the RECS, respondents were asked to first provide an estimate of square footage and then specify if that estimate includes their basement, attic, or attached garage. The square footage questions in the CAPI, web, and mail modes are shown in Appendix A.

Once all data were collected and processed, edit checks were performed to determine the quality of the reported square footage. If the reported square footage fell outside of a specified set of bounds, it was flagged for analyst review. Extreme outliers, housing units that were less than 100 ft^2 or greater than 5,000 ft^2 were automatically reviewed. Additionally, lower and upper bounds were set using the 5^{th} and 95^{th} percentiles of the square footage from the previous RECS cycle. A set of bounds were created for each

housing type: mobile homes, single-family detached homes, single-family attached homes, apartments in a building with two to four units, and apartments in a building with more than five units. For the 2015 RECS editing, percentiles from the 2009 RECS were used.

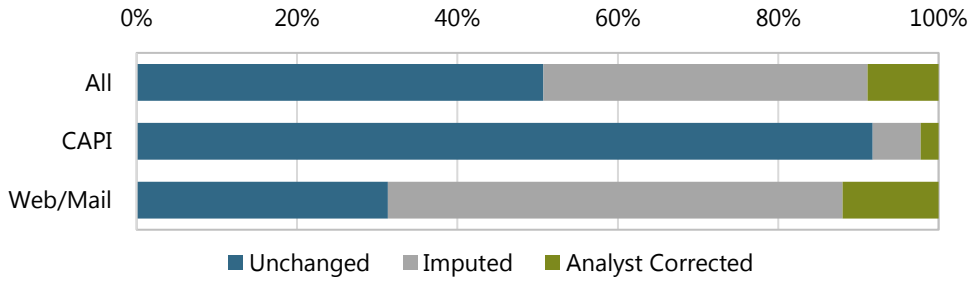
There were 5,686 RECS respondents. Of these, 950 respondents (16.7%) did not report square footage. Excluding these, 8.8% of reported square footage was flagged for further review. Figure 1 shows the frequency of reported square footage that was accepted, missing, or flagged for review by survey mode. Two-thirds of flagged square footage were from web and mail respondents.

Figure 1: Square Footage Status by Survey Mode



Household characteristics such as the inclusion of a basement, attic, or attached garage in the square footage can have a significant impact. During review, analysts used these household characteristics, administrative records, and geographic software to look at homes to verify estimates. After review, the square footage was left as is, sent to imputation, or corrected. If the square footage was considered unreasonable after review, it was sent to imputation. Corrected square footage was normally a typo where a zero should have been added or removed from the reported figure. For the 2015 RECS, half of the square footage remained unchanged after review. Analysts corrected 8.9% of the square footage, and the remaining were imputed. Figure 2 shows the distribution of the editing review results by mode. Most of the flagged reported square footage from CAPI respondents remained unchanged while over half of the square footage from web and mail respondents was sent to imputation. With competing work assignments, the review process took a team of analysts a few weeks to complete.

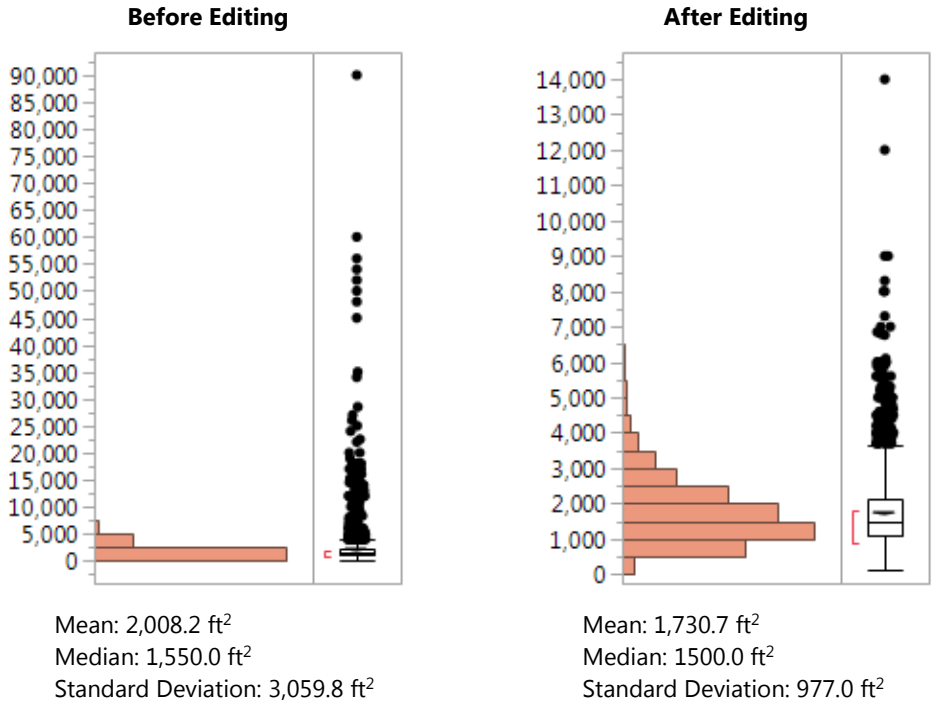
Figure 2: Distribution of Editing Results by Mode



How did the edit review process change the overall square footage estimates? Figure 3 shows the distribution of the estimated square footage before and after the editing process. Before editing, there were several estimates above 10,000 ft^2, and the standard deviation was quite large. The largest reported square footage was 90,000 ft^2. After editing, data quality greatly improved resulting in more reasonable square footage estimates. The

average square footage decreased by 277.5 ft^2, and there was 68% decrease in the standard deviation.

Figure 3: Square Footage Before and After Editing



3. Considering Alternative Outlier Detection Methods

Since the review process can be lengthy if there is a large amount of flagged estimates, alternative editing methods are considered. Instead of using the 5th and 95th percentiles of the estimated square footage for each housing type as the lower and upper bounds, even wider bounds that were 2, 2.5, and 3 standard deviations from the mean were used to flag square footage estimates. The mean from the previous survey cycle, RECS 2009, was used. Table 1 shows the editing bounds by housing type and method. Unlike the current editing process, many of the lower bounds from the alternative methods fall below zero. As a result, homes with very small estimated square footage would not be reviewed.

Table 1: Editing Bounds by Housing Type and Method

Housing Type	Original Method	2 Standard Deviations	2.5 Standard Deviations	3 Standard Deviations
Mobile home	(528, 1976)	(231, 1960)	(15, 2176)	(-200, 2392)
Single-family detached	(800, 5346)	(-217, 5612)	(-946, 6341)	(-1675, 7069)
Single-family attached	(716, 3664)	(-4, 3767)	(-475, 4238)	(-946, 4710)
Apartment (2-4)	(400, 2372)	(-114, 2346)	(-421, 2653)	(-729, 2961)
Apartment (5+)	(440, 1431)	(195, 1523)	(29, 1688)	(-136, 1854)

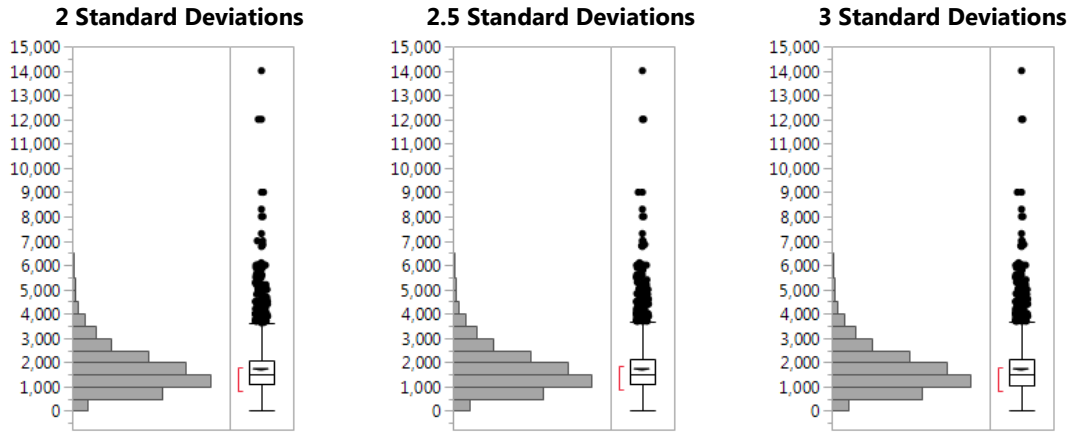
Since the bounds were much wider, the alternative methods flagged fewer cases. Table 2 shows the percentage of flagged square footage estimates by method. Using any of the three alternative methods reduced the number of flagged cases by 53% to 72%.

Table 2: Frequency and Percentage of Flagged Cases by Method

Method	Frequency	Percent
Original Method	418	8.8
2 Standard Deviations	198	4.2
2.5 Standard Deviations	137	2.9
3 Standard Deviations	118	2.5

Figure 4 shows the distribution of the reported square footage after the editing and imputation processes using the alternative bounds. The three distributions look quite similar, with similar means and standard deviations.

Figure 4: Square Footage Distributions for the Alternative Methods



Mean: 1,717.5 ft^2
Median: 1,500.0 ft^2
Standard Deviation: 993.2 ft^2

Mean: 1,714.7 ft^2
Median: 1,500.0 ft^2
Standard Deviation: 997.3 ft^2

Mean: 1,714.2 ft^2
Median: 1,500.0 ft^2
Standard Deviation: 998.1 ft^2

Table 3 shows the mean of the respondent reported square footage by housing unit type for each of the editing options. For single-family detached and single-family attached homes, the means from the three alternative bounds are identical.

Table 3: Mean Estimated Square Footage by Housing Type and Editing Method

Housing Type	Original Method	2 Standard Deviations	2.5 Standard Deviations	3 Standard Deviations
Mobile home	1,188.6	1,188.6	1,158.7	1,156.4
Single-family detached	2,011.7	2,001.3	2,001.3	2,001.3
Single-family attached	1,455.2	1,413.4	1,413.4	1,413.4
Apartment (2-4)	911.7	876.8	887.1	887.1
Apartment (5+)	872.3	864.5	849.9	846.3
All homes	**1,730.7**	**1,717.5**	**1,714.7**	**1,714.2**

Although the alternative methods yield similar mean square footage, their inability to capture extreme values of low square footage can prove to be problematic. RECS data are released at the housing unit level. Subsequently, the square footage should reflect the corresponding household characteristics, such as the number of bedrooms and bathrooms. For example, there is a single-family attached home with four bedrooms and three bathrooms with a reported square footage of 20 ft^2. After editing and imputation, the final value for this home was a much more reasonable 1,700 ft^2.

There were 31 housing units of all types where respondents reported a square footage of less than 100 ft^2. The lowest reported square footage estimate was only 2 ft^2 for a one bedroom one bathroom apartment in a building with five or more units. After editing and imputation, the updated values ranged from 150 ft^2 to 2,500 ft^2. The updated mean square footage of these units was 1,093.8 ft^2, which was a 3,000% increase from the previous mean of 35.9 ft^2.

4. Conclusion

More research should be conducted to strike a balance between minimizing the number of cases flagged for questionable square footage and maintaining a high level of data quality. The alternative methods yielded similar results, but they lacked the ability to detect extremely low square footage. It is not only important to review homes with low square footage to verify that the size of the home is consistent with other household characteristics, but to also ensure that uniquely small homes are not easily identified by data users.

While additional research is considered from the editing perspective, other survey improvements can aid in collecting better square footage estimates from respondents. CAPI and Web questionnaires can be updated to allow for immediate correction of extreme values by building edits directly into the questionnaire. When an unreasonable value is provided or entered for the estimated square footage, respondents would be asked to verify their input. Additionally, questions about the respondent's level of confidence and the source of their estimate can be incorporated into the questionnaire to aid in determining the quality of the estimate. If a respondent has no confidence in their answer and it's just a guess, then their response could automatically be sent to imputation without being flagged for editing at all. Also, editing analysts would be more confident in the reported square footage if they knew respondents consulted a reliable source for the estimate, such as a deed, lease, or real estate website. As the RECS moves forward in an increasingly complex survey environment, the square footage editing process should evolve as well to meet its needs.

Appendix A: Square Footage by Survey Mode

CAPI

To understand the usage of energy in your home, we need to know about its size and shape. About how many square feet is your home? Your best estimate will do.

_____

Does your estimate of square footage include your basement?

1. Yes
0. No

Does your estimate of square footage include your attic?

1. Yes
0. No

Does your estimate of square footage include your attached garage?

1. Yes
0. No

Web

About how many square feet is your home? Your best estimate is fine.

square feet

Back Next

Which of the following areas are included in your estimate of <*SQFTEST*> square feet? *Please select all that apply.*

☐ Basement
☐ Attic
☐ Attached garage

☐ I have at least one of these spaces but none are included in my estimate

☐ My home does not have any of these spaces

☐ Don't know

Back Next

Mail

15. About how many square feet is your home? Your best estimate is fine.

, square feet

16. Which of the following areas are included in your estimate of square footage in Question #15?

A. Basement

1 ☐ Yes
0 ☐ No
-4 ☐ Don't know
-3 ☐ Not applicable (my home does not have this space)

B. Attic

1 ☐ Yes
0 ☐ No
-4 ☐ Don't know
-3 ☐ Not applicable (my home does not have this space)

C. Attached garage

1 ☐ Yes
0 ☐ No
-4 ☐ Don't know
-3 ☐ Not applicable (my home does not have this space)