Evaluation of Methods to Form Segments from Census Blocks in Area Sample Designs

Jennifer Kali[1], Tom Krenzke[1], Ying Chen[1], Angela Chen[1], Jim Green[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract
In-person surveys often utilize a multi-stage sample design in which households are sampled within geographic areas called segments, improving cost efficiency by restricting the geographic range that data collectors must travel. Often, segments are formed by grouping neighboring census blocks until the number of housing units in the segment is large enough to support the household sample to be selected within the segment. A simple method to combine adjacent census blocks is to sort the census block file by the census block ID. Doing so often creates segments that are not contiguous, not complete (i.e., they contain holes), and not compact. Issues with contiguity and completeness create challenges for data collectors in determining which housing units to include in the sample frame. Less-compact segments increase interviewer travel costs. We will review alternative approaches to forming segments with three shape-filling curves – Peano, Hilbert, and Geo-hash, evaluating the segments formed by each sorting method according to contiguity, completeness, compactness, and between-segment variance, and will present a hybrid segment formation algorithm that utilizes all four sorting methods.

Key Words: area sampling, segments, multi-stage sampling, in-person surveys

1. Introduction

There are times in survey data collection when it is necessary to collect data in person, such as when a survey also includes an assessment of adult literacy skills or blood collection. In the case of in-person data collection, it is beneficial to cluster sampled persons to reduce the costs related to interviewer travel between their assigned cases. A common sampling technique for in-person data collection is called area sampling. Kish (1965) described the cost-variance trade-off for area sampling in which the smaller the clusters, the higher the between-cluster variance, but the lower the cost. In practice, the study’s budget constraints often dictate the size of the clusters.

Consider a national sample that includes a collection of blood samples from individuals. In such a study in which in-person data collection is required, data collection is quite expensive because data collectors must travel to each sampled dwelling unit. Therefore, instead of selecting a random sample of dwelling units in the entire nation, the sample is selected in stages. First, the nation is subdivided into fairly large geographic areas, such as counties or groups of counties, commonly called the primary sampling units (PSUs). Once a sample of PSUs is selected, smaller geographic areas are formed that are suitable for in-person data collection within each PSU. These smaller subdivisions are commonly called segments (or secondary sampling units). Once a sample of segments is selected, a list of dwelling unit addresses is created (or purchased from a vendor) within each segment, and a sample selected. Within the sample of addresses, the interviewer visits the selected dwelling unit and typically administers a screener questionnaire. During the screening process, individual members of the household are enumerated, eligible individuals are identified, and then a sample of individuals is selected for which blood can be collected.

Census blocks are often the building blocks for segment formation. The U.S. Census Bureau forms the census blocks and provides census data such as the number of people and households overall and by various demographic subgroups for each census block, which is useful during the segment formation and segment sampling processes. The purpose of segment creation is to create a small geographic area that is easy to field in a short amount of time, either by walking or driving, to increase data collection efficiency. Depending on the methodology used to group census blocks into segments, segment boundaries may be difficult to determine, which can become important for a listing operation or a coverage enhancement process, and may be oddly shaped, which may take longer to traverse. In this article, an algorithm is proposed that will combine census blocks into groups geographically close together to serve as the segment sampling frame in a manner that supports data collection efficiency. The goal of the algorithm is to create segments in a manner that balances cost, variance, and other operational aspects of implementing an in-person area probability survey.

2. Overview of Segment Formation

Segment formation follows a set of parameters that are specific to each study. One such parameter is the minimum number of sample units (either households or people) required within a segment, which is developed based on the study's sample design and operational needs using census demographic information associated with each census block. Another such parameter defines the segment boundaries. For a variety of reasons, study managers may prefer segments formed within specific geographic boundaries, which we call "hard boundaries." The proposed algorithm will create segments respecting the hard boundary, defined as a border that segments cannot cross. Because census blocks are the building blocks of segments, the hard boundary definition must follow census geography, such as block group, tract, county, or PSU.

A map of census blocks exists on a two-dimensional (2D) plane. The proposed algorithm identifies geographic neighboring blocks by creating a one-dimensional (1D) ordered list of blocks from the 2D plane. Adjacent blocks, according to this sorted list, are grouped until the minimum number of sample units (either households or people) is reached. The creation of the 1D block list can be done in several ways, which we will refer to as "sorting." The creation of segments hinges on the sorting method. Each sorting method creates a unique set of segments.

Consider the example in Figure 1, which shows the 32 blocks within one example's hard boundary. The study design calls for a minimum of 60 households per segment. The number of sampling units within each block is often called the measure of size (MOS) and is shown in the figure for each block. Many of the census blocks in this imaginary hard boundary have fewer than the required minimum of 60 households. In this example, blocks must be combined to create segments of sufficient size.

MOS = 13	MOS = 25	MOS = 61	MOS = 17	MOS = 10	MOS = 50	MOS = 53	MOS = 2
MOS = 15	MOS = 27	MOS = 100	MOS = 23	MOS = 37	MOS = 43	MOS = 18	MOS = 45
MOS = 20	MOS = 27	MOS = 33	MOS = 14	MOS = 46	MOS = 51	MOS = 73	MOS = 32
MOS = 12	MOS = 18	MOS = 16	MOS = 29	MOS = 31	MOS = 56	MOS = 16	MOS = 80

Note: 32 blocks; Measure of size (MOS)= # of housing units; Minimum MOS = 60

Figure 1: Example hard boundary with 32 census blocks

The numbers with the prefix "B" shown in Figure 2 represent the sorting method applied within this hard boundary. The sort order provides the definition of adjacency necessary for segment formation. Following to this sorted list of blocks B1 to B12, the algorithm groups adjacent blocks until the sum of MOS reaches the required minimum number of sample units and thus forms a segment. The sorting method shown in Figure 2 results in the set of segments shown in Figure 3.

B1 MOS = 13	B2 MOS = 25	B3 MOS = 61	B4 MOS = 17	B5 MOS = 10	B6 MOS = 50	B7 MOS = 53	B9 MOS = 2
B32 MOS = 15	B31 MOS = 27	B30 MOS = 100	B19 MOS = 23	B18 MOS = 37	B17 MOS = 43	B8 MOS = 18	B10 MOS = 45
B27 MOS = 20	B28 MOS = 27	B29 MOS = 33	B20 MOS = 14	B21 MOS = 46	B16 MOS = 51	B15 MOS = 73	B11 MOS = 32
B26 MOS = 12	B25 MOS = 18	B24 MOS = 16	B23 MOS = 29	B22 MOS = 31	B14 MOS = 56	B13 MOS = 16	B12 MOS = 80

Figure 2: Example hard boundary with 32 census blocks with sorting key assigned to each block

Applying a different sorting method to the same set of blocks will generate a different set of sorting key, which may result in a different set of segments, as shown in Figure 4. There could be many variations in segments depending on the sorting method chosen.

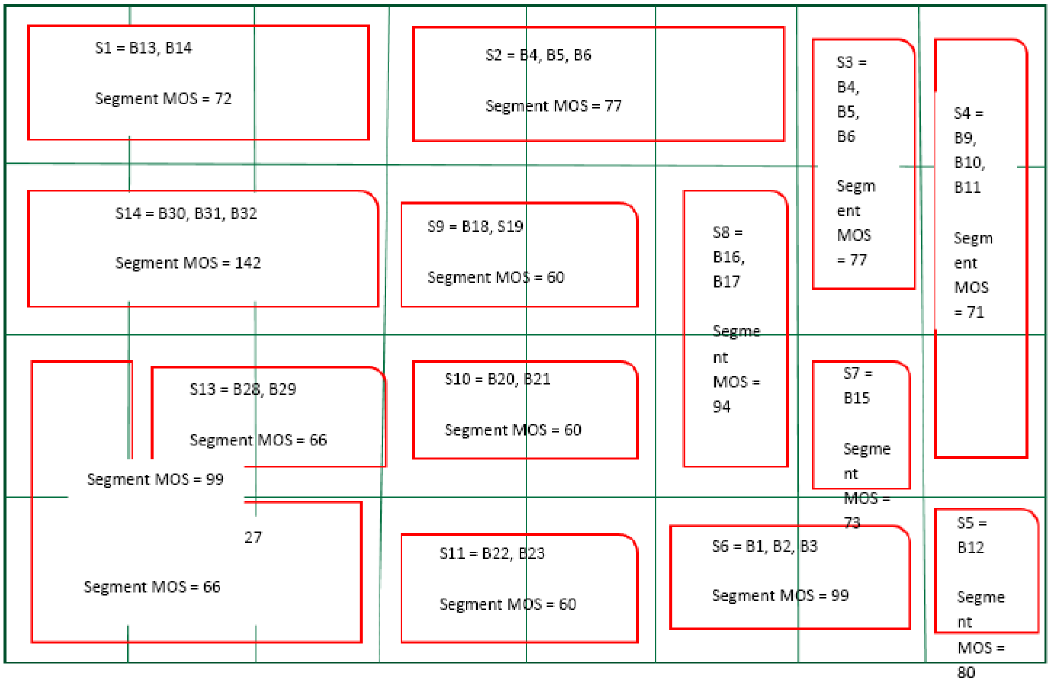


Figure 3: Segment assignment for example hard boundary with 32 census blocks based on sorting key shown in Figure 2

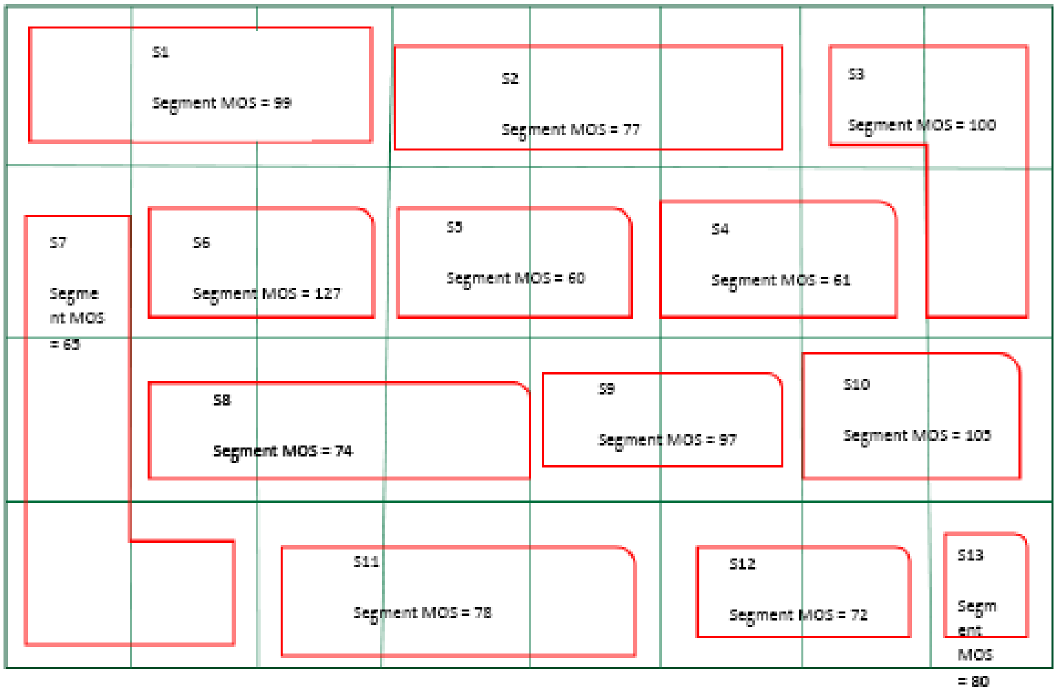


Figure 4: Alternative segment assignment for example hard boundary with 32 census blocks based on a different sorting key

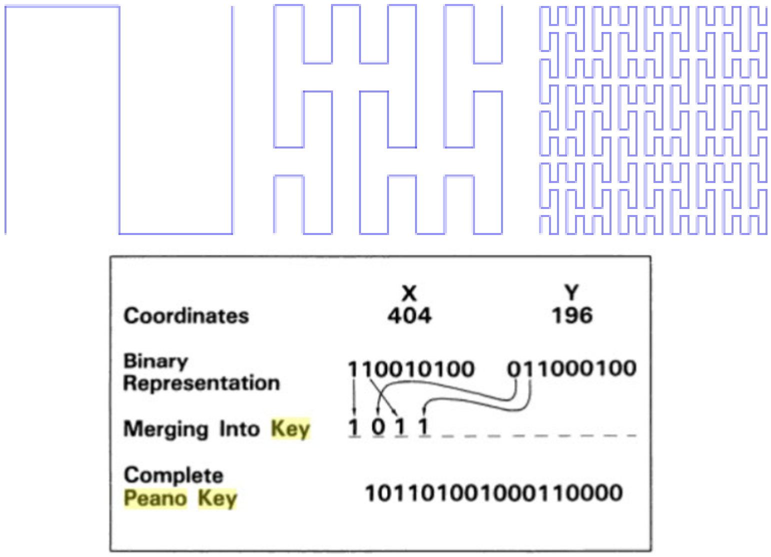
Each set of segments will have challenges that reduce the effectiveness of the cost reductions gained by clustering creating segments that are more time-consuming to navigate. By considering multiple sorting method used to form segments, the proposed algorithm addresses these challenges and attempts to meet the statistical needs of the study.

3. Sorting Methods

The proposed algorithm considers four sorting methods: Census block ID, Peano, Hilbert, and Geo-hash.

Census block ID: The census block ID is assigned by the Census Bureau for every census block. It is a numeric code that uniquely identifies all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data (United States Census Bureau, n.d.). The sequential ordering of the ID makes it a possible sorting method.

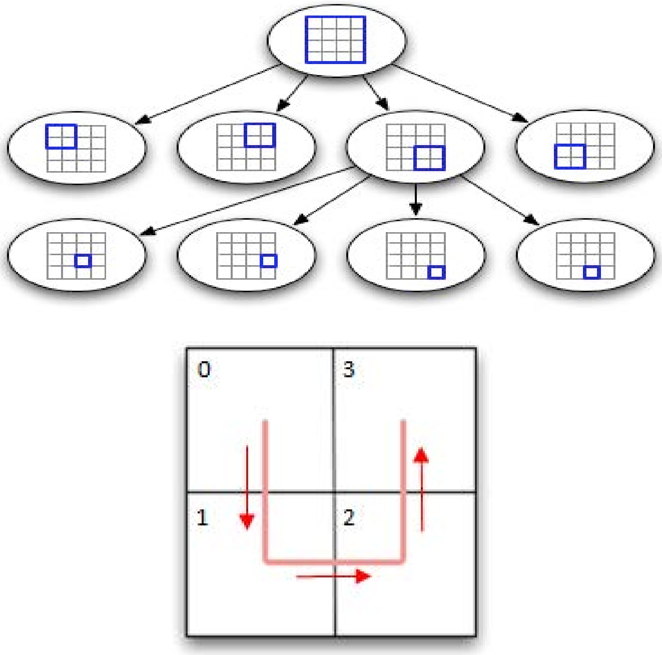
Peano: The Peano curve (Peano,1890) is a space-filling curve. The Peano key (illustrated in Figure 5) results from combining coordinates into a single key code composed of alternating longitude and latitude digits and used primarily for nearest-point searches. Garrett and Harter (1995) describe use the Peano curve for sorting geographic units for sampling. The top image in Figure 5 shows the pattern the Peano curve follows through a 2D space. The bottom image illustrates the creation of the binary Peano key from the latitude and longitude.



Source: Marx and Saalfeld, 1988

Figure 5: Peano Curve Sorting Key

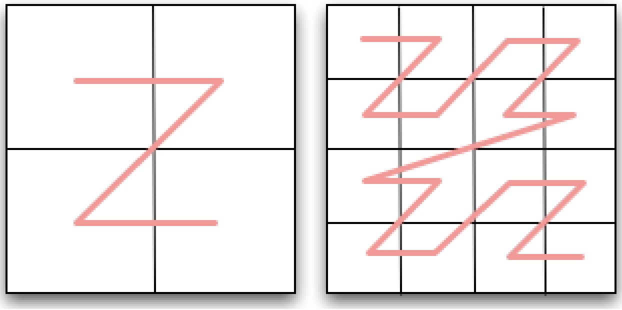
Hilbert Curve: The Hilbert curve (Hilbert, 1891) is a space-filling curve and a type of quad tree (illustrated in Figure 6). Quad trees create a geospatial index by grouping neighboring cells in a quadratic fashion following a root-and-node branching method until all units in the space have been indexed, as shown in the image on the left in Figure 6. The basic element of a Hilbert curve is a U-shape, as shown in the image on the right in Figure 6.



Source: Johnson, 2009

Figure 6: Hilbert Curve Sorting Key

Geo-hash: Geo-hash (Johnson, 2009), shown in Figure 7, is an application of a space-filling curve in the quad tree family similar to the Hilbert curve except that it utilizes a Z-shaped pattern instead of a U-shaped pattern. A geo-hash key is a fixed value based on subdividing the entire earth at once. Every latitude and longitude for the entire earth has been preassigned a geo-hash key.



Source: Johnson, 2009

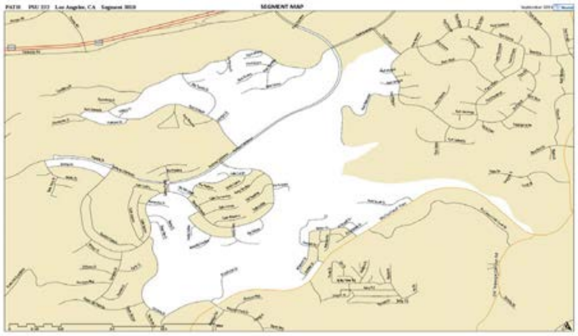
Figure 7: Geo-hash Sorting Key

Sorting the census blocks using these four different sorting keys will give different solutions and form four different sets of segments. The proposed segment formation algorithm utilizes all four sorting methods to create four unique sets of segments. The next section describes metrics to evaluate the four sets of segments.

4. Segment Formation Metrics

Segment characteristics, such as shape and associated population demographics, can be compared to determine which set of segments has features that best support the goals of the area sampling application—both regarding cost reductions related to data collection and reductions on the variance related to the clustering of sampled units. To this end, we consider eight metrics to evaluate the four sets of segments created using the four sorting methods.

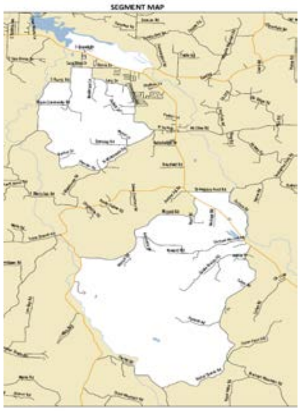
Integrity. A segment with integrity is complete without any holes. Figure 8 represents a segment with a hole. Segments without integrity are challenging for data collectors because the boundaries of the segment are difficult to determine. To measure the integrity, we count the number of holes within a segment.



NOTE: Segment is represented in white.

Figure 8: Example of Segment Lacking Integrity

Contiguity. A segment with contiguity has only one part. Figure 9 shows a segment that lacks contiguity, as it is made of two distinct parts. Segments lacking contiguity are challenging for data collectors for the same reason as segments lacking integrity—the boundaries of the segment are difficult to determine. They may also result in increased travel time compared with contiguous segments and increased data collection costs. To measure the contiguity, we count the number of segments that have multiple parts.

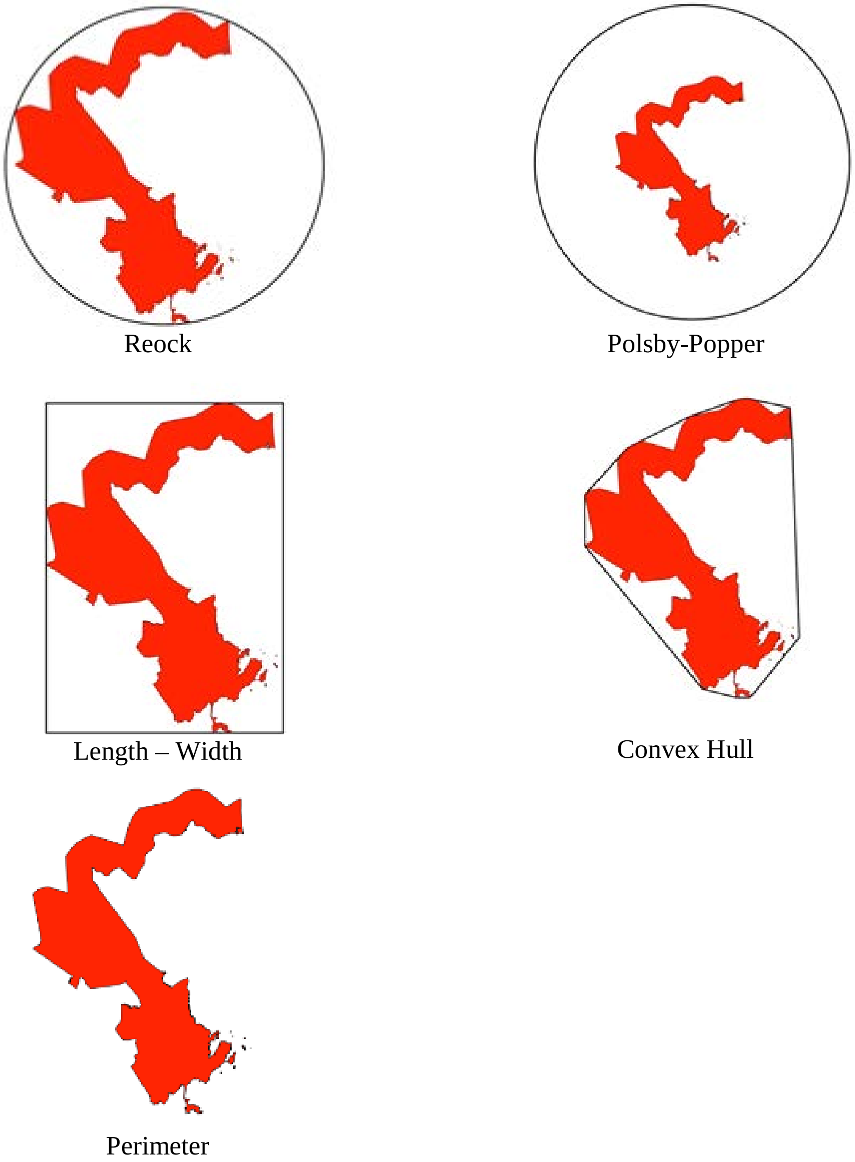


NOTE: Segment is represented in white.

Figure 9: Example of Segment Lacking Contiguity

Between-cluster variance. If there is a demographic characteristic that is important to the study, such as the percentage of Hispanic persons within the segment, it is beneficial to create segments that are heterogeneous within each segment and homogeneous between segments. Doing so will reduce the effect of the clustering on the variance of the estimates.

Compactness. The algorithm considers five commonly used compactness metrics: Polsby-Popper, Reock Score, Length-Width, Perimeter, and Convex Hull (Figure 10), which all illustrate compactness. Compactness is a measure used in the development of congressional districts. These measures compare geometric features of the geographic area (e.g., perimeters, areas) to the features of a related base geometric object (e.g., minimum bounding circle). More compact segments should result in reduced travel time and travel costs for data collectors. According to geometry, the most compact shape is a circle.



Reock

Polsby-Popper

Length – Width

Convex Hull

Perimeter

Source: Measuring Compactness, n.d.

Figure 10: Examples of Contiguity Metrics

5. Compactness

Polsby-Popper: The Polsby-Popper measure is the ratio of the segment's area to the area of a circle whose circumference is equal to the perimeter of the segment. (Measuring Compactness, n.d.)

Reock Score: The Reock Score is the ratio of the segment's area to the area of a minimum bounding circle that encloses the segment's geometry. (Measuring Compactness, n.d.)

Length-Width: The Length-Width Ratio is calculated as the absolute value of the difference between length and width in miles. (Caliper Mapping and Transportation Glossary, n.d.)

Perimeter: Measurement in miles of the perimeter of the segment. (Caliper Mapping and Transportation Glossary, n.d.)

Convex Hull: The Convex Hull score is the ratio of the segment's area to the area of the minimum convex polygon that can enclose the segment's geometry. (Measuring Compactness, n.d.)

6. Evaluation of Four Sorting Methods

Four segment formation sorting methods were evaluated based on block-level data that included two randomly selected PSUs per state for each of the 50 states. The hard boundaries in the formation process were census tracts, which are geographic areas comprising several blocks. Three characteristics[1] of persons were used for the between-cluster variance metric, and the minimum size to form a segment was 200 households.

Table 1 shows the eight evaluation metrics for each sorting method. The cells highlighted in pink and in bold text identify the best value for each metric. The value for each cell is the average value of each metric within each census tract.

[1] The characteristics considered were total household population counts for different race/ethnic groups: total non-Hispanic Black in combination with other races, non-Hispanic Black alone or in combination with one or more other races, and non-Hispanic White alone.

Table 1: Eight evaluation methods for each sorting method

Metric	*Desired Direction (Optimal Bound)*	*Block ID*	*Hilbert*	*Peano*	*Geo-hash*
Integrity/ Donuts	Low (0)	0.56	0.29	**0.26**	**0.26**
Contiguity/ Splits	Low (0)	1.44	**0.82**	0.96	0.96
Between Cluster Variance	Low (0)	43510	**42756**	43863	42720
Compactness:					
Reock	High (1)	0.29	**0.33**	0.30	0.30
Perimeter	Low (0)	24.39	21.50	21.75	21.91
Polsby-Popper	High (1)	0.29	0.31	0.30	0.30
Length-Width	Low (0)	2389.4	2065.5	2319.6	1970.6
Convex Hull	High (1)	0.60	0.68	0.64	0.63

Across the eight metrics, the Hilbert sorting method produced the best results for five of the eight metrics. For example, consider the second row of Table 1. On average, the census block ID sorting method has 1.44 noncontiguous segments per tract while the Hilbert method only has 0.89. The Geo-hash approach has the best results for three of the metrics and the Peano approach has the best results for one of the metrics (Integrity). The census block ID sorting method does not produce the best result for any metric, and it actually produces the worst results for seven of the eight metrics. Based on these results and the evaluation parameters, using the Hilbert curve sorting method will result in better segments overall. That being said, we proposed a hybrid method (described in the next section) that pools results from each sorting method.

7. Hybrid Method

The hybrid algorithm first creates four sets of segments within each hard boundary, derived from each of the four sorting methods. The methods are compared according to the eight metrics discussed above and the best set of segments for each hard boundary determined. The segment sampling frame is created by combining the best segment set for each hard boundary so that the sorting method will vary by hard boundary. The steps of the algorithm are shown in Figure 11 and listed below.

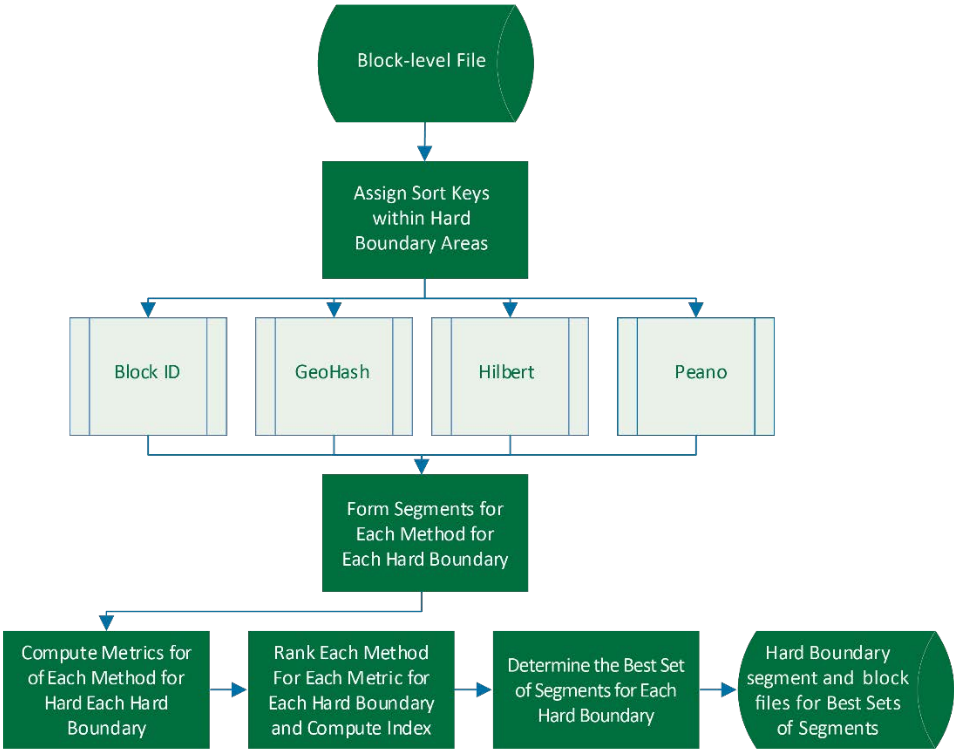

Block-level File
Assign Sort Keys within Hard Boundary Areas
Block ID
GeoHash
Hilbert
Peano
Form Segments for Each Method for Each Hard Boundary
Compute Metrics for of Each Method for Hard Each Hard Boundary
Rank Each Method For Each Metric for Each Hard Boundary and Compute Index
Determine the Best Set of Segments for Each Hard Boundary
Hard Boundary segment and block files for Best Sets of Segments

Figure 11: Flow Chart of Hybrid Method to Form Segment Sampling Frame

1. Start with a file of all census blocks within targeted areas, such as within sampled PSUs.
2. Define a hard boundary that is relevant to the study and follows census geography (e.g., block group, tract, county, PSU).
3. Within each hard boundary, create four sets of segments based on the four sorting methods.
4. Choose variables available at the census block level that are related to study outcomes.
5. Compute the eight metrics for all segments for all sorting methods.
6. Summarize the eight metrics to the hard boundary level.
7. For each hard boundary level metric, rank the set of segments created by each sorting method (i.e., 1 = set of segments with the smallest average perimeter, …, 4 = set of segments with the largest average perimeter).
8. Assign a level of importance to each metric based on the considerations for the study. Metrics can be given equal importance (1/8), or some metrics can be given a higher importance than others (i.e., Integrity has an importance score of 1/2, between-cluster variance is 1/8, while all others have an importance rating of 1/14). Metrics can be left out of computation by giving it an importance rating of 0.
9. Compute a hard boundary level score for each sorting method by summing the hard boundary ranks weighted by the importance ratings. The most optimal set of segments according to the selection criteria and evaluation formula is the set of segments that will form the segment sampling frame for that hard boundary.
10. The final segment sampling frame combines each hard boundary level segment frame into one larger file that will serve as the segment-level sampling frame.

8. Evaluation of Hybrid Method

The same input files and parameters that were used in the previous evaluation (with results resented in Table 1) were used to evaluate the hybrid method, with each metric given equal preference. Table 2 shows the number of hard boundary "wins" for each sorting method. A win means that the sorting method produced the best set of segments for that hard boundary. Viewing the table in a different way, the table provides the number of times the sorting approach was used in the creation of the final hybrid frame. The findings are consistent with the earlier finding that the Hilbert method creates the best sets of segments in general. That is, the Hilbert approach was used the most—for 45 percent of the hard boundaries (2,016/4,499). This means that the segments in the final segment sampling frame perform better than the Hilbert in 55 percent of the hard boundaries. The Geo-hash was used the second-most number of times and the Peano approach was used the third-most number of times. The census block ID method was used the least, for 15 percent of the hard boundaries. The benefit of the hybrid approach is that it can utilize the best performing method among the four sorting methods.

Table 2: Number of Hard Boundary Wins by Each Sorting Approach

Type	*Number of Hard Boundaries*	*Block ID*	*Hilbert*	*Peano*	*Geo-hash*
Count	4,499	674	2,016	835	974
Percent	100%	15%	45%	19%	22%

Table 3 provides the same results as Table 1, with a column added for the hybrid method. This table provides the average metric score across all sorting methods. The cells highlighted in pink with bold text indicate the original sorting method with the best results. The cells highlighted in green with italic text indicate the metrics in which the hybrid method outperformed the four original sorting methods. In seven of the eight metrics, the hybrid method outperformed the other sorting methods. For example, looking again at the Integrity metric, the census block ID sorting method produced an average of 1.44 integrity issues within a hard boundary, the Hilbert curve performed much better with an average of 0.89 issues per hard boundary, while the hybrid method performed even better with an average of 0.69 issues per hard boundary. The Hilbert curve outperformed the hybrid approach for the between-cluster variance metric. The between-cluster variance is the only metric that does not consider the shape of the segment. If the between-cluster variance was given a higher importance weight in the score function, the results would likely be different.

Table 3: Average Values of Evaluation Metrics for the Hybrid Approach and Each Sorting Method

Metric	*Desired Direction (Optimal Bound)*	*Block ID*	*Hilbert*	*Peano*	*Geo-hash*	*Hybrid*
Integrity/ Donuts	Low (0)	0.56	0.29	**0.26**	**0.26**	*0.23*
Contiguity/ Splits	Low (0)	1.44	**0.82**	0.96	0.96	*0.69*
Between Cluster Variance	Low (0)	43510	**42756**	43863	42720	*43546*
Compactness						
Reock	High (1)	0.29	**0.33**	0.30	0.30	*0.35*
Perimeter	Low (0)	24.39	**21.50**	21.75	21.91	*20.37*
Polsby-Popper	High (1)	0.29	**0.31**	0.30	0.30	*0.34*
Length-Width	Low (0)	2389.4	2065.5	2319.6	**1970.6**	*1867.9*
Convex Hull	High (1)	0.60	**0.68**	0.64	0.63	0.69

9. Summary

There are many ways to combine census blocks to create segments for area sampling. Ideal segments reduce data collection costs without incurring a large clustering effect on the variance of the estimates produced. Given the 100 randomly selected counties (two from each state), after evaluating the segments formed by each sorting method according to contiguity, completeness, compactness, and between-segment variance, the Hilbert curve sorting method creates the "best" segments, followed by Geo-hash.

The proposed hybrid approach creates segment frames within a hard boundary. Within each hard boundary, all four sorting methods are used to create four sets of segments and a "best" segment set chosen as the segment sampling frame. These hard boundary segment frames are then combined to form the final segment frame. This approach results in an optimal set of segments over what would result from any one of the four sorting methods discussed in this article.

The approach described here uses census blacks as the basis of the segment frame. We are working to expand the hybrid formation algorithm to utilize larger census geographic areas, such as block group, tract, and county, thus allowing the formation of larger segments and primary sampling units.

Acknowledgements

We would like to thank our colleagues at Westat for contributing to the development of this algorithm—in particular, Matthew Airola, Marcelo Simms, Katie Hubbell, Jeri Mulrow, and Michael Giangrande.

References

Caliper Mapping and Transportation Glossary [n.d]. “What are Measures of Compactness?” Retrieved from https://www.caliper.com/glossary/what-are-measures-of-compactness.htm.

Garrett, J., and Harter, R. (1995). “Sample Design Using Peano Key Sequencing in a Market Research.” Business Survey Methods.

Hilbert, D. (1891). Über die stetige Abbildung einer Linie auf ein Flächenstück.Mathematische Annalen 38, 459–460.

Johnson, N. (2009, November 9). “Damn Cool Algorithms: Spatial Indexing with Quadtrees and Hilbert Curves” [Blog post]. Retrieved from http://blog.notdot.net/2009/11/Damn-Cool-Algorithms-Spatial-indexing-with-Quadtrees-and-Hilbert-Curves

Kish, L. (1965) Survey sampling. John Wiley and Sons, Inc., New York.

Marx, R., and Saalfeld, A. (1988). “Programs for Assuring Map Quality at the Bureau of the Census.” Proceedings of the Annual Research Conference. Retrieved from https://books.google.com/books?id=fMa4i8nS_YoC&pg=PA253&lpg=PA253&dq=peano+key+census+blocks&source=bl&ots=smYZCUF0t0&sig=ACfU3U1ZJy9f0PaVkes-I0_ukthA0H3hag&hl=en&sa=X&ved=2ahUKEwi4s5nGsMPrAhU7hXIEHfteAooQ6AEwEnoECAEQAQ#v=onepage&q=peano%20key%20census%20blocks&f=false

“Measuring compactness.” [n.d.] Retrieved from https://fisherzachary.github.io/public/r-output.html

Peano, G. (1890). “Sur une courbe, qui remplit toute une aire plane.” Mathematische Annalen *(in French)*, 36 (1): 157–160. doi*:*10.1007/BF01199438.

United States Census Bureau. [n.d.] Understanding Geographic Identifiers (GEOIDs). Retrieved from https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html.