Introduction to the American Community Survey Public Use Microdata Files (PUMS) Files
Transcript
February 22, 2017
Javier Gomez
American Community Survey Office

This presentation was presented as a webinar to the general public on February 22, 2017 by Javier Gomez of the U.S. Census Bureau's American Community Survey Office. A link to the recording is available at https://census.webex.com/census/lsr.php?RCID=6bef72af851afd99b87ad8d04cf6a18d

The transcript of the webinar follows. Slide references and links have been added to the spoken text as appropriate.

Coordinator: Welcome everyone and thank you for standing by. At this time all participants will be on listen-only until the question and answer session of today's conference, at which time you may press Star 1 to ask a question. Today's conference is being recorded. If you have any objections, please disconnect at this time. I would now like to turn the meeting over to your host, Mr. Javier Gomez. Sir, you may begin.

Slide 1: Title Slide

Javier Gomez: Thank you. Good afternoon and welcome to this presentation of Introduction to the American Community Survey Public Use Microdata Files – PUMS. My name is Javier Gomez and I work for the Outreach and Education Branch, in the American Community Survey Office.

Slide 2: Outline

Before I start I would like to mention that this presentation is not an overview of the that American Community Survey. Hopefully you are already familiar with how the survey is collected and how to find the more than 1000 standard

pre-tabulated data on the American FactFinder. If not, I would absolutely recommend that you take a look at one of our Introduction to the ACS Webinars.

If you put Census Data products on a scale of easiest to use to the most complicated, the PUMS files will be near the most complicated end. Working with PUMS data generally involves downloading large data sets into a local computer and analyzing the data using statistical software such as R, SPSS, Stata or SAS, or becoming comfortable with using DataFerrett.

I just want to warn you that this presentation is not going to turn you into a PUMS expert user overnight. However, after walking you through an overview and explanation of the available PUMS Geographies, methods to access the files, some answers to common questions and, of course, introducing you to our many resources, I think you will be comfortable getting started with these files.

Slide 3: Why use PUMS?

If you are already familiar with the ACS program and its data, then you know that with the more than 11 billion estimates produced annually, we meet many of the needs of data users, but not quite all of them. The U.S. Census Bureau produces the Public Use Microdata Sample (PUMS) files precisely so that data users with needs that are not met by standard products, can conduct their own analysis. On the screen you can see some examples of why a data user might turn to PUMS and I'll give you a real-life user case.

Let's say that I have observed that in my metro area a lot of the jobs that are traditionally held by teenagers they now seemed to be filled by adults and seniors. Perhaps I want to investigate this issue in the data. The standard

tables produced by the Census get very detailed. There is for example, table B14005, which you can get Sex by School Enrollment by Educational Attainment by Employment Status for the Population 16 to 19 years. This is very detailed, but I still have questions that are not supported by this specific table.

Perhaps I want to know if 16 and 17 years old are employed at different rates than 18-year-olds. Or maybe I want to do a more detailed analysis. I might have a theory that teenagers who speak a language other than English are most likely to work while in school, so I can look to see if there is any correlation.

Or perhaps I want to test a theory that when all the adults in a household have a lower income, the teenager is more likely to work. So I can merge a person and housing unit file to create a variable that tells me whether there is a working teenager in the household and see if there is a correlation between that variable and the median household income. These are all interesting ideas to look into, but the data is not available in the standard products.

Slide 4: What are PUMS Files?

Before we go any further, let’s go over the definition first. Public Use implies that these files have been created specifically for the public. These files differ from the Survey Microdata: Identifying information has been removed, some categories have been modified, either the extreme values have been grouped together (Top and bottom coding) or categories are broader in general. However, public use also means that these files come with a suite of data user guidance and reference materials, and are provided to you at no cost.

Microdata: These are individual records of survey responses with identifying information removed. Users create the estimates, tables and Margins of Error

themselves. And Sample: These files do not include every record of every person who responded to the ACS. Only a select few that in turn, are representative of the population. This means that PUMS estimates will not exactly match the American FactFinder estimates, and that is something that I will talk about a little bit later in this presentation.

The ACS samples 3.5 million addresses per year. The 1-year ACS PUMS file contains about 1% of all of the US households. The 5-year ACS PUMS file is the equivalent of five 1-year files, so it includes about 5% of all of the US households.

Slide 5: Summary Data vs Microdata

Let’s look now at an example of this Microdata and what makes it different from the Summary Data that you may see in American FactFinder or other sources.

In aggregated tables, or summary data, the individual records are categorized and weighted to create an estimate for the larger population. For this example that the statisticians at the Census Bureau have taken the records of all respondents who live in Alaska, 25 years or older, male and with a Bachelor’s degree, group them together and weighted them to create an estimate of all the males 25 and over in Alaska who have a Bachelor’s degree. The statisticians have also calculated and provided the Margin of Error for these estimates.

By contrast, the Microdata provides a sample of the record the statisticians used. Here you can see that one person who responded to the ACS in Alaska is male, 52, and attained a Bachelor’s degree. To create an estimate, the data user must take these raw materials and do all the work the statisticians did in the example above.

Slide 6: Summary Data vs Microdata

Although the data user's work with this Microdata will not be easy, there are definitely benefits to using the Microdata. Let's work through some of the pros and cons of each of them.

Benefits: Summary File data is much easier to use; there are more than a thousand tables that have already been designed and produced with the Margins of Error. Summary File data is also available for various small geographies, which is something that I will also talk about in a moment.

The Microdata on the other side requires significant work, but the detail available is amazing The person-level files include about 250 variables per record and the housing unit file includes about 200 variables. The files also include many useful constructed variables, such as property status, subfamily identification, etc.

Limitations: The summary data includes a lot of tables, but not every need can be met by the existing categories and topic combinations. The Microdata is complex, it's a smaller sample and has fewer geographies. The Microdata also has collapsed codes, that is broader categories, for some variables – for example, race, Hispanic origin, ancestry, place of birth, etc. Also, the Microdata has no geographies smaller than what we call PUMAs – also, I'll get into more detail into that topic in a moment.

Slide 7: PUMS Availability

One final note about the difference between Summary Data and the Microdata – we release the Microdata a little bit later than the Summary Data. Typically, the newest ACS Summary Data is available one month before the PUMS

Microdata. This means that the 1-year PUMS is released in October, one month after the September Summary Data release, and the 5-year PUMS is released in January, one month after the December Summary Data release.

However, you can prepare yourself to process these files early as we begin releasing the documentation you will need about one week before the release.

Slide 8: Multiyear (5-year) PUMS Files

I mentioned earlier that the 5-year PUMS is the equivalent of five 1-year files, so it includes about 5% of all the US households. So the question is, why would you wait until January for the 5-year PUMS when you could just merge the five most recent 1-year files in October? The short answer to this is yes, you can merge the files on your own if you want to, but we do some nice standardizations for the 5-year PUMS that might be worth waiting for. For example, new weights are produced for these records so that the weighted population matches the latest population estimates. Dollar amounts have an adjustment factor to standardize them to the latest year in the multi-year file, so that no one is comparing apples to oranges. Other coding schemes are updated, for example, ancestry, patient, industry and place of birth, with lots of categories.

Multi-year file takes the lowest-common denominator approach, so that all years are coded the same.

Slide 9: 2-Dataset

Let’s now discuss PUMS geography.

Slide 10: 3-Geographies

To ensure that confidentiality of ACS respondents, the Census Bureau has to balance geographic details with details in the data. As I mentioned before, there are more than 250 variables on a PUMS personal record. This means that we cannot identify as many small geographies in the PUMS as users might hope. We can put the region, division and the state on the file, but the only other geography is something called Public Use Microdata Area – PUMA.

PUMS is not designed for statistical analysis of small geographic areas, but the PUMA can still be used for focus analysis in planning of cities of over 100,000 in population and many metro areas. For example, Baltimore City, with a population of over a million, is subdivided into six separate PUMAS. However, not all places will be easy to analyze.

Slide 11: Public Use Microdata Area (PUMA)

As I mentioned before, PUMA is an area where the population of over 100,000 in – population of over 100,000, large enough to meet the disclosure avoidance requirements. It is identified by a five-digit code that is unique within each estate, and they nest within state or state equivalence.

PUMAs are redefined after each Decennial Census. In most states, the State Data Center defines the boundaries of the PUMAs, but in some states the Census Bureau Regional Geography staff defined the PUMAs.

It is important to mention that PUMAs redefined after the 2010 Census were first used in 2012 ACS PUMS files. Multi-year files contain dual PUMAs vintages, for example, the 2010 to 2014 ACS PUMS files.

PUMAs are built on Census Tracks and Counties, and can be combined to create rough approximations of towns, counties or cities for analysis.

When you look at PUMA maps on the Census Web site, you need to know that there are two types of PUMAs mentioned: 1% PUMAs, also called Super PUMAs, and 5% PUMAs. The PUMAs used by the ACS PUMS is the same as the 5% PUMAs used by the Decennial Census PUMS.

Slide 12: Public Use Microdata area (PUMA) Maps

As with many geographic concepts, seeing PUMAs on the map may help you to understand them better. These two maps are shown at the same scale. You can see that most of the PUMAs in Wyoming, with a population of almost 600,000, are larger than PUMAs in New Jersey, with a population of almost 9 million. The point here is that PUMAs are really built on population and not geographic size.

The Missouri Census Data Center has a fantastic Geographical Correspondence engine that can match PUMAs to other geographies of interest. You can also use the links at the bottom of this slide to explore the PUMAs in and around your geography of interest.

Slide 13: Outline

Now let’s move to how does a user get to the PUMS file.

Slide 14: American FactFinder

The Census Bureau’s American FactFinder offers the PUMS files in both SAS and CSV formats. You can find the PUMS files at the American FactFinder’s site. You can see the Web site address at the bottom of this slide.

Once there, select advanced search from the top menu, click on topics on the left panel, and then look for product types, which you will find Public Use Microdata Sample.

Slide 15: American FactFinder

Here you will find the different PUMS data sets available to you. Choose a data set and format. As I mentioned, we provide them in SAS format and CSV format.

Slide 16: American FactFinder (cont'd)

Choose US or a state population or housing unit. Open the zip file, and I always recommend that you check first the ReadMe file. And then, open your PUMS file.

Slide 17: Census Bureau FTP Site

You can also download the files through the Census Bureau's FTP site. If you look at the Web address at the bottom of this slide you will land at a Web site looks like that screenshot on the left. Choose the data year and file. You will note as you can see on the screenshot on the right, that in this case the files have a "_h" and a "_p" suffix. The "_h" designates the housing units file, and the "_p" designates a person file. The last two letters of the file name are the state abbreviation.

Slide 18: DataFerrett

If you do not have any statistical software you can also use the Census Bureau's DataFerrett software application. These tools searches and retrieves

PUMS data. It can recode variables and can create complex calculations. DataFerrett also allows you to download only specific variables, a benefit if you don't have the space for a huge file.

Slide 19: DataFerrett Assistance

If you need additional assistance working with DataFerrett, we recommend the resources on dataferrett.census.gov. However, we also have produced a couple of video tutorials hosted on the ACS Web site, and step-by-step guidance on the "What PUMS Data Users Need to Know" Compass Handbook.

Slide 20: Outline

On the next few slides, I am going to go through some of the most common questions that we get from data users.

Slide 21: How do I put PUMS files together?

One of the most common questions is how do I put PUMS files together? Now that you are ready to download a PUMS file, you should know that PUMS files containing data for the entire United States, in contrast to individual states, are separated into multiple data files.

For example, as you can see on the top screenshot, if you download the 2011-2015 ACS 5-year PUMS files of United States Population records, you will note an "a", "b", "c", and "d" files. These files contain about one-forth of the population records in the particular data sets. These files are so large that we have to release them in pieces and the users then have to concatenate.

In other cases, some data users will need to use the household and person items together. For example, in order to analyze how the number of rooms in a home varies by a person's age, the merging of the household and person files will be required. This merger must rely on the serial number (SERIALNO) variable, which is the same in the household and person files.

On the screenshot on the bottom-right, you will find two easy commands to concatenate files, and to combine population and housing files by serial number. You can find instructions on how to concatenate and merge files on the PUMS Read Me file.

After combining the records, limiting the number of records you are processing by selecting those of interest will often increase processing speed.

Slide 22:Which weight should I apply?

Most users know that they need to weigh the data and statistical packages usually have commands to apply the weight, but knowing which weight to use can be tricky. First of all, what is a weight? A weight determines how many people or households are represented in a one-sample case. The ACS PUMS have complex weighting methodologies as described on the Accuracy of the PUMS document, but I am going to give you a quick overview.

The PUMS assigns an initial weight to each housing unit address record. The PUMS initial weight is that ACS full sample final weight times the sampling interval. Then PUMS weights are ratio-estimated to agree with ACS for a few characteristics. Those characteristics are: persons in household by sex by PUMA, housing units by vacant/occupied by PUMA, persons in Group Quarters by institutional/non-institutional by State.

Finally, the weights are rounded to integers. Replicate weights, those numbered 1 to 80, are used for calculating the standard errors.

Slide 23: Why don’t my PUMS estimates match AFF?

Why don’t my PUMS estimates much AFF? There are several reasons why a PUMS estimates may not match AFF. The PUMS is a sample of the full ACS responses. The difference in sample size in a year is about two million cases, which will cause the estimates not too match exactly. The PUMS files have some extreme values grouped together, where we use top and bottom coding, and the categories are broader in general.

I would like to point now that the purpose of PUMS is to create estimates that are not available on AFF (American FactFinder), not to recreate AFF. If your goal is to check your work, we do have files called “PUMS Estimates for User Verification” that data users can use to see how we have created estimates using the PUMS. In a few minutes in this presentation I will walk you through an example on how to use the Estimates for User Verification files.

Slide 24: How Do I Use Dual Vintage PUMAs?

The multi-year PUMS files with years before 2012 and after 2012 have both the Census 2000 PUMAs and the Census 2010 PUMAs. Unfortunately, this does not mean that each record has both PUMA codes. Rather records from data years before 2012 have the Census 2000 PUMA codes, and records from 2012 and later have the Census 2010 PUMA codes.

The first question that probably comes to your mind is, why? The reason is, it’s a disclosure risk, because the data users could potentially take an old PUMA and a new PUMA that barely overlap and figure out which cases are in

that sliver. In that case, the user will now have geographic detail and case detail. There are three solutions to this problem: One, wait until the 2012-2016 5-year file, which comes in 2018. Two, use State-level estimates. Or three, if you can be comfortable with geographies that are not an exact match, which we call fuzzy boundaries. If this is your case – if you can be comfortable with fuzzy boundaries, then determine which 2000 and 2010 PUMAs most closely approximate your area. The next step would be look at the pre-2012 records in the 2000 PUMAs. Step number three, look at the 2012 and later records in the 2010 PUMAs. And finally, add these two figures.

Slide 25: Use Caution…

Some final speedbumps that you might run into. Three million records sounds like a lot, but when you start splicing it by filtering variables and PUMAs, you may quickly narrow your focus to too few cases. Please make sure that you are calculating Margins of Error so you have a sense of the reliability of the estimates. When the number of unweighted cases is too small you can also consider adding cases from neighboring geographies, broadening categories or creating multi-year files.

Extreme values have been grouped together. PUMS is not the data set to use to determine how many millionaires, mansions, centenarians, etc live in your state. Likewise, it is unwise to assume things like “no one spends more than…” when your value is on the extreme end of a range. That said, we do provide the Top and Bottom codes for each PUMS file, so you will know that people older than 94 in Florida for example, are not in fact, lying about their age.

Slide 26: Outline

Next I want to point out where you can find some additional resources about PUMS.

Slide 27: ACS Main Page

The American Community Survey Web site has a lot of information about the program, data products, including PUMS, and helpful information and tools for data users. The left navigation on this page will get you to the sections of interest. Today I want to point out two sections: Data, and Technical Documentation.

Slide 28: PUMS Data Page

If you click on Data you will see an option for PUMS data. This page includes links to all of the PUMS data available. A note here: the 2000-2004 data does not include every state, only a selection. And the 1996-1998 PUMS files are only available on DVD upon request.

Slide 29: PUMS Technical Documentation

If you click on Technical Documentation and then on PUMS Documentation, you will find general information about PUMS, confidentiality, Frequently Asked Questions, file structure and DataFerrett. On the PUMS Technical Documentation page, the URL at the bottom of this slide, you will find the resources listed on the slide for each data year released.

Slide 30: PUMS Data Dictionary

The data dictionary is where you will find a list of variables, definitions and values. The dictionary is very helpful, even for those variables that seem intuitive. For example, you may correctly assume that that two-digit code

under AGEP is years of age, but you may not be sure what AGE=0 means, which is actually under one year.

I would also like to mention that PUMS provides Estimates for User Verification files. These files contain PUMS estimate for selected housing and population.

Live Demo: User Verification Files with DataFerrett

It is as this point that I want to switch to a live demo and show you how to work with the Estimates for User Verification files using DataFerrett. What I am going to do is get out of this presentation for a moment. One second please.

So this is the ACS Web site as I was talking about in a moment. If you go to Technical Documentation you will see PUMS Documentation and Technical Documentation. You will see that it is divided by the data, by the years, so in 2015 you will find here, on the bottom right PUMS Estimates for User Verification in three different formats, SAS, LST or CSV, and also for 5-year estimates and 1-year Estimates.

What I am going to do is open the CSV format for the 2011-2015 5-year Estimates, PUMS Estimates. Actually, I have already went ahead and downloaded it and I opened it with Excel. As you see here, we have a selection of different estimates for population and housing for the entire nation and then the states.

In this way you can work with PUMS files and check your work and see if you are actually doing the estimates the correct way. What I'm going to do here is just take an example, let's say population for the entire nation, and

broken down by age – 0 to 4, 5 to 9 and 10 to 14 – just for this example. And I'm going to do it using DataFerrett.

If you go to DataFerrett's Web site, you will see the button here on the right, "Launch DataFerrett", which will open a new window that will look like this. On the left, you will see all the different data sets available. You see American Community Survey and then you see the Public Use Microdata Sample, divided by 5-year Estimates or 1-year Estimates. Since we are looking at 2011-2015, we are going to use 5-year Estimates 2011-2015. Click on it and view variables.

Keep in mind that we are trying to estimate a population for the nation so what we need is to pick Geographies and to pick the Population variables. After selecting those two, I am going to click on such variables. If you are not sure which variables you need to do your calculations, I always recommend that you look at the PUMS Data Dictionary. This is where you will find the full list of variables and values included within each of the variables. Going back to DataFerrett, I am going to start filtering this by adding my first parameter, Geography. I am going to click on it and then "Select Highlighted Variable".

Since I am trying to estimate the population for the entire nation, I am going to select all of the States. I am going to select all of them and then just drag them to the right, and then finish.

Next, as we are trying to estimate population by age, I need to select the Age variable. I click on it and "Select Highlighted Variable". One difference between AFF and PUMS is that AFF has pre-tabulated tables, so values are already grouped together. In PUMS you will see all the difference responses

that we got, so what we want to do here is select all the years that – all the values that come within this variable.

These are pretty much the only two variables that we need, so we are going to go to step two, Data sets/Make a table. Because we are trying to estimate population by age, we need to click this variable, age, and recode it. That means we group some values together. So as you noted, I click on the variable and then click on recode variable.

This new window allows you to group together some of the values and you can also rename the newly created variable. What I am going to do is name this “Age Groups” and then start to group together some values. For example, from 0 to 4, hit “recode” and you can also rename this as “0 to 4”. And I’m going to continue the same process by selecting “5 to 9”, hit recode and then “10 to 14” and the same, hit recode and I’m going to rename this.

I could continue going on and grouping together more values, but this is all I need for the sake of this example. What I am going to do here is just rename the rest of the values and say “15 or over”. Click “Ok” and now we have created recoded variables. From here we click on “Make a Table”. Something important to mention is that PUMS, DataFerrett, I am sorry, automatically brings the proper weight for this estimates that you can see here, person’s weight has been added. As we click “Ok” a new window will appear with a spreadsheet similar to Excel and on the right we have the variables that we have selected. What I am going to do is select the State, drag them to the spreadsheet and then I am going to take that recoded variable, the one that we created “Age Groups”, and I am going to drag it to the spreadsheet as well.

You see now we have all the States in this row, row number four. We have the total, which will mean for the entire Nation, but the data is not there yet. It

will not come until you click on that green button “Go Get Data”. And so now we have the estimates for the entire Nation and even for all of the States. This one here will get a total population for the Nation and this is broken down by Age Groups, 0 to 4, 5 to 9, 10 to 14 and 15 or older. And you can check this values against the User Verification Files that we have here. And now you can confirm that in fact we are doing the right calculations here.

Slide 31: Source Us!

After this, I am going to go back to my presentation. Finally, to wrap up this presentation I want to say please help us to reach new users by sourcing us. Also, I would like to encourage you to connect with us. You can sign up and manage alerts on the ACS website via GovDelivery. Visit our website or connect on the various social media platforms using the hashtag ACSData. You can also email acso.users.support@census.gov with any questions you may have.

Slide 32: Continue the Conversation #ACSData

We want to remind you that there is a Data Users Group specifically for users of the ACS data. The ACS data user group was formed in partnership with Population Reference Bureau. It is a great way to learn from your peers and how to use the ACS data for all kinds of applications. Go to acsdatausers.org to learn more, including how to sign up to be one of the over 1800 users in the ACS Online Community. There is even a group specifically for using the ACS PUMS files.

Slide 33: American Community Survey Data Users Group

Also, I want to mention that we will have a 2017 ACS Data Users Conference, which will take place in May 11 and 12, 2017 at the Patent and Trademark

Office in Alexandria, Virginia. Registration for this conference will be open next week.

As a reminder, we have Data Dissemination Specialists throughout the Country who can provide data workshops locally. If you are interested in a workshop please contact CLMSO at, I am sorry at census.askdata@census.gov.

Slide 34: Questions?

Thank you very much for listening to this presentation. At this point I can take, we can take any questions that you may have. Also, I am providing on the screen the email where you can send us questions that you might have.

Questions & Answers Section

Coordinator: Thank you. We will now begin the question and answer session. If you would like to ask a question please press Star 1. You will be prompted to record your name. Please be sure to unmute your phone. Once again, if you would like to ask a question press Star 1 and we will pause for just a moment to allow those questions to start coming through.

And our first question comes from (name removed). Your line is open.

(Question #1): Hi, I wanted to ask whether PUMAs completely cover the country. I know that, you know, for instance, tracks did not completely cover the country until about 1990, I think. And I am wondering if PUMAs completely cover the country. And also, whether they conform to stateliness or where they can cross over stateliness the way MSA boundaries do.

(Sirius Fuller): Hello, my name is Sirius, I work on the PUMS data with Javier at Census Bureau. The answer is, yes, the PUMAs boundaries do cover the entire nation. There are no gaps, and they are contained within state boundaries. They do not cross state boundaries. And if you're looking at PUMA codes you need to merge them with the state codes because you could have the same PUMA code in two different states, but they represent two different geographies.

(Question #1): Great, thank you.

(Sirius Fuller): Thanks.

Coordinator: And our next question comes from (name removed). Your line is open.

(Question #2): Thank you, Javier. A couple related questions if I may. My kind of basic understanding of PUMS is that, is it not essentially a sample of a sample? It is a sample of the ACS sample. Is that more or less correct?

(Sirius Fuller): Yes, so the ACS is a sample of the population and PUMS is a sub-sample, but it, so as Javier said, it has some things to correct like confidentiality, but the weights will make it representative of the United States.

(Question #2): Right, so the correct term, I guess, it sounds like it is a sub-sample of the bigger sample?

(Sirius Fuller): Yes. Yes.

(Question #2): And then from there, and I am just trying to get a sense of things, so if I take one PUMA boundary, and let's just say it has an even 100,000 population – if I overlay that same boundary into the ACS file, how, in a ratio terms, what

percentage is that sub-sample of the sample? In other words, is it 50% of the ACS? Or is it more or less than that?

(Sirius Fuller): It depends. On AFF there are unweighted counts by different Geographies and there are PUMAs for ACS and then you can use the PUMS file to just do an unweighted count by PUMA if you wanted to compare.

(Question #2): Well I am just trying to get a big picture understanding to what degree is it a sub-set, generally speaking. Is it most of the ACS sample or is it just a little bit of it? Or somewhere in the middle? I think you said it was like two-thirds.

(Sirius Fuller): It depends if you are looking at weighted or un-weighted. Go ahead, Javier.

(Question #2): Or are you saying the weighting kind of takes care of that…

(Javier Gomez): You mean the sample?

(Question #2): Pardon?

(Javier Gomez): You mean the size of the sub-sample?

(Question #2): One year, five year…

(Javier Gomez): So the ACS takes a sample of 3.5 million addresses of the entire nation. The PUMS on the other hand has a 1% of all the population of the country on a 1-year release. If you look at 5-year ACS PUMS files, what we do is we aggregate by 1-year sample size so that will account for 5% of the US population.

(Question #2): Which I guess is very similar to ACS?

(Sirius Fuller): Yes, the US is designed – sorry, excuse me – the ACS is designed to be roughly a 2.5% sample. So like sort of a ballpark ratio.

(Question #2): Oh okay, so that would be about half then?

(Javier Gomez): Yes.

(Question #2): Okay, thank you. That helps. And then a final question, so I tend to work in larger scales, smaller territories. So I deal with block – ACS Data Block Groups and Tracks. The Margins of Error can be, you know, very very high. Within one PUMS what type of margin of error are we looking at? Is it relatively high or lower to working with say block groups or tracks? All other things equal.

(Sirius Fuller): That's a broad question. Do you want to answer?

(Question #2): It's not a fair question?

(Sirius Fuller): No, it is a fair question. So it depends, so the ratio or the margin of error to the estimate is, I assume, what you're looking at. So if you're looking at small geographic areas like the blocks or tracks group for ACS, you can get large Margin of Error for the Estimates.

(Question #2): Right.

(Sirius Fuller): If you are looking at rare populations you can also have that. So if you're looking at PUMA levels the Margin of Error should not be too bad. If you are looking at some of the like total population, but if you are looking at, you know, native Hawaiians in Vermont, you can have large…

(Question #2): …very high.

(Sirius Fuller): Right, it could be relatively high. You know, it depends on…

(Question #2): So in some cases it might be higher and some cases lower than the ACS because it – those pre-defined tables only slice and dice so much.

(Sirius Fuller): Right. If you slice down to very rare estimate, you know, characteristic then…

(Question #2): Well let us just say we are looking at something very simple total population within a given one PUMA territory. Is that going to be higher – a lower margin of error than a block group out of the ACS total population?

(Sirius Fuller): Relative the estimate will be lower. It will be because it is a larger sample size.

(Question #2): Okay.

(Sirius Fuller): The Block Group, in general, will have the largest Margin of Error relative to estimate. So that is all things, as you said, being equal.

(Question #2): Right. Okay, I appreciate the information. Thank you.

(Sirius Fuller): You are welcome.

Coordinator: And our next question comes from (name removed). Your line is open.

(Question #3): Hi, thank you. My question is regarding the extremes being masked, particularly with regard to income levels. Is there a documentation on what the thresholds are for masking or is that identified somehow in the outputs?

Javier Gomez: I am sorry, can you please repeat the question?

(Question #3): Yes, the question – your presentation talked about extremes – the extremes that are masked. So that, for example, you cannot find out all the number of millionaires, I think, was the example you used.

Javier Gomez: Yes, top and bottom.

(Question #3): Yes, top and bottom. So I am wondering what the threshold is for masking. At what point are those Data masked? So if we are looking for poverty level residents in a PUMA, would that be masked because they are – that is an extreme?

(Sirius Fuller): Yes, so not all variables would have top or bottom coding. There is a list of the variable which have top and bottom coding, but it's not going to be that the data is not available. It would just be, you know, income over a certain amount will all be assigned the same value. Or age over, you know, a certain value will always be assigned to the top-coded value. And then top and bottom coding file will tell you the variable – which variable it is and what the – you know, over what value will be top coted and what the top coted value is.

(Question #3): Okay.

(Sirius Fuller): Does that make sense?

(Question #3): Thank you.

Coordinator: Our next question comes from (name removed). Your line is open.

(Question #4): Thank you very much. I appreciated the presentation. I have what I think is a very basic question. I'm a grant writer for an agency and we serve an age range of 18 to 24, but rarely does the Census data – is it divided into that 18 to 24. Usually it is 20 to 24 and then, you know, 16 to 19 or something like that. So would – am I correct in understanding that I could use this PUMS – this PUMS to figure out the data for 18 and 19 year olds and then combine that with the data that's already there for the 20 to 24 year olds?

(Sirius Fuller): Well, yes, you have to be careful – you do not want to combine PUMS and ACS data, but you could create your estimate for 18 to 24 year olds using the PUMS data.

(Question #4): So do not combine PUMS with the ACS? Because I think earlier, you were saying, you had some reason there, but I could use it for 18 to 24 – now this is for – the city of Detroit. Would that be prohibitive as far as my gathering all this data – from your examples it looked like I am looking at individual answers to the Census data. So would it be realistic for me to even try to do this for an area like the city of Detroit?

(Sirius Fuller): Probably – yes, the PUMAs boundaries – you could probably get a good approximation of Detroit. I would also, we do not represent the grant writers and so this is, I would double check with the grant writers to see if, you know, how they feel about using the PUMS data versus using the ACS data. You know, I do not want to mislead you, but you could definitely use the PUMS data and for a large geographic region like Detroit you should be able to get close geographic boundaries.

(Question #4): Did you say it would be reasonably accurate?

(Sirius Fuller): Yes, it should be.

(Question #4): Okay, thank you very much.

(Sirius Fuller): You are welcome.

Coordinator: Once again, if you would like to ask a question please press Star 1 and record your name.

(Javier Gomez): I would like to mention at this point that we have a brief survey after the question and answers. If you could please complete it after this presentation, that would be highly appreciated.

Coordinator: Once again, if you would like to ask a question please press Star 1 and record your name. We had another question pop in. Just a moment. And, (name remoevd), your line is open.

(Question #5): Yes, this is (name removed). Why aren't MOEs, Margin of Errors, calculated and delivered via DataFerrett?

Javier Gomez: I am sorry – your question, why are the margins of error not calculated on DataFerrett?

(Question #5): Yes.

(Sirius Fuller): So they can be. You can select the different replicate weights and use them to calculate margin of error. If you look at the PUMS Accuracy of the Data there is also a method to approximate the Margin of Error with design factor methodology. I think they might be working on adding it in the future. I am not exactly sure. You can – yes, it may not be – I know it's a desire to add that

to the future, but I don't – you might have to contact them to see where it is in the process of if they're – or if there's resources available to do that.

(Question #5): Thank you. I just want to voice my approval if that ever gets put in.

(Sirius Fuller): Thanks.

Coordinator: We have another question from (name removed). Your line is open.

(Question #6): Thanks again. I have in my notes here that you mentioned a handbook earlier on in the Webinar. What PUMS Users Need to Know. Can you tell us again how we can get that handbook?

Javier Gomez: Yes.

Coordinator: Our next question comes from (name removed). Your line is open.

(Question #7): Hi, yes, my question is about computing the standard error using the design factor method. The variable in size of geographic area – is that number going to come from the table that I would make out of DataFerrett? Like the total number of responses that I get. Do I take that number from there or would I take it from the documentation that provides the number of people in the area? I forgot what you called that – the estimates for verification.

(Javier Gomez): Let's answer this question and then go back to the document from the question before.

(Sirius Fuller): Okay. So you can – the N will be the total number of persons or households or housing units, depending on what you are calculating and I believe all three

are in the estimates for User Verifications, but you can also calculate that right from the PUMS data.

(Question #7): Okay, so if I am creating a table of employed people in a certain county, that number – the table that I make in DataFerrit gives me a total. So would I use that number or I need to get it from the estimates for user verification?

(Sirius Fuller): It would be the total – it would not be the total of your estimates for N, it would be the total of the Geography.

(Question #7): Okay, perfect. And so if it's for a PUMA area, how do I – or a combined area – a combined number of PUMAS.

(Sirius Fuller): Right, then you would have to calculate using the PUMS data because it's not published in the estimate for user verification.

(Question #7): Okay, okay, I hear what you are saying. So then I need to go and find out how many – get the estimate of how many people are in that PUMA area? Right?

(Sirius Fuller): Correct. Yes, in addition to your estimate of unemployed or whatever you are doing.

(Question #7): Okay. Wonderful, thank you so much.

(Javier Gomez): Before we take any other questions, let me answer to the user who asked before about "What PUMS Data Users Need to Know". You can find it if you go to the ACS Web site, go to library and Educational Materials. Yes, Educational Materials. Scroll down and you will see the handbooks for the data users and then right here you find "What Public Use Microdata Sample

Data Users Need to Know". Once again, ACS website, Library, Educational Materials.

Coordinator: Are you ready for the next question?

(Javier Gomez): Yes.

Coordinator: Okay, our next question comes from (name removed). Your line is open.

(Question #8): Thank you for taking my question. I got two very short questions. The first one is about the poverty variable in the PUMS files. The poverty variable contains some missing values. Is that missing due to the respondents and non-respondents or is it due to skips and is there any strategies about inputting the missing values for the poverty variable?

(Sirius Fuller): So the poverty variable – the (unintelligible) would be missing variables if there is no – I believe that is poverty in relation to income. I think it is.

(Question #8): Yes.

(Sirius Fuller): There are missing variables because they are people who are not in the universe for that variable. So it is okay that there are missing values and those should not be in the estimate for poverty when you are calculating that.

(Question #8): Okay. All right, the other question is that we – we are currently using the PUMS – trying to use the PUMS to do some analysis with the congressional district areas. Even though the congressional district data is available in the American FactFinder, but the congressional district identifier is not in the PUMS. I was just – I asked the question separately via private emails to you guys. I do not know which one of you answered, but I was just wondering if

the congressional district identifier can be added to the PUMS since it covers a larger areas than the PUMAS currently available in the PUMS.

(Sirius Fuller): Yes – no, it will not be – only PUMA areas under sub-state will be provided because even though the Congressional District is larger, as Javier said earlier, there might be a sliver of geography that could be created because of the differences. And so you can – you'll have to use the PUMAs to approximate the Congressional District and I think you mentioned earlier, Javier – go ahead.

(Javier Gomez): Yes, as I mentioned before, the Missouri Data Center has a really good engine that can approximate different geographies using the PUMAs. That's an option that you might want to take a look at.

(Question #8): Yes, that is one way I have been using it so far, but I was just wondering if there's kind of an easy to use identifier – Congressional District identifier in the PUMS rather than going into that sort of process of converting PUMA to congressional districts. But, anyway, thank you for your answers.

(Javier Gomez): Thank you.

Coordinator: Once again, if you would like to ask a question, please press Star 1 and record your name. It looks like we have no more questions coming through at this time.

(Javier Gomez0: All right. Thank you very much.

Coordinator: And does that conclude the conference for today?

(Javier Gomez): Yes, please. Thank you everyone and thank you for your help.

Coordinator: Thank you. That does conclude today's conference. Thank you for participating. You may now disconnect.

END