# Building a Segmentation Model to Target the 2010 Census Communications Campaign

**Nancy Bates and Mary H. Mulry**[1]
Census 2010 Publicity Office, U.S. Census Bureau, Washington, DC 20233
Statistical Research Division, U.S. Census Bureau, Washington, DC 20233

## Abstract

For the 2010 Census Communications Campaign, the U.S. Census Bureau has developed a research-based audience segmentation framework. This paper presents our findings from a macro-level segmentation study designed to help target markets and effectively deliver media messages planned for the 2010 Census. First, we performed a tract-level factor analysis using demographic, housing, and socioeconomic characteristics and mail response in Census 2000. This defined the underlying constructs behind the hard-to-count populations. The factor analysis revealed three non-correlated dimensions highlighting three distinct factors that describe populations with low mail return rates in 2000. We next performed a cluster analysis to identify mutually exclusive segments of the population according to propensity to mail back a census form. The cluster analysis revealed eight distinct groups varying across the entire spectrum of mailback propensities from high return rates to low in 2000 and each with unique demographic, housing, and socioeconomic characteristics.

**Key words**: Factor analysis, cluster analysis, mail return rate

## 1. Introduction

For the 2010 Census Communications Campaign, the U.S. Census Bureau has developed a research-based audience segmentation framework. This paper presents our findings from a macro-level segmentation study designed to help target markets and effectively deliver media messages. This analysis uses tract-level variables correlated with mail nonresponse. These data allow for segmentation of the population according to indicators related to mailback behavior.

The 2010 Census marketing campaign has three goals: to increase mail response rates, to increase accuracy and reduce the differential undercount, and to increase cooperation with door-to-door enumerators. Like most campaigns, resources for the 2010 Census are limited and must be allocated effectively to achieve the campaign goals and maximize return on investment (ROI). The first goal -- increased mail response -- is also a critical measure of ROI. Mail returns are a much cheaper way to count households and also provide better quality data than data collected during personal visit followups (Hillygus, Nie, Prewitt and Pals, 2006). The Census Bureau has estimated that a single percentage increase in mail returns translates to roughly 75 million dollars saved in personal follow-up costs. The mail response rate in Census 2000 was viewed as a success at 67 percent, much higher than the forecasted 61 percent (U.S. Census Bureau 2000). Some of this success was attributed to that fact that Census 2000 was the first to use paid advertising.

The approach for the market segmentation for 2010 presents a departure from the Census 2000 communications campaign. In 2000, the audience segmentation model (a.k.a. the Likelihood Spectrum™ ) was built solely upon consumer survey data (Baron and Billia, 1999). The Likelihood Spectrum™ divided the population into three broad segments: those likely to respond by mail, those undecided, and those unlikely to respond by mail.

For the 2010 Census market segmentation, we first performed a factor analysis using demographic and mail response characteristics from Census 2000 to define the underlying constructs behind the hard-to-count populations. The factor analysis revealed three non-correlated dimensions highlighting three distinct factors that set the foundation for understanding populations with low mail return rates in 2000. These factors were Economically Disadvantaged, Single Unattached Mobiles, and High Density Ethnic Enclaves.

[1] Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are the authors' and not necessarily those of the U.S. Census Bureau. Do not quote or cite without permission.

Subsequently, we performed a cluster analysis to identify mutually exclusive segments of the population according to propensity to mail back a census form. The cluster analysis revealed eight distinct groups varying across the entire spectrum of mailback propensities from high return rates to low in 2000 and each with unique demographic, housing, and socioeconomic characteristics.

## 2. Data: The 2000 Planning Data Base (PDB)

The source of the data used in this study is the U.S. Census Bureau 2000 Planning Database (PDB). This is a tract-level database that is publicly available and contains a range of housing, demographic, and socioeconomic variables correlated with mail response (Bruce and Robinson 2006). The 2000 Census is the data source and the PDB contains all tracts with population and housing units in the Census 2000 mail universe. After excluding nonrepresentative tracts[2], our dataset for analysis contained 62,708 tracts within the 50 states. This file was then merged with a Census 2000 operational file that contained the mail return rate for each tract. The mail return rate is defined as the percentage of occupied housing units eligible to receive a mail form that returned a form. This yielded a macro-level indicator of behavior by tract.

In addition to housing and socioeconomic indicators, the PDB also contains a "hard-to-count" (HTC) score (Robinson, Johanson, and Bruce 2007). HTC scores range from 0-132 for any given tract. This score is highly correlated with mail return rates and is constructed from twelve variables:

- % vacant units,
- % non-single family attached/detached units;
- % renter occupied units;
- % units with >1.5 persons per room ;
- % non-spousal units;
- % units without phone;
- % people below poverty level;
- % units receiving public assistance;
- % people unemployed;
- % linguistically isolated households, and
- % moved within last year.
- % adults without high school education

## 3. Factor Analysis

A first step in building a targeted model is to define the underlying constructs behind areas deemed as HTC. To address this, we performed a principal components factor analysis using the twelve PDB variables that make up the HTC scores (see Table 1). The analysis revealed three distinct factors (sometimes referred to as unobserved variables) highlighting three different population segments all hard to count by mail. The three factors were subsequently labeled:

(1) The Economically Disadvantaged,
(2) The Unattached/Mobile Singles, and
(3) High Density Areas with Ethnic Enclaves.

The Economically Disadvantaged factor had high loadings on vacant housing, poverty, public assistance, unemployment, less than a high school education, and absence of a phone. This factor has the largest negative correlation with mail return rates (-.56). The average mail return rate in tracts scoring high on this factor was far below average at 63.5 percent (the national average of representative tracts in the PDB was 75.4%). Tracts scoring high on this factor also had a high correlation with percent Black and a moderate correlation with percent American Indian and Alaskan Native (AIAN). Tracts with high Economically Disadvantaged scores had an average HTC score of 75 (well above the national average of 33). In summary, this factor reflects struggling underclass populations and underserved communities.

The second factor (Unattached/Mobile Singles) is distinct from the first with high factor loadings on non-spousal households, renters, multi-unit structures, and residential mobility within the last year. Tracts loading high on the Unattached/Mobile Single factor also had below average mail return rates (66.5%) and

[2] Nonrepresentative tracts are tracts that are sparsely populated or have a large percentage of group quarters or a large percentage of vacant housing units -- in all about 4% of all 2000 Census tracts.

a fairly strong negative correction with mail return rate (-.48). Tracts closely aligned with this factor did not indicate a strong correlation with any one race or ethnic group. In summary, this factor tends to reflect mobile, single adults, many of whom do not have children and may be living on their own for the first time

**Table 1.** Factor Analysis of Tract-Level Planning Database with Census 2000 Data with the HTC variables

| | Factor 1 –<br>***Economically disadvantaged*** | Factor 2 –<br>***Unattached/mobile singles*** | Factor 3 –<br>***High density w/ethnic enclaves*** |
|---|---|---|---|
| Underlying housing and social characteristics: | - Vacant housing<br>- Poverty<br>- Public Assistance<br>- Unemployment<br>- Less than high school education<br>- No phone | - Multiunit structures<br>- Renters<br>- Nonspousal households<br>- Persons moved in last year | - Crowded housing<br>- Few vacant houses<br>- Linguistic isolation<br>- Less than high school education |
| Correlated demographic characteristics: | - High correlation with % Black<br>- Moderate correlation with % AIAN<br>- Moderate correlation with % pop <18 (children) | - No strong correlation with any one race/ethnicity ( diverse )<br>- Moderate *negative* correlation with % pop <18 (absence of children) | - High correlation with % Hispanic<br>- Moderate correlation with % Asian or NHOPI<br>- Moderate positive correlation with % pop <18 (children) |
| % variance explained by each factor: (cumulative =74.7%) | 46.2% | 14.7% | 13.8% |
| Average 2000 mail return rate for tracts with high factor score: (national average mail return rate=75.4%) | 63.5% | 66.5% | 67.2% |
| Average hard-to-count score for tracts with high factor score: (national average HTC score=33) | 75 | 65 | 75 |
| Pearson correlation coefficient with 2000 MRR | -.56 | -.48 | -.21 |
| Number of tracts and % of total tracts with high factor scores | N=7051<br>(11.2%) | N=4073<br>(6.5%) | N=3758<br>(6.0)% |

The final factor (High Density Area with Ethnic Enclaves) loaded high on only three HTC variables: crowded housing, linguistic isolation, and less than high school education. Tracts with high scores on this factor had below average mail return rate (67.2%), an above average HTC score (75), a strong correlation with percent Hispanic and some correlation with percent Asian or Native Hawaiian/Other Pacific Islander (NHOPI). The underlying construct with this factor appears to be densely populated ethnic enclaves -- some with limited English language proficiency.

In summary, our factor analysis groups *variables* into distinct underlying factors – in our case we use the twelve variables that make up the HTC score. The analysis revealed three noncorrelated dimensions highlighting three different population segments all hard to count by mail. This serves as the foundation for understanding the below average mailback population and how they represent three distinct constructs.

## 4. Cluster Analysis

Following the factor analysis, we performed a cluster analysis also using data from the 2000 tract-level Planning Data Base. Unlike factor analysis, a cluster analysis groups *objects* (in our case tracts) with similar characteristics into relatively homogenous subsets. The cluster analysis groups each and every tract into one of several mutually exclusive clusters creating a multidimensional classification typology. The goal is to produce a macro-level market segmentation based on propensity to mail back a Census 2000 form. Unlike the factor analysis which serves to illustrate the underpinnings of the hard-to-count populations, the cluster analysis encompasses the entire spectrum of mailback propensities from high mail return rates to low. The two techniques are complimentary since both perform clustering functions, but with slightly different purposes.

There are many ways to perform cluster analysis. Our study uses the SAS procedure FASTCLUS to perform a disjoint cluster analysis based on distances computed using the 12 Hard-to-Count score variables in the PDB. Each observation (i.e., a tract) is assigned to one and only one cluster. The FASTCLUS procedure uses Euclidean distances so the cluster centers are based on least-squares estimation. The method is sometimes called the k-means model, since the cluster centers are the means of the observation assigned to each cluster.

For our analysis, we requested eight mutually exclusive clusters and a maximum number of 100 iterations.[3] We settled upon these parameters after several rounds of exploratory analysis using fewer clusters and iterations. Eight clusters seemed to satisfy our requirements by producing distinct enough groups that could be logically named according to their differences from (and in some cases similarity to) one another. In three instances, pairs of clusters appear closely related to one another with homeownership/renter status as the distinguishing feature. The eight groups ranged in size from the largest (representing 35% of all occupied housing units) to the smallest (reflecting only 2% of all occupied units).

Table 2 and Figure 1 clearly illustrate that the clusters capture where the low, medium, and high mail response tracts were located.

**Table 2.** Mail Return Rate, Number of Tracts, and Occupied Housing Units by Cluster

| # | Cluster Name | 2000 Mail Return Rate | Total Occupied Housing Units | | Number of Tracts |
|---|---|---|---|---|---|
| | | | Number (in millions) | Percent | |
| 1 | All around average I (homeowner skewed) | 77.3% | 36.5 | 35% | 21,174 |
| 2 | All around average II (renter skewed) | 74.2% | 16.5 | 16% | 8,957 |
| 3 | Econ. Disadvantaged I (homeowner skewed) | 66.5% | 6.6 | 6% | 5,230 |
| 4 | Econ. Disadvantaged II (renter skewed) | 58.0% | 3.0 | 3% | 2,574 |
| 5 | Ethnic Enclave I (homeowner skewed) | 69.8% | 3.4 | 3% | 2,440 |
| 6 | Ethnic Enclave II (renter skewed) | 63.6% | 2.5 | 2% | 1,754 |
| 7 | Young/mobile/singles | 67.1% | 8.0 | 8% | 4,073 |
| 8 | Advantaged Homeowners | 83.2% | 26.8 | 26% | 16,506 |

Note: The mail return rate is the percentage of occupied housing units eligible to receive a mail form that returned a mail form.

[3] The algorithm converged in 9 iterations.

*Cluster 1: All Around Average I (homeowner skewed)*

This group has the largest number of occupied housing units and had the second highest mail return rate in 2000 (77.3%). Tracts in this cluster are close to average on every one of the hard-to-count variables. Around 28% of the housing units are not single-family structures, only one-quarter are renters, and slightly less than half (45%) are in non-spousal households.

In Cluster 1, unemployment, poverty, education and mobility levels are all close to national averages. The tracts are fairly representative of the national average racial breakouts but have above-average percentage of non-Hispanic whites (80%) slightly below-average Blacks (9%), 2% Asian or Native Hawaiian/Pacific Islander (NHPI) and 1% American Indian/Alaska Native (AIAN). Tracts in this cluster contain about 7% Hispanics which is well below the national average. Around one-quarter of the population is under age 18 and about 15% are over 65. This group is the largest cluster representing about 36.5 million occupied housing units (about 35% of the total). This cluster has the largest percentage of rural tracts (on average around 37% are rural[4]).

*Cluster 2 - All Around Average II (renter skewed)*

Cluster 2 is also somewhat unremarkable and "average" on most of the hard-to-count variables. About the only distinguishing characteristic is an above average number of households renting and in multi-units. This group of tracts is slightly more racially diverse than Cluster 1 (12% black, 11% Hispanic, and 69% non-Hispanic white) and is also much more urban and densely populated. However, like Cluster 1, this group is relatively large (represents around 16% of all occupied housing units). Tracts in this cluster had close to average mail return rates in 2000 (74.2%).

*Cluster 3 – Economically Disadvantaged I (homeowner skewed)*

This cluster reflects households that are economically disadvantaged, but not as much as Cluster 4. One noticeable difference is that this cluster has fewer renters than Cluster 4 (less than half rent – 46%). Nonetheless, households in these tracts have a high percentage in poverty, receiving public assistance, and adults without a high school education. Above average unemployment and non-spousal households are also characteristics of this cluster. Blacks comprise about one-half (49%) of the population in these tracts – the second largest black population next to Cluster 4. This cluster has above-average number of children (29% are younger than 18). This group represents about 6% of the total occupied housing units. The overwhelming majority of tracts in this cluster are urban (92% urban on average). This cluster had mail return rates in 2000 that were well below average (66.5%).

*Cluster 4 – Economically disadvantaged II (renter skewed)*

Cluster 4 had the lowest mail return rate of any group (58.0%). Close to three-quarters of the households in these tracts contain non-spousal renters in multi-units (especially 10+ units). These tracts also have the highest poverty, public assistance, and unemployment of any cluster. This cluster most closely resembles Cluster 3 but has far fewer homeowners (on average, 81% of households are renters). Like Cluster 3, this group contains a higher-than-average percentage of Blacks (54%) but also has an above-average percentage of Hispanics (21%). This cluster reflects the most urban of all clusters (99.9% urban on average). This cluster represents about 3% of the total occupied housing units.

*Cluster 5 – Ethnic enclave I (homeowner skewed)*

This cluster is characterized by above average crowding and poverty, public assistance, unemployment and low education. However it also contains a *below-average* percentage of non-spousal households and above-average percentage of children. It looks most like Cluster 6 with the following differences: lower occurrence of linguistic isolation, lower mobility, higher homeownership, and fewer Asians. This cluster is

[4] "Urban" is defined as housing units located within urbanized areas or urban clusters. Urban areas consist of areas containing 50,000 or more while urban clusters consist of areas containing at least 2,500 but less than 50,000. "Rural" consists of areas located outside of urban areas and urban clusters (U.S. Census Bureau, 2001).

also less urban and less densely populated than Cluster 6. This group is predominantly Hispanic (61%) with 24% non-Hispanic white, 8% black, and 5% Asian or NHPI. Tracts in this cluster had below average mail return rates averaging around (69.8%).

*Cluster 6 – Ethnic enclave II (renter skewed)*

Cluster 6 had the second lowest mailback rate at 63.6%. This cluster has above-average presence of children and is characterized by multi-unit structures with at least 10 units. This group is exclusively urban, the most densely populated of clusters, and characterized by crowded housing. On average, half of persons residing within this cluster lack high school degrees. These tracts are predominantly comprised of Hispanics (59%) and Asians (11%) with only 19% non-Hispanic white, 9% Black, and 1% AIAN.

This cluster contains tracts with high levels of linguistic isolation (on average, around 31%). In some tracts, this ranges as high as 79% of households where Spanish is spoken at home and no household member 14 or older speaks English very well. Likewise, other tracts have as high as 74% of households where an Asian/Pacific Islander language is spoken at home and no household member over 14 speak English very well. This group is overwhelmingly renters (75%). It also has high rates of poverty, unemployment, and public assistance. This is the smallest of the 8 clusters, representing only 2% of the total occupied housing units. As such, increases to response rates will yield a smaller number of actual mail forms compared to the other clusters.

*Cluster 7 – Young/single /mobiles*

This cluster had a similar mail return rate as Cluster 3 but looks very different. The overwhelming majority of households are non-spousal renters located in multi-units (especially structures with at least 10 units). The people in these tracts have higher than average education along with very high mobility. The tracts are densely populated and almost exclusively urban. These tracts have a below average percentage of children (17%). This cluster has a relatively high percent of group quarters (4%), possibly reflecting college campuses. These tracts probably include a disproportionate share of younger singles in school or just out of school and into the workforce for the first time. This cluster is racially diverse with above-average percentage Asian (7%) and the majority non-Hispanic white (59%) followed by black (17%). This group represents about 8% of the total occupied housing units.

*Cluster 8 – Advantaged homeowners*

The tracts in Cluster 8 had the highest mail back rate (83.2%) in 2000. As such, these tracts have a very low percentage of renters, few multi-units structures, very low levels of poverty and unemployment, low mobility, and few non-spousal households. This cluster is indicative of stable homeowners who reside in spousal-households in single-unit houses, about one-quarter of which are located in non-urban areas. This group of tracts is the least racially diverse of all clusters with 85% non-Hispanic white and only 4% black, 5% Hispanic, 4% Asian or NHPI and less than 1% AIAN. It is also the least densely populated cluster as measured by population per square mile. This group is the second largest behind Cluster 1 reflecting 26% of the total occupied housing units.

## 5. Summary and Discussion

The factor analysis revealed three constructs for the hard-to-count population that proved helpful in performing and interpreting the cluster analysis. The result was a market segmentation for planning the 2010 Census Communications Campaign.

The cluster analysis revealed eight distinct segments, each with varying levels of mail return behavior in 2000 and each with unique demographic, housing, and socioeconomic characteristics. Five of the clusters exhibit characteristics of the underlying factors uncovered in the earlier analysis (i.e., Economically Disadvantaged I and II; and Ethnic Enclave I and II; and Single/unattached/mobiles). These five clusters have below-average mail return rates in Census 2000 and together comprise 22% of occupied housing units. The mail return rate for the two All Around Average clusters had about average mail return rates and these two combine to cover 52% of the occupied housing units. The remaining cluster (Advantaged Homeowners) have 26% of the occupied housing units and an above average mail return rate.

In summary, the groups emerging from the cluster analysis present contrasting socioeconomic and demographic pictures according to propensity to mail back a census form in 2000. It is interesting to note that some of the clusters have very similar mail return rates and HTC scores yet look very different once we more closely examine the characteristics that compose the tracts – this is the type of detail that should help inform the communications contractor as they develop tailored media messages and delivery strategies.

It is also of interest that the clusters mirror in many ways the “stairstep” typology of household characteristics correlated with mail return documented by Word (1997). Word noted that in the 1990 Census, White, non-Hispanic owners in spousal households had the lowest non-mailback rate (13.2%) while Hispanic renters in non-spousal households had the highest non-mailback rate (64.3%). In keeping with this typology, our highest mail return cluster (the Advantaged Homeowners) had the highest percentage White population, lowest percentage of renters, and lowest percent of non-spousal households. In contrast, the cluster with the lowest mail return rate (Economically Disadvantaged II – Renter Skewed) had the lowest percentage of Whites, highest percentage of renters, and highest percent of non-spousal households.

Since the PDB is now seven years old, we plan additional micro-level analysis using American Community Survey data to validate and supplement the macro analysis as we approach the 2010 Census.

## Acknowledgments

The authors wish to express their gratitude to Marie Pees for providing various decennial Census data files necessary for the cluster and mail return rate analyses. Finally, thanks to our reviewers Pat Cantwell, Aref Dajani, and Gregg Robinson for very helpful comments on earlier versions.

## References

Baron, S. and Billia, D. (1999). Building a Surrogate for Predicting Census Participation.” A paper presented at the 54th Annual Conference for the American Association for Public Opinion Research, St. Petersberg, Florida, 1999.

Bruce, A. and Robinson, J.G. (2006). “Tract-Level Planning Database with Census 2000 Data.” U.S. Department of Commerce, U.S. Census Bureau: Washington, D.C. http://www.census.gov/procur/www/2010communications/library.html

Hillygus, D.S., Nie N., Prewitt K. and Pals, H. (2006). *The Hard Count: The Political and Social Challenges of Census Mobilization.* The Russell Sage Foundation: New York: New York.

Robinson, J.G., Johanson, C. and Bruce, A. (2007). “The Planning Database: Decennial Data for Historical, Real-time, and Prospective Analysis”, paper presented at the 2007 Joint Statistical Meetings, Salt Lake City, Utah.

SAS® Procedures Guide, Version 8. SAS Institute, Inc.: Cary, N.C.

U.S. Census Bureau (2001) “Technical Documentation: Census 2000 Summary File 1, Census 2000 Geographic Terms and Concepts”, pp. A-22.

U.S. Census Bureau (2000) **“**’Well Done, America!’ Nation Achieves 67 Percent Response Rate in Census 2000, Two Points Higher Than 1990”**.** Press Release dated September 19, 2000. U. S. Census Bureau. Washington, DC. http://www.census.gov/Press-Release/www/releases/archives/census_2000/000657.html

Word, D.L. (1997). “Who Responds/Who Doesn’t? Analyzing Variation in Mail Response Rates during the 1990 Census.” U.S. Census Bureau, Population Division Working Paper no. 19, July 1997.

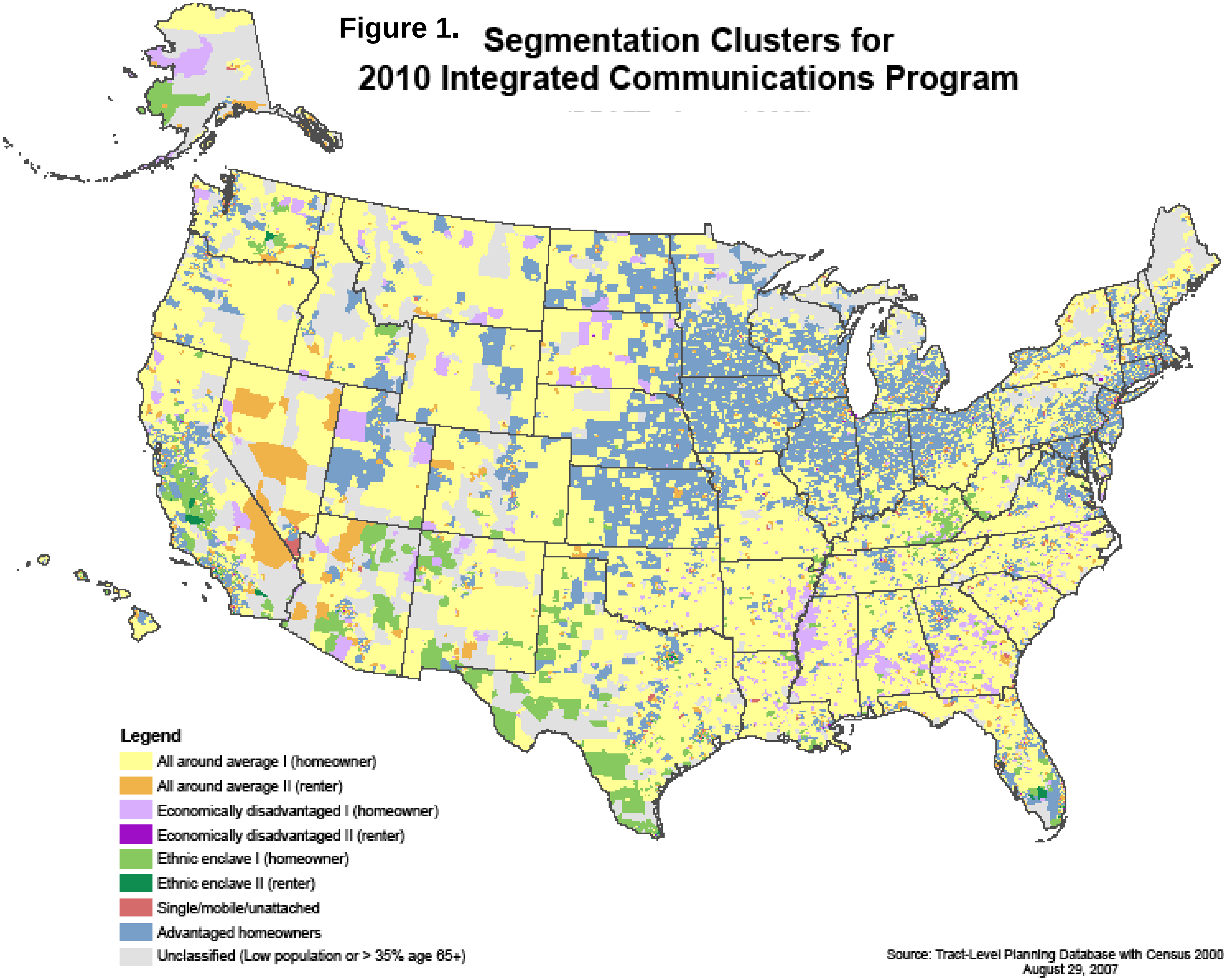
Legend
🟨 All around average I (homeowner)
🟧 All around average II (renter)
🟪 Economically disadvantaged I (homeowner)
🟪 Economically disadvantaged II (renter)
🟩 Ethnic enclave I (homeowner)
🟩 Ethnic enclave II (renter)
🟥 Single/mobile/unattached
🟦 Advantaged homeowners
⬜ Unclassified (Low population or > 35% age 65+)

**Figure 1.** Segmentation Clusters for 2010 Integrated Communications Program

Source: Tract-Level Planning Database with Census 2000 Data
August 29, 2007