Joint Statistical Meetings - July 2018

Will Administrative Data Save Government Surveys? A View From Alaska

Eddie Hunsinger, State Demographer
Alaska Department of Labor and Workforce Development

ALASKA DEPARTMENT OF LABOR & WORKFORCE DEVELOPMENT

I'm a demographer for the State of Alaska – most US states have some staff who work on population data (projections, and many states make their own estimates as well), and partnerships programs to help ensure accurate and useful data from the US Census Bureau – I lead a three-person team (Eric Sandberg, Liz Brooks, and me – so grateful for their expertise) that does these things in Alaska.

In Alaska we actually have and get to use a terrific administrative data resource: the Alaska Permanent Fund Dividend (PFD) applicant database – we geocode (a huge task that makes all its use possible, and all thanks to Eric Sandberg) it and use it extensively each year to produce population and migration statistics (and colleagues use it for many employment stats too, but I'll focus on population and migration data). But from it we also see a lot of potential pitfalls in use of administrative data for population statistics (including demographic characteristics as well as population totals), and especially why **it's so critical to ensure we have a census/benchmark that is complete/right.**

I'll first talk briefly talk some more about the Alaska PFD data, which I think offers an interesting case study, and how we use it to create population and migration statistics, and issues/shortcomings with it. **Then I'll briefly cover some concerns I have from Alaska, as a data user and partner, over potential for expanded use of admin records as part of the decennial US Census**.

I think that Administrative Data can be useful to improve government information and statistics, but they should be used with a lot of care toward fully and clearly improving quality and systems, and uses and shortcomings should be communicated well and openly.

Screen image from https://pfd.alaska.gov/

Alaska has a statewide Permanent Fund Dividend (PFD) program, to which any resident (meeting certain eligibility requirements, but even including children) may apply each year and receive a dividend payment. The payment can vary quite a bit from year to year, but is often around $1,000 to $2,000 (I think probably closer to 1,000 in the ~near-term outlook). The number of applicants statewide is typically around 90 percent of the estimated usual resident population. Applicants each have Unique IDs in the system and SSNs are provided, which we can use to link them from year to year and to some other data sets. Fields include physical address as well as mailing address, date of birth, place of birth, sex, uniform military... **Sounds so ideal!**

But, some finer print:

-The residency requirement for eligibility is one calendar year (except for newborns) – this means that it takes between one and two years to become eligible, and a large population of recent movers to the state is lagged, and that many (who leave before they become eligible) are never picked up. Further, the residency definition is different than "usual place of residence" (decennial census and population estimates) or "current place of residence" (ACS data collection).

-Many special populations (military, seasonal workers in mining, seafood, tourism…more young working-age) are not well-covered by the PFD data – largely related to eligibility, but also I think knowledge about the program and eligibility.

-Alaska has a lot of "non-city-style" addresses, and/so even though physical addresses are provided, many are relatively unstable and difficult to geocode.

-A person who just didn't apply (or used a different address) in one year will appear inaccurately to

be someone who moved (or died for didn't apply).
-The share of the state's population (based on independent population estimates from the Census Bureau) that applies for a PFD seems to be declining some in recent years. I'm not sure why really – more short-term residents? Military outreach/education about the PFD? Residents less fixed in Alaska (maybe with recent retirements)? General program participation trends? I don't know at this point.

Screen images from http://laborstats.alaska.gov/

Really, the PFD data has a very different objective than population or migration estimation, and misses a lot of people, and some groups missed more than others – a big problem.

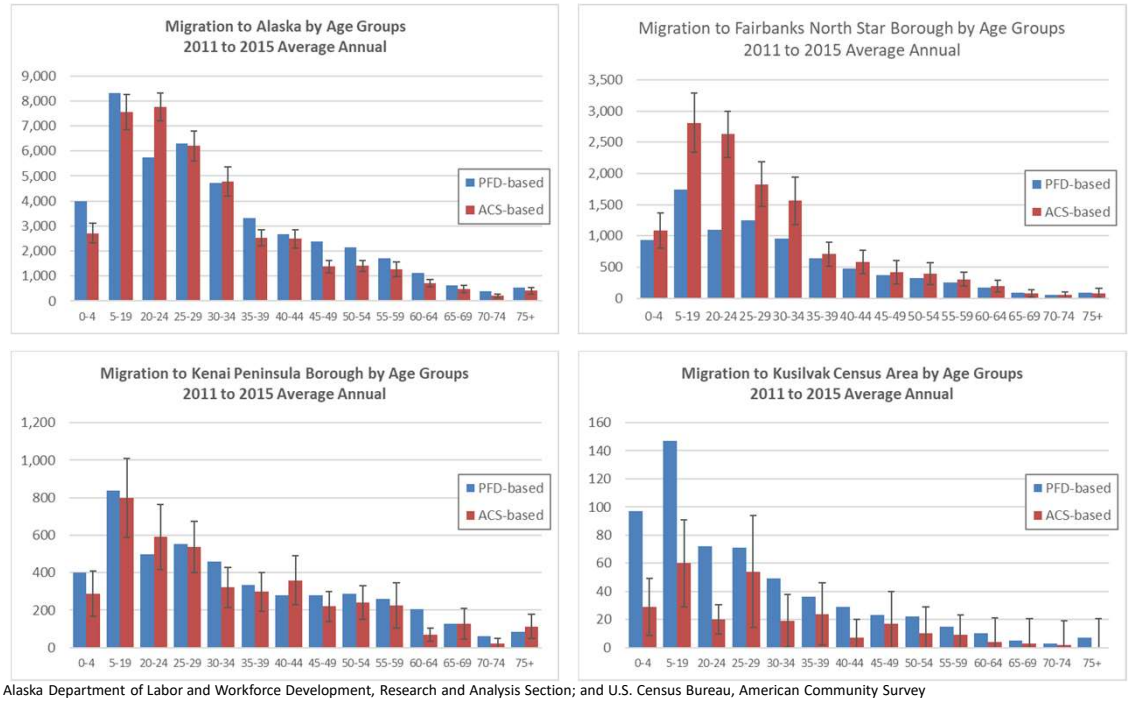
But we still use it in three different ways that we feel comfortable with for population and migration data production – and these will likely be obvious but as someone from a state, who works with a lot of end-users, I think so important to keep up front in consideration of Administrative Data for Government Surveys:

(1) Where we can benchmark it to a reliable "complete" count – underscoring the importance of a complete census for the most used or valuable variables. **This is what we do for our population estimates**: "census ratio" method with control/adjustment to an independent total, and we make annual estimates for more than 1,000 geographic entities in the state.

(2) Where we can effectively supplement and improve (not just substitute) information and not harm information. This has been harder to apply than one might think though, because a lot of the PFD information that isn't found elsewhere may be poor or incompatible (physical address names, for instance – even SSNs, for instance). But **we have found some use and value for vetting of some ACS data**, on place of birth for instance. And we think the Census Bureau may be able to use it to supplement information, but as a supplement to existing resources it would (hopefully) be very marginal – just records/information that aren't found by other resources (eg survey response, IRS) *and* are reliable.

(3) Where a more raw form of it is useful to some end-users (no doubt fewer, but some), and we can describe for users readily, plainly, and specifically what the data is. **This is what we do for**

migration – “PFD-based migration.” I think another neat/useful and similar example is US IRS Tax Statistics migration data, which are published online.



Sources: Alaska Department of Labor and Workforce Development, Research and Analysis Section; and U.S. Census Bureau, American Community Survey

So, we thought it would be interesting/neat to look at **some PFD-based migration data, and especially how it compares to American Community Survey** (ACS) ("residence one year ago" question) based migration data.

***Generally* (for totals we tabulate each year), we tabulate the migration data that we publish in a very straight-forward way and don't adjust it** (so again I think it has some similarities to US IRS Tax Statistics – including that it's incomplete but there's some clarity on what it represents): **We compare the geocoded physical address of each applicant in subsequent years**. By this we can see how many people changed their physical address and how with some detail (between each borough and census area each year, for instance), and how many people are new applicants (by migration or birth) or are no longer applicants (by migration or death). (Many thanks to colleagues Eric Sandberg and David Howell for all of their work and expertise to develop and produce of this data for Alaska!)

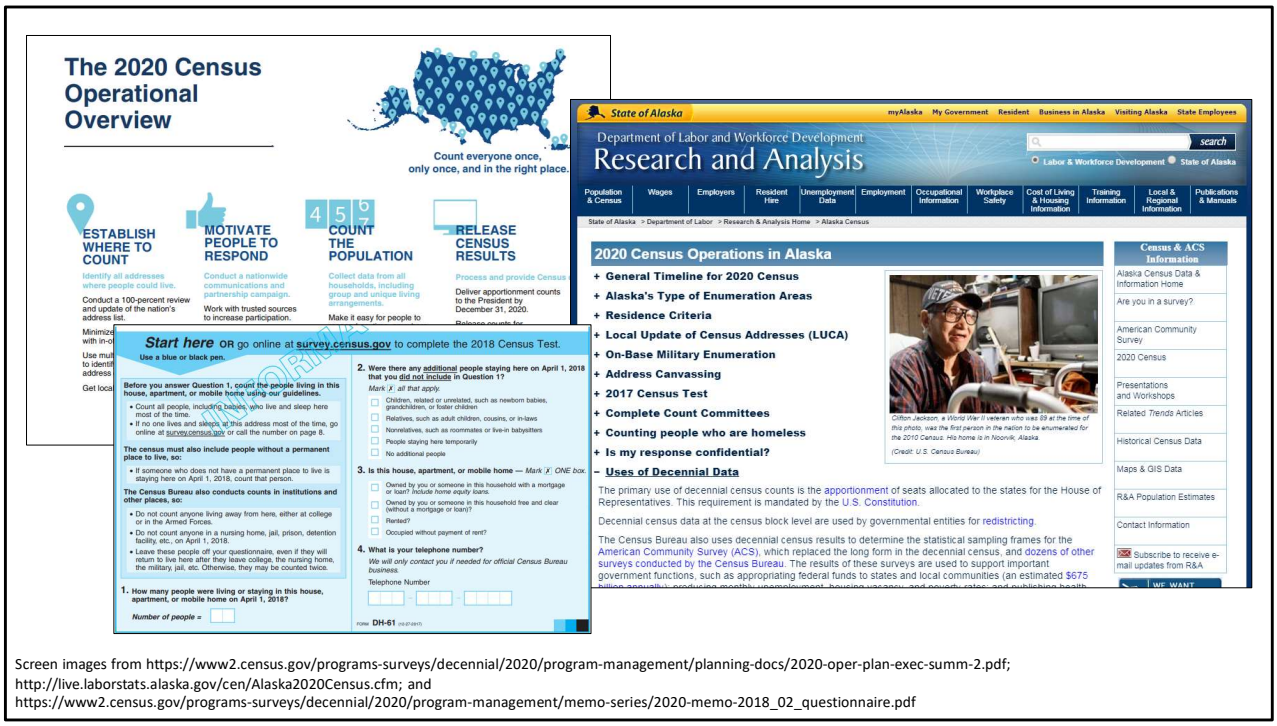
These graphs are in-migration by age groups (note 5 to 19 is combined) to selected areas, from a special migration tabulation that we produce every five years (blue/left is PFD-based, red is ACS-based), and for which we do (for the out-mover side) make a very simple (and I especially hope clear-for-users) adjustment for deaths by age.
The graph to the top-left covers Alaska/statewide (The other areas are "boroughs and census areas" – these are the same geography as counties in other US states- and I selected them just to cover areas with very different characteristics:

-Fairbanks North Star Borough (top-right) (total population ~100,000) is a major population center in Interior Alaska – it's also home to significant military and student populations
-Kenai Peninsula (bottom-left) (population ~65,000) is south of Anchorage and has larger seasonal populations and a few population centers – snowbirds and generally a higher median age (~42 for usual residents)
-Kusilvak Census Area (bottom-right) (population ~8,500) covers 13 remote communities in southwest Alaska, with no large population center – population there is the youngest in the state (median age ~24), over 90 percent Alaska Native, and relies significantly on subsistence

Some important issues with the PFD-based data:
-So, from earlier comments, there are features of the PFD data that can push/bias them higher than actual migration, such as people not applying in one year (or applying in a different place) even though they didn't move.
-And there are features that will bias them to be lower than actual migration, such as people don't reside in the state long enough (even up to two years) to be eligible - I imagine it for the statewide age 20 to 24 year old populations, and I believe can see here it for the Fairbanks populations (military, students and young workers).
-These biases must cancel each other out some, but it's not readily known where or how much.
-The different factors causing errors affect some areas more than others, and maybe some of these could be guessed at or estimated by an expert, but not readily with precision.
-And so important, I think: they don't in their basis *aim* to represent fair/equitable coverage for different areas.
-The goal of the PFD applicant database isn't to track migration, so some valuable options to improve the data are missing.
-And the data aren't useful for everyone – some users may need a more conclusive (less-biased) estimate, or an estimate with adjustments for more of the biases.

The ACS data must have their own non-sampling error too (in addition to the sampling error):
-Here (as big as it is for Kenai Peninsula and Kusilvak Census Area) we see just the sampling error (which would be inconsequential for the PFD-based sample size).
-ACS data does have its issues with residency determination I believe – where the questionnaire uses a "current residence" concept, while the estimates they're weighted to use a "usual residence" concept, and this could have some effect.
-And the challenges of low response or poor response – but it's not clear to me that follow-up efforts and imputations based on the sample are best *replaced* by administrative data, which have their own biases. I think testing should demonstrate improvements from expanded use of administrative data by type of areas *before* a major shift is committed to.

Like the graphs by age here, comparing PFD-based and ACS-based total migration between boroughs and census areas ("county-to-county" migration) also showed many differences that were bigger than the ACS sampling error.



Screen images from https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan-exec-summ-2.pdf; http://live.laborstats.alaska.gov/cen/Alaska2020Census.cfm; and https://www2.census.gov/programs-surveys/decennial/2020/program-management/memo-series/2020-memo-2018_02_questionnaire.pdf

So again, I think administrative data is often valuable and helpful in its raw form or for a supplement. But for many uses, a very reliable and fair benchmark – the decennial Census in the US – is essential: It is the starting point for countless data resources in the United States.

Of course the Census Bureau has much better experts than me who can describe the different parts of the decennial Census, but I want to note a few general aspects of it and admin-data-related issues that I think are important to consider, and related issues that we are especially concerned about as data users and partners in Alaska:

I think of the decennial US Census (as it exists currently) as really two censuses – first, **canvassing**, in which the Census Bureau works to complete (really update/improve) its database (Master Address File) of every housing unit and living facility in the country, and second, **enumeration**, in which the Census Bureau works to count the people at every housing unit in the country. When they don't get a response, they have to find out why, and ultimately determine whether the unit is vacant or not (**Non-Response Follow Up**), which is very costly. **There are at least a few places in which the Census Bureau may use administrative data in this process, and that I have some concerns about as a data user:**

(1) To improve the Master Address File (this is done extensively, for instance, with "Delivery Sequence File" data from the US Postal Service).

Admin records like the Delivery Sequence File data are incomplete and don't cover all housing units

at all or evenly, but in *past* censuses it's been used effectively as part of a combination of methods and resources for Master Address File improvement, which included nationwide in-field canvassing too. **My concern here is about the prospect of different methods for canvassing being *taken away* from areas where they're needed** – replaced by admin records – even if reduced to a small share of the national population, it can be a very big share of a given state or county.
Related, and more info: https://www.forakergroup.org/wp-content/uploads/filebase/census_documents/Alaska-White-Paper-Types-of-Enumeration-Areas.pdf

(2) To count people, particularly in this Census for places like "group quarters" living facilities (this is done extensively in recent Censuses with data (rosters, etc) from facility contacts, and the Census Bureau plans to expand it (part of Operational Plan)).
This puts a lot of weight on individual respondents (interpreting and responding for large facilities), and, additionally, it seems that quality suffers - review of the data from 2000 and 2010 showed a great deal of missing information.
Related, and more info: https://www.forakergroup.org/wp-content/uploads/filebase/census_documents/Alaska-White-Paper-Group-Quarters.pdf

(3) To help determine whether a unit is vacant or not (there has been research on this for use in the 2020 Census).
Some Census Bureau testing on this (2016 Census Test) showed that large shares of administrative record vacant (21 percent) or delete (29 percent) units were in fact occupied, and I'm not sure how use of it would help toward improved counts.
Related: https://www2.census.gov/cac/nac/meetings/2017-11/mule-admin-records.pdf

Basically I'm especially concerned about transitions to use of admin data that are at the cost of data quality and coverage for different areas.

So again, I think that Administrative Data can be useful to improve government information and statistics, but they should be used with a lot of care toward truly improving quality and systems, and uses and shortcomings should be communicated well and openly.

Thank you!

Contact: eddie.hunsinger@alaska.gov; 907-269-4960
Website: laborstats.alaska.gov

Information about the Alaska Permanent Fund Dividend
pfd.alaska.gov/
pfd.alaska.gov/Division-Info/Annual-Reports

Alaska population and migration data
laborstats.alaska.gov/pop/
laborstats.alaska.gov/pop/migration.cfm

Information about the Census in Alaska
forakergroup.org/index.php/sector-voice/ensuring-good-census-data/
laborstats.alaska.gov/cen/Alaska2020Census.cfm

ALASKA DEPARTMENT OF LABOR & WORKFORCE DEVELOPMENT

Info about the PFD data
https://pfd.alaska.gov/
https://pfd.alaska.gov/Division-Info/Annual-Reports

Alaska population and migration data (many thanks Eric Sandberg, Liz Brooks, and David Howell)
http://live.laborstats.alaska.gov/pop/
http://live.laborstats.alaska.gov/pop/migration.cfm

Info about the Census in Alaska (many thanks Liz Brooks and Alaska Census Working Group)
https://www.forakergroup.org/index.php/sector-voice/ensuring-good-census-data/
http://live.laborstats.alaska.gov/cen/Alaska2020Census.cfm