Virtual Listing: GIS Approaches to Improve Survey Listing Efficiency

Michael Giangrande[1], J. Michael Brick[1], David Morganstein[1], Katie Lewis[2]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850
[2]EIA, U.S. Energy Information Administration, 1000 Independence Ave., SW, Washington, DC 20585

Abstract
As the availability and strength of geographic information systems (GIS) tools and data increase, the opportunities to use them to improve survey efficiency grow as well. In this paper, we discuss the use of various geospatial resources, such as those provided by Google and other commercial vendors, to conduct a virtual listing of commercial buildings. The virtual listing method was used to construct a frame for the 2018 Commercial Buildings Energy Consumption Survey, a survey conducted periodically for the U.S. Energy Information Administration. We discuss the advantages we found by having office staff remotely list areas using GIS tools. Also, we identify challenges we overcame as well as some that remain. The research design included a dual listing activity where specific areas were independently listed both virtually and using traditional on-the-ground listing methods. The resulting coverage, accuracy, and efficiency estimates for both approaches are evaluated.

Key Words: GIS, geographic information systems, listing, area probability sample, virtual listing, geospatial

1. Background

The 2018 Commercial Buildings Energy Consumption Survey (CBECS) is conducted by the U.S. Energy Information Administration, the independent statistical agency within the U.S. Department of Energy. Westat is the data collection contractor for this round of CBECS. The main objective of the study is to examine the current levels of energy use in commercial buildings across the United States, as well as to provide estimates of trends in energy use over time. The study collects statistics on characteristics of sampled commercial buildings, equipment in the building that is related to energy use, the quantity of energy used in the building, and associated cost of energy use.

2. Sampling

Most of the sample of commercial buildings for the 2018 CBECS was selected using a multistage area probability sample[1]. The first stage divided the entire nation into 700 large geographic regions (counties or groups of counties) called primary sampling units (PSUs).

[1] Several lists of large buildings, which are not discussed in this paper, were used to supplement the area probability sample.

We sampled 151 PSUs for CBECS using a probability proportional to size (PPS) procedure where the size was correlated to energy use in the PSU.

The sampled PSUs were then divided into 8,800 smaller geographical areas called secondary sampling units or segments. We sampled 764 segments, again using a PPS sampling procedure with a size measure correlated to the number and size of the buildings in the segments. In the third stage, we will select about 12,000 commercial buildings for the study. To accomplish this third stage, a frame of commercial buildings in the sampled segments is required, and no such frame currently exists. This research is related to the method of creating this frame of commercial buildings within the sampled segments.

CBECS has been conducted since 1979, and for each of the 10 previous rounds, the sample frame of commercial buildings in the sampled segments had been constructed using traditional field listing. This field listing involves sending "listers" to the sampled segments to list all the commercial buildings in the segments. Because of high costs of frame construction, for some of the rounds the lists of buildings from a previous round of CBECS were used and partially updated, even though the lists were several years old at that point.

3. CBECS Listing Requirements

Due to the specific and specialized listing requirements for CBECS (for example, applying a uniform definition of a building), on-the-ground field listers needed to be extensively trained on the process. This requirement made it difficult to recruit listers who were local to their assigned segments. In this regard, the commercial building listing operation was more complex and costly than the corresponding operation for listing households in household surveys. Expensive travel to the segments and within the segments often was required. The on-the-ground listers must completely canvass the segments, some of which are as large as 5,000 square miles.

CBECS listers were also required to estimate each building's total square footage, because building size is an important determinant of energy usage. Larger buildings typically use more energy than smaller buildings; the frame of buildings in each segment was stratified by size, and then larger buildings were sampled at higher rates than the smaller buildings. Approximate methods for estimating the building's square footage included training the listers to count their paces as they walked along the side of a building, or extrapolating based on the number of parking spaces observed in front of a building. The listers multiplied the length of the building by the width of the building, then multiplied by the number of floors in the building to report an estimated square footage.

4. Virtual Listing

Virtual listing utilizes GIS data and tools available in Westat's home office, enabling listing to occur remotely on a workstation, without sending data collection staff to the field. Virtual listing was developed to alleviate some of these costly, time consuming, and complex tasks by taking advantage of the new GIS resources and tools currently available. Having a team of virtual listers working in the same field room on the Westat main campus greatly reduced the costs associated with travel and helped standardize lister performance. Listing segments virtually enabled canvassing large areas quickly, and this virtual environment helped to create a more consistent and timely review of lister activity. In addition, we implemented tools in our Virtual Listing System (VLS) to automate the calculation of square footage.

The idea for virtual listing resulted from research conducted after the 2012 CBECS. Lewis (2013) studied virtual identification of commercial buildings by selecting a sample of buildings that were listed in the field. She conducted a test to see how many of those buildings were identifiable in Google Earth and Google Street View™. While Lewis was able to identify around 80% of the buildings, the research suggested that in 2012, Google products alone were insufficient to yield the quality required when identifying CBECS-eligible buildings.

In 2017, Westat revaluated the available Google products and the possibility of combining them with other data sources. Our research found that Google had made significant improvements to Google Earth, to their coverage of Street View™ (Ibarz and Banerjee, 2017), and to the frequency of their aerial imagery. This review also found that other enhancements were needed for this technology to be more capable for conducting a complete listing. We decided to add content to the mapping environment by adding geographic data layers that we built using other commercial data sources of businesses and structures. The goal was to have point of interest layers that, once clicked, displayed specific information about the buildings and businesses from those commercial sources such as the addresses and business types of the buildings. We also wanted to be able to store listed buildings in the VLS as polygons, allowing for the automated calculation of square footage. Creating this enhanced, custom virtual listing environment resulted in the development of our VLS. Unlike Lewis's approach, where she searched for addresses from buildings that were found via traditional listings and verified them within Google Earth, the VLS offers a comprehensive environment in which complete virtual canvassing of the segment areas is conducted.

The VLS is a fully integrated, custom, web-based system for observing, cataloging, and documenting CBECS eligible commercial buildings. The VLS transports the lister virtually to their assigned segment and enables them to canvass their segment remotely, using Google's 2-D and 3-D aerial satellite views, as well as panoramic street images from Google Street View™. The system includes custom geographic layers that are used as a reference for identifying buildings. Once a building is identified as being CBECS eligible, the lister lists the building by tracing the outline of the building's roof. The building's outline, or "footprint," is stored in the VLS along with other relevant building characteristics. Figure 1 is a screenshot of the VLS where the lister added a footprint to the building in the center of the screen (yellow polygon) and added the relevant building characteristics in the property panel on the right. The VLS automatically returns the buildings square footage once the number of floors is entered. After the lister saves the property panel, the building footprint turns blue and the building geography and attributes are stored in the VLS.

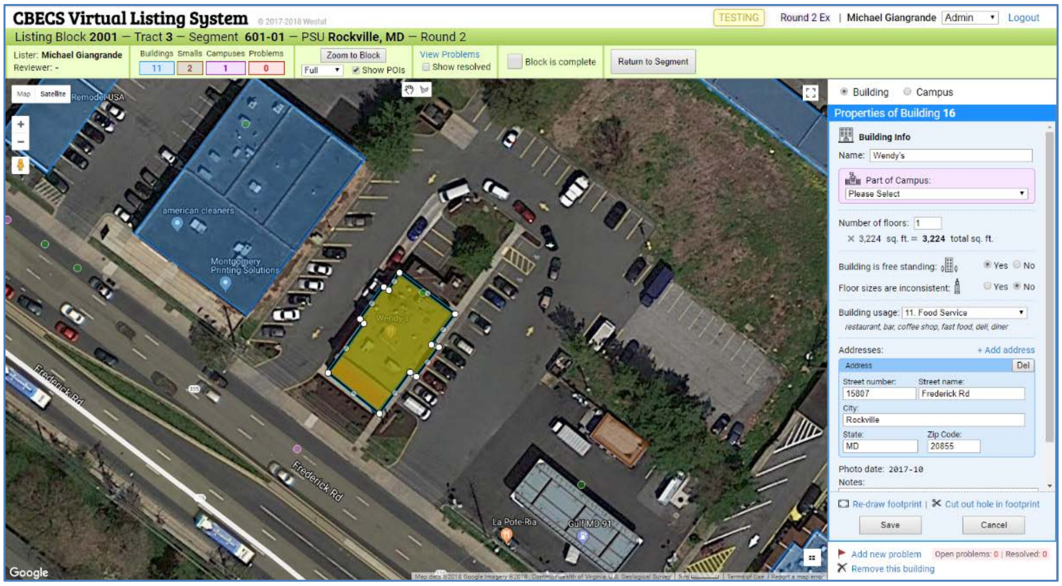

Figure 1: Screen shot of VLS building entry showing the building footprint and property panel.

5. Evaluation Design

The first step of the sample design called for virtually scanning all the sampled segments and classifying them as being either ineligible for virtual listing or as eligible for virtual listing. The availability of Google Street View™ and Google 3D, as well as the question of how current the imagery was in each segment, factored into the assessment of whether to not the segment was able to be listed virtually. Of the 764 segments sampled for the entire 2018 CBECS, only 26 or 3.5% of all sampled segments were ineligible for virtual listing. All other segments were eligible for selection for the test described below.

Of the segments that were eligible for both virtual and on-the-ground listing, an initial test was conducted that directly compared the two listing methods in a random sample of 50 segments. In these 50 sampled segments both types of listing were done independently. In other words, the virtual listers and the on-the-ground listers were completely unaware of the activities of one another. Both the virtual listers and the on-the-ground listers compiled a frame for each of the 50 segments. This sample of 50 segments is used here to assess the quality of the listing methods. The goal of this paper is to describe the quality of the listing in these segments so as to quantify the viability of virtual listing. This sample of segments provides a method for adjusting the estimates to account for any VLS under coverage using a method similar to a frame enhancement method (see Kalton, Kali, and Sigman, 2014).

6. Outcomes

We begin by describing some of the resources required to list for the two methods. Both the on-the-ground (OTG) field and VLS staff listed the same 50 segments in about 2 calendar months; however, the hours worked and production were quite different. It took the OTG listers over 4,000 hours to complete the 50 segments, while the virtual listing completed the same segments in 1,700 hours. By eliminating travel and square footage estimations, the virtual staff were able to enter buildings at a much faster rate, averaging 6.3 minutes per building, while the field staff averaged 15.3 minutes per building. Since the hourly rates of both the VLS and OTG listers were comparable, the cost differential between the methods is largely associated with the number of hours worked. The other main component of the direct cost is travel. Travel costs for the OTG staff, including the cost of flights, hotel stays, per diem, and car rentals for the 2-month field period, exceeded $125,000. There were no travel expenses for the virtual listing. Table 1 below highlights the cost differences between the two methods.

Table 1: Costs of on-the-ground and virtual listing in the 50 segments.

	On-the-ground listing	*Virtual listing*
Number of hours worked	4,008	1,700
Number of buildings listed	15,159	16,563
Number of minutes per building	15.3	6.3
Travel costs	Over $125K	$0

There are also indirect costs of listing such as reconciliation or review of listings by project staff. Reconciliation involves searching for addresses of buildings that were listed without street numbers where this information was not available in the field. Because virtual listers had multiple point of interest reference data layers, some of which included addresses, as well as access to the internet, the virtual listing yielded more complete data. As a result, virtual listing required substantially less reconciliation than the OTG listings.

Another indirect cost is supervision of the listers. All virtual listing was done in a central field room. This made supervision of listing easier and supervisors could manage multiple listers. The centralized facility also enabled answers to building eligibility questions to be immediately shared among all listers. Field listers were on their own to a much greater extent, having to make contact with their field supervisor after they had completed their listing activities for the day. Information learned by one OTG lister was not necessarily passed on to all other listers.

Another difference between the two methods was the ability to compute or estimate square footage of the buildings. As noted above, with OTG listing this is a complex, time-consuming process that sometimes results in errors in size categorizations. Virtual listing increases the accuracy in the square footage calculations since the VLS had an automated approach.

The main quality issue that the sample of 50 segments allows us to evaluate is the coverage of the listings. Published research on traditional listing coverage such as Eckman and Kreuter (2013) focuses primarily on residential household listings. Nevertheless, it is reasonable to assume that many of the challenges of household listing also apply to commercial building listing. The sample of 50 segments permits a direct evaluation of coverage of the two methods. This approach is different from the preliminary work of Lewis (2013) because both methods of listing can be compared at the same time rather than one-way coverage of virtual listing compared to building counts from an OTG method.

Evaluating the relative coverage of the two methods requires identifying which addresses are on both files and which are only on one of the files. This classification involves matching addresses from the virtual listing frame of buildings and the OTG listing frame, and this type of matching is always challenging, mainly due to formatting differences across addresses. While some of the matching was automated with the help of text-matching algorithms and location-based approaches, some manual matching had to be done, especially in rural areas, where addresses may be less standardized or missing. The information on addresses from the virtual listing was generally more standardized because the addresses primarily came from point-of-interest data sources or the internet. The OTG listers had to rely on only what they noticed in the field (e.g., street numbers on buildings or alternate road name information, which may or may not be present/visible). To be thorough, this matching process also had to happen in both directions (from virtual to OTG and from OTG to virtual) individually.

The analysis in Table 2 focuses on three primary outcomes after matching: identified in VL but missed in the OTG listing, identified in OTG listing but missed in VL, and found in both. The table shows the number and share of buildings with each outcome, both overall and by size strata. The table shows that 18.4% of all listed buildings were not identified by the OTG listing method and only 10.9% of all listed buildings were overlooked by the virtual method. For both listing methods, the highest percentage of missed buildings were the smallest ones (buildings as small as 501 square feet were supposed to be listed but only those greater than 1,000 square feet are eligible for the full CBECS study). Since energy usage is highly correlated with building size, the coverage of larger buildings is of much greater concern than that of smaller buildings.

Comparing the percent of the buildings missed by method, we see that the VLS missed a smaller percentage of buildings in every size category compared to the OTG. This finding suggests that replacing the OTG method with VLS (in those segments for which it is feasible) would improve coverage of buildings for CBECS. As noted above, for the 2018 CBECS, the VLS is being used in conjunction with OTG listing in a subsample of segments, rather than replacing it, to further improve the coverage for CBECS.

Table 2: Number of buildings listed and percent missed by method, by building size

	Total	*Percent*	*Missed in OTG*	*Percent*	*Missed in VLS*	*Percent*	*Found in Both*	*Percent*
Overall	18,588	100.0	3,429	18.4	2,025	10.9	13,134	70.7
Building Size (square feet)								
A - 501 - 10,000	13,690	100.0	2,955	21.6	1,751	12.8	8,984	65.6
B - 10,001 - 25,000	2,653	100.0	296	11.2	162	6.1	2,195	82.7
C - 25,001 - 50,000	1,066	100.0	98	9.2	59	5.5	909	85.3
D - 50,001 - 100,000	659	100.0	39	5.9	32	4.9	588	89.2
E - 100,001 - 200,000	328	100.0	24	7.3	9	2.7	295	89.9
F - 200,001 +	192	100.0	17	8.9	12	6.3	163	84.9

7. Future Work

The analysis is not yet complete. A sample of the buildings in the 50 segments that were missed in the virtual listing but found in the OTG listing will be included in the CBECS survey sample. We suspect that some, perhaps even many, of these will not be eligible for CBECS. For example, some of the buildings may meet the definition of a building for CBECS. Similarly, the eligibility of some sampled buildings missed in the OTG listing will be determined in the data collection. When this data collection step is completed, a more robust coverage assessment of each listing method can be made.

Data collection will also provide a means for capturing an actual reported square footage from the buildings in the study that are sampled and complete the survey. For buildings that were listed in both methods, comparing this reported square footage to the listed values will allow us to measure the validity of the automated square footage captured in the VLS compared to OTG listing.

8. Summary

In summary, we consider the virtual listing methodology and the Virtual Listing System itself to be a significant process improvement for CBECS. Along all the important dimensions considered – time, cost, quality, and coverage – virtual listing of commercial buildings for CBECS is an improvement over the previous OTG method.

We also learned important lessons in the initial implementation of virtual listing. Virtual listing is not an assignment that is well suited for everyone. It became clear that virtual listers require a higher technical aptitude and a stronger spatial awareness than the average OTG lister. For the virtual listing of the remaining segments for the 2018 CBECS, we targeted staff with GIS backgrounds. We have noticed an increase in productivity and quality due to better targeting staffing skills and aptitudes.

References

Eckman, S. and Kreuter, F. (2013). Undercoverage rates and undercoverage bias in traditional housing unit listing. *Sociological Methods & Research*, *42*, 264-293.

Ibaraz, J. and Banerjee, S. (2017). Updating Google Maps with Deep Learning and Street View. *Google AI Blog*, https://ai.googleblog.com/2017/05/updating-google-maps-with-deep-learning.html.

Kalton, G., Kali, J. and Sigman, R. (2014). Handling frame problems when address-based sampling is used for in-person household surveys. *Journal of Survey Statistics and Methodology*, *2*, 283-304.

Lewis, K. (2013). Exploratory research on the use of Google Earth to create a sampling frame of buildings, 2013 Federal Committee on Statistical Methodology Research Conference, Washington DC.