

American Community Survey Summary File Comments Submitted by the Association of Public Data Users, Inc.

The Association of Public Data Users, Inc. (APDU) appreciates the opportunity to assist the Census Bureau in developing a Summary File from the American Community Survey (ACS) that meets the needs of the more sophisticated data user interested in processing large amounts of ACS data. APDU is an organization of about 70 data user, intermediary, and producing organizations and individuals joined by the common purpose of maintaining and improving the quality and usefulness of public data.

APDU's comments are categorized by the following topics: file structure/ technical comments; ACS products; and master summary file.

File Naming Convention

There are several problems with the file naming convention that make the files difficult to use.

- While the file names indicate the type of data in the file (estimate, margin of error, or standard error), they do this in a way that makes it difficult to work with one type of data. The indicator is 11th character in the file name. It would be much easier for data users to work with these files if this was earlier in the file name, after a standard portion.
- The file names all have the non-standard extension of "2005-1yr." This is problematic because some programs (e.g., SPSS) don't readily recognize non-standard extensions and there are 450 files that need to be renamed for each state. Is there an easy way to rename files in batch process?

A better naming convention would be tyyyyppggssss.txt where:

t = type of file (e=estimate, m=margin of error, s=standard error)

yyyy = reference year for the data, e.g., 2005

p = period covered by the file (1=1-year, 3=3-year, 5=5-year)

gg = geographic area (state or US) covered by the file

ssss = file segment number (0001 through 0150)

Technical Documentation

The technical documentation provided with the prototype file was of such poor quality that it made it nearly impossible to test the actual files. There are several specific problems that need to be noted here.

- The technical documentation does not contain the record layouts; instead it contains links to individual spreadsheets containing the layout for one table.
- There is nothing that indicates which tables are in which files.

Ideally, the technical documentation should contain all of the information included in the Census 2000 Summary File technical documentation. However, some chapters, such as chapter 6 – Summary Table Outlines, can be deleted since they duplicate the more detailed

information included in the data dictionary chapter. It is critical that the data dictionary identify which segment a particular table is located on. Having information in the technical documentation about how users should calculate error levels when aggregating data would be extremely useful to data users.

Additionally, since many data users use the record layouts to prepare computer code to read the summary files, it is critical that the record layouts, especially the variable/cell names and descriptive labels be available electronically in a single file.

File Structure

There are a number of issues related to general structure of the files included in the summary file package. This section discusses these particular issues.

Estimate, Margin of Error, and Standard Error files

The prototype summary file contains three different types of files – estimates, margin of error, and standard error files. One of the primary uses of the Summary File is to aggregate geographic areas or data cells together into larger areas or categories. This can more easily be done by working with a file containing just the estimates. Of course, when this is done, the error levels go down, but not in a way that allows them to be easily calculated at the same time as the estimates. **We believe that the best approach is to keep the estimate, margin of error, and standard error files separate from each other.**

Given the simple mathematical relationship between the margin of error and standard error, it is unclear why the Census Bureau is providing both of these measures in the summary file package. The type of user who is most likely to work with these files can easily calculate one measure from the other. Since more statistical calculations are based on the standard error than the margin of error, keeping just the standard error files would be more useful. Additionally, providing just one of these measures would reduce the number of individual files in the package from 450 to 300. **APDU recommends designing the final ACS Summary File product so that it contains the files reporting the standard errors and that the margins of error files be excluded from the product.**

Multi-year files

There has been discussion about whether the Census Bureau should create products that show the 1-year, 3-year, and 5-year estimates side by side. This may be a way for the Census Bureau to educate data users, especially the less experienced or sophisticated ones, about the reduction in the margins of error for these estimates as the periods covered are increased.

APDU believes that educating data users about the advantages of the multi-year estimates compared to the single-year estimates is critical. However, this is most likely best done through specially prepared educational materials containing sample comparison tables and graphs. Most of the data users working with the summary files will already understand this concept.

APDU recommends that there be separate summary files containing the 1-year, 3-year, and 5-year estimates.

Sample Code

The suggestion has been made that the Census Bureau provide sample code that users can duplicate or modify to combine the individual summary files into a single “master file.” Given that there are many different software programs used by the data community, it is impossible for the Census Bureau to provide code for all of these programs. Additionally, for the Bureau to provide code for some programs and not others would be a tacit endorsement for selected commercial programs. While this is a laudable service, the Census Bureau’s efforts are probably best spent on other activities that the Bureau is uniquely in a position to perform.

History has shown that the data user community routinely produces this code for many software programs and is more than willing to share this code with other users.

APDU recommends that the development of computer code to read the summary files be left to the data user community, but that the Census Bureau support this effort by producing a single file containing all of the data variables and descriptive labels.

Data Tables

While the Census Bureau may consider the design of the tables included in the ACS Summary File beyond the appropriate scope for these comments, APDU feels this is a critical part working with these files and takes this opportunity to comment on these.

Table complexity

The detailed tables on American FactFinder, which were replicated in the summary file, can be extremely difficult to use and are subject to large amounts of suppression. A prime example of this problem is the difficulty of calculating a simple unemployment rate in the current products independent of age and gender. Currently, the data user must aggregate 26 cells to calculate this extremely commonly used measure. Given the current suppression rules, there is a strong likelihood that the one table that can be used to calculate this is going to be suppressed. In the past, the Census Bureau provided tables that allowed the user to calculate this by working with just four cells. **The tables included in the ACS products (both the summary files and on American FactFinder) need to be redesigned with the common data uses in mind. This generally means that simple 1- and 2-way distributions (such as, employment status by gender) need to take precedence over 3- and 4-way tables (like employment status by age by gender).** Data users looking for the more detailed tables can prepare their own from the PUMS files.

Race, Ethnicity, and Ancestry Based Tables

Many data users are interested in the characteristics of very specific race, ethnicity, or ancestry groups. The current ACS tables do not allow these groups to be studied. In the past, products like the Census 2000 Summary File 4 met this need for many data users, by providing data on groups such as the Chinese or Hispanic subgroups like Mexicans. While users can use the PUMS file to get as these data, a product similar to the Census 2000 Summary File would be very useful to these data users. **APDU recommends that the Census Bureau develop an ACS product similar to the Census 2000 Summary File 4.** There are alternate ways of providing this data, but they lead to even more complicated data tables that most likely be suppressed under the current rules.

Summary

APDU believes that there should be a summary file product for the American Community Survey to meet the needs of sophisticated data users that cannot be satisfied with the current products available through American FactFinder. In our opinion, there are a number of strengths and weaknesses in the prototype ACS Summary File. In order to improve the quality of these summary files, APDU recommends that:

- The Census Bureau change the file naming convention to the one described above
- The technical documentation should contain all of the information included in the Census 2000 Summary File technical documentation. However, some chapters, such as chapter 6 – Summary Table Outlines, can be deleted since they duplicate the more detailed information included in the data dictionary chapter.
- It is critical that the data dictionary identify which segment a particular table is located on.
- Having guidance in the technical documentation on calculating error rates when aggregating would be extremely useful.
- It is critical that the record layouts, especially the variable/cell names and descriptive labels be available electronically in a single file.
- The best approach is to keep the estimate, margin of error, and standard error files separate from each other.
- There be separate summary files containing the 1-year, 3-year, and 5-year estimates.
- the development of computer code to read the summary files be left to the data user community
- The Census Bureau develop an ACS product similar to the Census 2000 Summary File 4
- The tables included in the ACS products (both the summary files and on American FactFinder) need to be redesigned with the common data uses in mind

APDU looks forward to working with the ACS staff towards a better product for all. If you have any questions about our recommendations, please feel free to contact our president, Leonard Gaines, at lgaines@empire.state.ny.us or (518) 292-5312.

Thank you for giving us the opportunity to comment on this product.