**Session 1 – Tools of the Trade: PUMS**

**Leader: Rob Pitingolo**

Present: Rob Pitingolo, Laura Simmons, Vicki Mack, Denise Linn, Jennifer Newcomer, Mary Buchanan, Lisa Pittman, Megan Swindal, Maia Woluchem, Nic Moe

R. Pitingolo - I have a lot to offer - I use PUMS to run a lot of analyses and it's one of my specialties

L. Simmons - I used it for my masters thesis…

V. Mack - It's been a couple years since I’ve used it.

D. Linn - I'm here to learn!

J. Newcomer - I've used it a fair amount, particularly recently. But I'm always nervous with using the margins [of error]. So just here to get into the mechanics of using the margins.

M. Buchanan - I've used it a little bit but using it quite a bit recently

L. Pittman - Never used it but I've asked my part-time people to start using it. I could always go to the tutorials but I'd love if you could talk about it as well.

M. Swindal - I'll parrot Denise. Used it in grad school and I'm rusty.

R. Pitingolo - There's lots of PUMS so let's talk about ACS. PUMS just take all of the individual records and just putting them into one file. I've never gotten the ACS form but in the decennial census, you're writing all sorts of things. The PUMS is that data in its simplest form. Each person is one row. Each column is the info about that person. Our D.O.B and address, and since the ACS has so many questions, each question is the column. It's fairly simplistic in that sense.

M. Buchanan - Is there a way to connect the population and housing PUMS?

R. Pitingolo - I use IPUMS ([https://www.ipums.org](https://www.ipums.org/)), the same people who create the historic census file. Jeff Matson works at the same university as those people but he has nearly no interaction with them. They have data for both housing units and the people in those housing units. To get at those groups, PUMS gives you a lot of different weight sets, and they have a bunch of technical variables. It's all in the same data file. How you run your analyses depend on whether you want to uses the houses or the people. What's something that people want to do?

L. Pittman - I understood that it's de-identified? But I imagine it probably has the same problems as block group data?

R. Pitingolo - I feel like the reason why it's not been a topic yet is that the data is not at the neighborhood level, it's at the PUMA level. A PUMA (<https://www.census.gov/geo/reference/puma.html>) is an area containing 100,000 people. So in Western New York, sometimes four counties equal one PUMA. So it turns a bunch of people off at first. If I have a county level definition, sometimes the counties and the PUMAs don't match

M. Woluchem - Occasionally outside parties will publish crosswalks between PUMAs and traditional geographies like tracts or counties.

V. Mack - There are some certain variables that you can't find on FactFinder. You can do all sorts of cross-tabs here that you can do only with PUMS but not with ACS data.

R. Pitingolo - It could be useful for a citywide analysis. Most cities in NNIP are more than 100,000 people so it could be helpful and the cross-tabs can also be helpful. But yes, it does have the same problems as ACS because it's a survey so there are lots of margins of error. A lot of times they are large and sometimes your estimate will not tell you that much. The more cross tabs you do, the larger the margin of error.

M. Swindal - So what are some really useful pieces of data for PUMS? For NNIP partners.

V. Mack - We haven't used it. Tulane uses it to do opportunity youth data and that's one example.

L. Pittman - So that's available at the PUMA level?

V. Mack - For a different age group than the typical ACS. The benefit is that you can get all the detail.

J. Newcomer - There was a grant application that we used it for?

R. Pitingolo - A common one is to create your own age cuts.

J. Newcomer - We've also used it is in developing some various methods for calculating housing affordability across areas and doing that by number of persons in the household. And we did that in the counties. But when we start having to split it out among the different types of communities it's tough. We have some resort counties and the housing stock composition is very different in those counties. The allocation method is one of the ways we can mistakenly skew the reality. So we're interesting to write some guidance about data buyer.

R. Pitingolo - So housing affordability is something that would be really cool for NNIP partners. If you're looking at just AMI or percent of housing spent on housing, now we're looking at specific families. When I was doing it, I was looking at roommates or group housing, something really common in DC. That household could have above 100k but clearly this isn’t a typical high-income household. The nuance there is so important. You can see whether people are roommates versus shacking up.

M. Buchanan - You said you use IPUMS, I've gotten it from the Census website.

R. Pitingolo - The stuff directly from the Census is good but it's less detail. It’s available earlier but if you need additional detail, the relationship file is created by the intermediary. And the weights and other stuff. You need that time until they process the data and add it all in. Do you want me to show you what it looks like? <<Demonstrates on computer>>

R. Pitingolo - the IPUMS people have processed the data for SAS and STATA so you can get the data pre-formatted and skip a step. But if you get the raw data file, you have to process it.

M. Swindal - Some of us use R and some use SPSS.

R. Pitingolo - You get a free account and have to agree to use the data for good and not evil. As an example, let's do number of toilets per household, in Iowa. [Selects data on IPUMS website]. [Referring to special settings in the “code” section…] So it's always helpful to use the code. To get a gut check, you can check to see how many raw observations there are using this option. In this example, there are 14,000 people who don't have flushing toilets here. So this is helpful.

So now we need where people live - we want it in household-level, geographic, and by state. So two variables in my cart right now (“toilets” and “state”).

L. Simmons - So by state, is that an approximation of just an aggregation of PUMA?

R. Pitingolo - If you asked for the county-level variable for New York City, it would return a variable and return a county that it needs. If you picked a PUMA in Upstate New York, the value for county would be like “999”. It wouldn't give you a county because the puma is larger than the county.

[Moving to the sample selection]

So here we could get the data for every survey that you could possibly ask of someone. I'll do a one year ACS in 2013. The platform preselects a bunch of data for you that it thinks you will need. I only selected State FIP and “toilet” variables but here I got everything (all necessary weights).

So you create your data extract and you can define it. I'll get a .csv, a rectangular file so that every row is a survey response. You can actually select a subset of the data if the data file is too large by searching for ‘select cases’. In this case, I only want Iowa. But if you have a ton of variables and you're pushing like 10 GBs, you'd limit it your selection. But when I run this, I usually run everything and write a line of code to select specifics.

M. Buchanan - is that also where you would select PUMA as well?

R. Pitingolo – [adds puma to the variable list]. It is good practice to also add a description of the extract. Everything you've ever saved in PUMS will always be here which is really helpful.

Ok so the .csv is ready so I'm gonna extract it and put it on the desktop. [Opens file in Excel]

So first, I'm interested in knowing the cases where there's no toilet in the house. [Filters the data to show only those cases]. If we want to know how many households we have that fit this description, we would need the household weight. Multiply each household by the weight. If we add them up, we have the total number of households without toilets. Now we have the answer about who doesn't have plumbing but we have so many follow up questions. Now what variables do I need to answer these questions? Probably the race, gender, education, run it by PUMA, see if there are children, see if there are rural and urban difference, etc.

M. Buchanan - So within the household data set, we have the number of people in each house?

R. Pitingolo - The serial number represents a household of people, so this household [picking an example household] is a household of five because there are five of the same serial numbers in a row for this variable. “Pernum” represents the order of the people filling out the form. We could figure all that out if we have the variable for age to see what kind of a household we imagine they are. We also know that their PUMA is 1800. If I had a shape file for PUMA, I could figure out whether they're urban or rural. Find a variable that you want, you could add on to here.

So here we have 70 survey responses representing 8,000 households, which is a huge margin of error when you finally run it. You might find based on that to go to the three or five-year sample. Obviously, we want to use the smallest year range so it's recent, but it's useless because 70 results is not covering 8,000 people.

L. Simmons – What’s our population? Household or housing units?

R. Pitingolo - So housing units is occupied housing plus vacant units. What I pulled was just a household file and it's only occupied housing units. If you want the housing units you need to add an extra step on the extract.

M. Buchanan - what you summed was the person weight. If you used the household weight?,

R. Pitingolo - The household weight only applies to the first person in the household. I actually want to take away all of the people in the household that aren't pernum=1. I want to keep only pernum number 1 and now I multiply the household weight times all the ones. And if it were income, I would only do the first person, so in the end it's 3319 households. So 70 responses for 3319 households.

N. Moe - So you could look at the household weights, and only do the householders. If you did the person weight, you're doing other things with the population?

R. Pitingolo - So if I'm looking that household of five, each person has different person weights. If you add up all those person weights, it would be way bigger than the household weights. At some points, they used other indicators to create those weights. I didn't download the race the age the sex, etc. those things go into the weights, but we don't see that. We can download the replicate weights to see where they got this from.

N. Moe - I could see a situation where you're picking the average household size if you were basing it off of the number of columns. Or you could use the weighted data, but the ratio wouldn't make sense. That's why I'm curious.

R. Pitingolo - What I would normally do is that before I start digging, I want to check that what I downloaded is actually what I downloaded. This data represents all the people in Iowa. If I just add up the “perwt” (person weights), it's 3,090,000. If that population is correct, then I downloaded what I think I downloaded. That's the first gut check you should always do. The thing about the weights is that it's really hard to understand but it's necessary to run any analysis no matter how simple. So if I limited this down to first person and added the households, I'd have about 1,000,036 households. It doesn't take a lot of extra time to do that but it does save a ton of time. We might to run a percentage of minorities among these households. If we're running it off of the wrong subset, it might not be the right mean.

M. Buchanan - have you done any trend comparisons? Like 2013 to 2014?

R. Pitingolo - This is one year but you can do one, three and five.

M. Buchanan - Is the five year that much more accurate?

R. Pitingolo - It's the same as the tract level. It has a smaller margin of error but if that margin is fifty percent then it's no good anyway. The margin of error will change based on the population, like doing a particular PUMA but not a particular age range.

M. Buchanan - I'm asking about change over time?

R. Pitingolo - So you can do it but you probably won't find any statistically significant differences. I had one case where the error bars were overlapping in every single year. So we don't know that there's a growth.

M. Buchanan - We used the Survey of Business Owner PUMS, and we were interested in the immigrant business owners. I think that there's been a new one in 2012. So there are a bunch of interesting PUMS resources.

R. Pitingolo - They're segregated. They don't want people mixing and matching.

R. Pitingolo - one thing that I see all the time is that ACS will tell you - where do you live? And the last one is where did you live last year? A different house in the same city, a different state, the same state, etc. And then some shop will say “Let’s see where people in Denver came from!” That's something that a lot of people do and it's really clickbaity but it's a good example. It's not really NNIPish but if you need a blog that you need content for, it's helpful. We were trying to work with a local DC journalist because she had an appetite for neighborhoods. We know that there are young professional group houses. What do the people look like? Are they mostly homogenous race and low income? It didn't really pan out but it's interesting. It's also not at the neighborhood level.