Exploratory Research on the Use of Google Earth to Create a Sampling Frame of Buildings

Katie Lewis

U.S. Energy Information Administration
1000 Independence Ave, SW, Washington, DC

Abstract

Since its inception, a majority of the sampling frame for the Commercial Buildings Energy Consumption Survey (CBECS), conducted by the U.S. Energy Information Administration (EIA), has been created or updated using traditional area listing. The frame is expensive to maintain because field workers travel to selected area segments and list all eligible commercial buildings in those segments. For each building, field workers record information for locating the building at the interview phase and variables needed for sampling.

This paper discusses exploratory research on one potentially less expensive alternative method for creating the frame: the use of Google Earth™ software and data. In this project, EIA selected a sample of segments listed in the Fall of 2012, then a sample of listed buildings within those segments, and attempted to reproduce a subset of the variables the field workers recorded for those buildings. This paper discusses the methods, comparisons of the frame variable values, the nature of the Google Earth images (e.g., their ages), and the difficulties encountered, before closing with discussion and recommendations for further research.

Background

The Commercial Buildings Energy Consumption Survey (CBECS) is a national sample survey of buildings over 1,000 square feet in which at least half of the floorspace is used for a commercial purpose. Buildings with a commercial purpose include any building that is not residential, industrial, or agricultural. This definition is inclusive of building types that may not traditionally be considered "commercial," such as schools, correctional institutions, and churches. CBECS collects and publishes information on the stock of U.S. commercial buildings, their energy related building characteristics, and their energy consumption and expenditure. The first CBECS was conducted in 1979 and is typically conducted on a quadrennial basis. The tenth and most recent CBECS began data collection via computer assisted personal interviewing (CAPI) and computer assisted telephone interviewing (CATI) in April 2013 for reference year 2012. The 2012 CBECS is expected to have about 7,000 completed interviews from a sample of about 12,000.

Because there is no existing comprehensive list of commercial buildings in the U.S. to sample from, EIA must create a sampling frame. There are two parts of the frame: the area frame and the list frame. The area frame, which accounts for about 80% of the total frame, is created with traditional area listing. In the fall of 2012, trained field staff (referred to as “listers”) created the area frame for the 2012 CBECS by walking or driving through selected areas and recording information about every commercial building. The list frame, which accounts for about 20% of the total frame, is comprised of administrative lists of large buildings greater than 200,000 square feet in size.

Multi-stage area probability sampling was used to select the areas that were listed to create the area frame. The U.S. was first divided into primary sampling units (PSUs), which are counties or groups of counties, and a sample of these was selected. The selected PSUs were divided into secondary sampling units (SSUs), which are Census tracts or groups of Census tracts, and a sample of SSUs was selected. If the SSUs were too large, they were further divided into segments, which are smaller geographic sections of Census tracts or groups of tracts.

Prior to the 2012 CBECS, the most recent area frame was created in 2003 for the 2003 CBECS, and the 2012 CBECS sample design was based on the 2003 design. The 2012 area frame included three types of segments: segments new to the 2012 sample that required complete building listing (33% of the total segments), segments

listed in 2003 that were updated using dependent listing (7% of the total segments), and segments listed in 2003 that were not updated and required no field work (60% of the total segments). Approximately 90 trained listers worked to create the first two types of segments in the fall of 2012.

In the new segments, the lister recorded every eligible commercial building within the segment boundaries. In the 2003 listed segments that were selected for updating using dependent listing, the lister checked the 2003 listing and added new eligible commercial buildings that were not listed in 2003 and deleted buildings that no longer existed. For every building added to the frame, regardless of segment type, the lister recorded information for locating the building (address or description of building location if address is not available, building name, and establishment names if applicable) and information used for sampling purposes. The two main variables used for sampling stratification are building size (in square footage) category and building activity category, which are highly correlated with energy consumption. The listers were trained to estimate these building characteristics and record them on the listing sheets. Figure 1 below shows the CBECS building square footage and building activity categories for listing and sampling.

Table 1: CBECS Building Size and Activity Categories

Building size (square footage) categories

A	501* to 10,000
B	10,001 to 25,000
C	25,001 to 50,000
D	50,001 to 100,000
E	101,000 to 200,000
F	200,001 +

Building activity categories

1	Retail, entertainment and recreation, food sales (convenience store, liquor, retail bakeries, specialty food, etc.), post offices, automobile repair/service/maintenance/sales, offices, assembly halls, auditoriums, religious worship
2	Education, lodging (hotels, motels, dorms), nursing homes, public order & safety (courts, police & fire stations, prisons)
3	Food service (restaurants, bars, coffee shops, fast food, deli, diner, etc.), health care (inpatient & outpatient, hospitals, dental clinics, medical clinics, mental health clinics, veterinary, etc.), laboratories, laundromats, dry cleaners
4	Warehouse (refrigerated and non-refrigerated), storage, vacant

* Buildings estimated to be above 500 square feet are listed and included in the frame to avoid listers erroneously excluding buildings that may actually be above the 1,000 square foot minimum for inclusion in CBECS. Selected buildings that are actually 1,000 square feet or less are screened out in the interview phase.

The area frame is expensive to create and maintain due to the field listing: a large percentage of the 2012 CBECS budget was field listing costs. EIA is continually exploring methods for reducing costs, and one such idea is to use mapping services widely available on the internet to create the area frame in place of field listing. Google Earth™ (http://www.google.com/earth/) is one example of software that has features amenable to area listing: satellite views, a tool to measure distances that can be used to estimate length and width of a building (components of square footage), and Street View™, which gives the user panoramic views from the street at ground level. Street View can be used to gather building activity and the number of floors in the building, another component of square footage. The primary focus of this research was to determine if the two stratification variables, square footage category and building activity category, could be collected with the software and how they compared to what the lister recorded in the field.

Methodology

EIA selected a sample of segments new to the 2012 CBECS and a sample of 2003 update segments, then systematically sampled listed buildings within those segments and attempted to gather some of the same information the lister recorded in the field with Google Earth for those buildings.

Twenty new segments were selected out of 257 total new segments listed in the field. Every tenth building in these sampled segments was systematically selected for data gathering from Google Earth, for a total of 408 buildings. Fifteen 2003 update segments were selected out of 50 total that were dependent listed in the field. In those sampled segments, every fifth newly added building (a building not on the 2003 frame that the lister added) was systematically selected for data gathering, for a total of 123 buildings. The newly added buildings were the focus of the update segment analysis because EIA was specifically interested in how the Google Earth method would perform and the its coverage of new buildings. The sample size was limited by staff resources for this project.

For each building sampled, the researcher did the following:

1) Searched for the building on Google Earth using the address or description of the building location. If the building could not be found, if there was no Street View available for the building, or if the images were too poor to gather any information, this was noted and the researcher moved onto the next sampled building without taking the remaining steps.
2) Viewed the building in Street View to determine the number of floors, and to record the Google Earth Street View image date (the date the image was taken).
3) Measured the square footage and assigned a building activity category with tools available in Google Earth. The ruler tool was used to draw lines to measure the length and width of the building, in feet, which was multiplied by the number of floors to obtain the square footage. If the building was not a square or rectangle or made up of multiple rectangles, standard geometric techniques were used to calculate area when possible. The square footage point estimate was mapped into the appropriate category shown in Table 1. The researcher then assigned a building activity category by using the Street View image. If square footage and/or building activity category could not be determined, this was noted.

The researcher did not see the lister's estimates of square footage and building activity to avoid biasing the estimates.

All Google Earth data was collected in December 2012 using Google Earth version 6.2.

Results

Prevalence of Inability to Collect Data on Google Earth

As described in the Methods section, if no information about the building could be captured - because the building could not be found, there was no Street View image, or the image was too poor to use at all - this was noted. It was also noted when either square footage or building activity could not be estimated. Table 2 below shows the frequency of the inability to gather data by segment type.

Table 2: Frequency of Inability to Collect Data

	Building segment sample type	
	New (*n*=408 buildings)	**Update (*n*=123 buildings)**
No information could be gathered about the building	20%	26%
Square footage could not be estimated*	2%	7%
Building activity could not be estimated*	24%	32%

* Denominator does not include buildings where no information could be gathered

Rate of Agreement with Lister for Variables Estimated

When the Google Earth data collector was able to estimate square footage and/or building activity category, their estimate was compared to the lister's estimate. Table 3 below shows the rates at which the two estimates agree and disagree by segment type.

Table 3: Square Footage and Activity Category Agree/Disagree Rates

Segment type	Item	Agree rate	Disagree rate
New	Square footage category	86%	14%
Update	Square footage category	89%	11%
New	Activity category	94%	6%
Update	Activity category	86%	14%

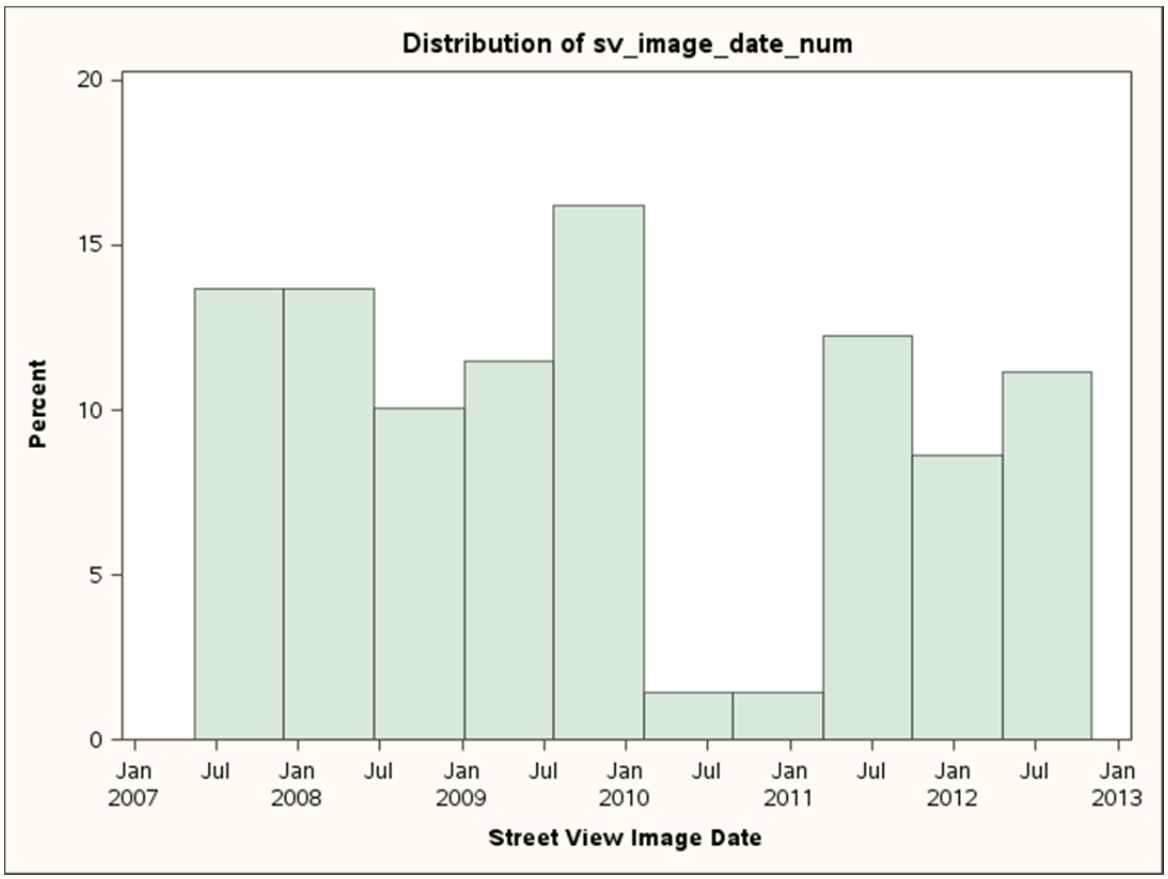
Of the buildings where the square footage category estimates were different (i.e. the Google Earth estimated category was above or below the lister's estimated category), a high percentage were off by only one category. For example, in the new segment buildings, of the 45 buildings where the square footage category estimates were different, 41 (91%) had a difference of only one category. The Google Earth estimate was more often higher than the lister's estimate: in 28 of the 45 (62%) new segment buildings where there was a square footage category disagreement, the Google Earth estimate was higher. The disagreements were more prevalent in smaller buildings: 78% of the new segment square footage disagreements were estimated to be in category A (501 – 10,000 square feet) or B (10,001 – 25,000 square feet) by the lister.

There was no pattern in the buildings where there was a disagreement in the building activity category estimates between Google Earth and the lister.

Street View Image Date Statistics

As discussed in the Methods section, the data collector recorded the date the Street View image was taken as part of the collection process. The average age of all Street View images of new segment sampled buildings, as of the end of data collection for this research project (January 1, 2013) was 3.2 years old. The average age of all images of the update segment sampled buildings was 3.6 years old. Figure 1 below is a histogram of the Street View image dates for the new segment sampled buildings.

Figure 1 - Histogram of Street View Image Dates in Sampled Buildings



Discussion

This project was a helpful first step in exploring the idea of using Google Earth to replace in-person field listings. Before attempting to list a segment of buildings from scratch on Google Earth, EIA wanted to determine if it was possible to estimate the two vital sampling stratification variables, square footage and building activity category, using the software. As a result of this research, the prospect of using Google Earth does not seem promising as a replacement for traditional listing in the area segments as a whole. The first and most important point to consider is that no information could be gathered on Google Earth on a high percentage of the buildings sampled (20% of the buildings sampled in the new segments and 26% in the update segments). This could indicate some potential undercoverage issues when compared to field listing.

When the Google Earth images were usable, square footage category was estimable in over 90% of the buildings, and the estimates on Google Earth agreed with the listers' 86% of the time in the new segments and 89% of the time in the update segments. However, building activity was much more difficult to estimate on Google Earth; nearly a quarter of the sampled new segment buildings and a third of the update segment buildings were not estimable for building activity.

It is important to note that when the Google Earth and lister's estimate disagreed for square footage or building activity category, it was not clear which estimate was correct. The Google Earth data collector could have made a mistake, the lister could have made a mistake, or potentially neither was a mistake, but the building changed size or activity in between the time the image was taken and the time the lister was in the field looking at the building. For example, a hospital could have added a large wing in the time after the Street View image was taken but before the lister arrived, bumping up its true size category, which would appear smaller if using Google Earth. Similarly, a

vacant building could become occupied by a restaurant, changing its building activity category. The average age of the Street View images was over three years, increasing the probability that the square footage and/or building activity could be out-of-date from what currently exists.

The age of the images also presents a problem for capturing newly constructed buildings on the CBECS frame. New buildings are important to include on the frame because they have different features and use energy differently than old buildings. If a Google Earth listing procedure does not adequately cover new buildings, the estimates of the building characteristics and consumption will be biased.

More research is needed on the dates of the images. The sample chosen for this project is small, concentrated around a small number of geographic areas (segments), and is not reflective of all images on Google Earth as a whole. It is also possible that many of the images were updated after this project was complete in December 2013.

The success of gathering information on these sampled buildings was entirely dependent on the quality of Google Earth coverage of them – the satellite images, Street View images, the date the images were captured, and the placement of addresses. Examples of problems encountered by EIA are: dated images, portions of segments with poor or no Street View coverage, grainy/blurry/sun-glared images, buildings too far from the street to see building activity or number of floors, very tall buildings too close to the street to determine the number of floors, maps placing addresses in the wrong location, and difficulty determining building boundaries when buildings are attached.

At this time, EIA does not believe that Google Earth should replace traditional field listing, however it may be promising in certain areas. Further research is needed to determine the qualities of a segment that would make the listing using this method comparable to the in-person field listing. These areas can then be listed with Google Earth and compared to an up-to-date in-person field listing to see how building counts and distributions of frame variables compare to listings. Because the ultimate goal of the listing is to provide the CBECS interviewer with information to identify a sampled building, EIA needs to be certain that the address information collected in Google Earth is accurate and clear.