Using 2010 Census Coverage Measurement Results to Better Understand Possible Administrative Records Incorporation in the Decennial Census

Andrew Keller and Scott Konicki[1]
U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233

Abstract

The 2010 Census Coverage Measurement (CCM) program evaluated coverage of the 2010 Census and produced components of census coverage results that included estimates of correct enumerations, erroneous enumerations, imputations, and omissions of the national household population. A goal of the 2010 CCM program was to inform decisions for the 2020 Census. As part of the 2020 Census, the Census Bureau is researching the possible use of administrative records (AR) to provide a status and count for some nonresponding addresses. The goal is to understand the ramifications of using AR on coverage and quality in the decennial census. In general, this research demonstrates how the 2010 CCM can be another tool by which AR usage can be evaluated.

Key Words: Administrative Records, Census Coverage Measurement, Components of Census Coverage

1. Introduction

To meet the strategic goals and objectives for the 2020 Census, the Census Bureau must make fundamental changes to the design, implementation, and management of the decennial census. These changes must build upon the successes of previous censuses while also balancing cost containment, quality, flexibility, innovation, and disciplined and transparent acquisition decisions and processes. In the 2010 Census, the Nonresponse Followup (NRFU) operation included about fifty million addresses requiring up to six contacts each, totaling about $1.6 billion (Walker et al. 2012).

For the 2020 Census planning, Mule and Keller (2014) laid out the many issues and different potential ways that administrative records (AR) could be used in an adaptive way in the NRFU operation. The Census Bureau implemented tests in 2013, 2014 and 2015 that used AR to reduce the number of contacts during the NRFU operation.

- Walejko et al. (2014) document an adaptive design pilot test in October 2013 conducted in Philadelphia, Pennsylvania. The pilot test was of a small sample of addresses that were in the NRFU universe in the 2010 Census. This was the first step to test the feasibility of using AR to reduce the number of contact attempts during NRFU.

[1] The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

- The 2014 Census Test was conducted with a Census Day of July 1, 2014 in parts of Montgomery County, Maryland and the District of Columbia. Keller et al. (2016) documented how basic rules were developed to identify occupied and vacant addresses through the use of AR. One of their findings was that improvements could be made by using predictive modeling approaches as compared to rules.
- The Census Bureau conducted research and developed predictive modeling approaches that used logistic and multinomial regression predictions. Linear optimization approaches were then applied to maximize the AR determination given constraints. This new approach was implemented in the 2015 Census Test in Maricopa County, Arizona (Morris et al. 2016).

In addition to the mid-decade census tests discussed above, the development of possible AR models has been guided by comparing models retrospectively against 2010 Census results. For example, running a simulation on 2010 Census data, we counted how many addresses identified as vacant by the AR model were actually vacant during the 2010 census. Essentially, this type of analysis treats 2010 Census results as 'truth'. However, a difficulty underlying the evaluation of AR modeling is the inherent error in census results. Although the analysis using the 2010 Census results as 'truth' provides a solid basis for assessing model performance, it is not the only way model performance can be measured. It is possible that census quality could be improved using AR data that is not reflected by solely comparing the modeling results against 2010 Census 'truth'.

Estimating census coverage error has traditionally been the focus of the coverage measurement program. Specifically, the 2010 Census Coverage Measurement (CCM) program evaluated coverage of the 2010 Census. This research folds in 2010 CCM results to provide an additional understanding of the ramifications of using AR modeling in lieu of NRFU contacts.

Keller and Fox (2012) provide the 2010 components of census coverage, including estimates of correct enumerations, erroneous enumerations, and omissions for the national household population. Within that document, they provide coverage component estimates for persons by major demographic groups, census operational areas, states, large counties, and large places. Section 2 discusses how AR models have been developed and how AR data would be incorporated into the NRFU operation. Section 3 provides an example simulation with AR data and shows how CCM information can be used to glean information about the quality of the AR models.

2. Administrative Records Modeling for NRFU

For the 2015 Census Test conducted in Maricopa County, Arizona and the 2016 Census Test conducted in Harris County, Texas and Los Angeles County, California, the Census Bureau identified occupied and vacant units using AR data and models. In this paper, we describe a national-level application of the same models that we applied during the 2016 Census Test. For the simulation in Section 3, we used the 2016 methodology to fit our AR models on a sample of the 2010 Census NRFU universe. We then applied the fit to the entire 2010 Census NRFU universe. See Morris et al. (2016) for specific details about the modeling approach and dependent and independent variables. Following the modeling, the NRFU address universe was split into three categories:
(1) units identified as occupied using AR (AR Occupied)
(2) units identified as vacant using AR (AR Vacant)

(3) addresses identified as no determination (No Determination).

2.1 Nonresponse Followup Contacts

This section gives an overview of the NRFU contact strategy related to enumerating some addresses with AR. This strategy was laid out in the release of the 2020 Operational Plan (U.S. Census Bureau 2015) and implemented in the 2016 Census Test. Note that the 2016 Census Test also included an AR Delete (Not a Housing Unit) category. For the purposes of this paper, AR Delete cases are grouped with AR Vacant cases. The rationale is that operationally they are treated the same as AR Vacant cases even though they will not be part of the final housing unit count. For the 2016 Census Test, before the NRFU operation began, a NRFU address may have received up to four mailings before and after Census Day. These mailings included letters encouraging the household to respond on the internet, two postcard reminders, and a paper questionnaire. If the address did not respond to these, a decision was then made about how many times to contact the address during the NRFU operation.

Figure 1 shows the flowchart of the contact strategy related to AR cases for the NRFU operation. Addresses determined to be AR Vacant received no contacts during the NRFU operation. While these units did not receive any NRFU visits, a postcard was mailed to them during the 2016 Census Test at the beginning of the NRFU operation. This allowed people at occupied addresses to self-respond by going online and filling out the internet questionnaire or dialing the questionnaire assistance phone number.

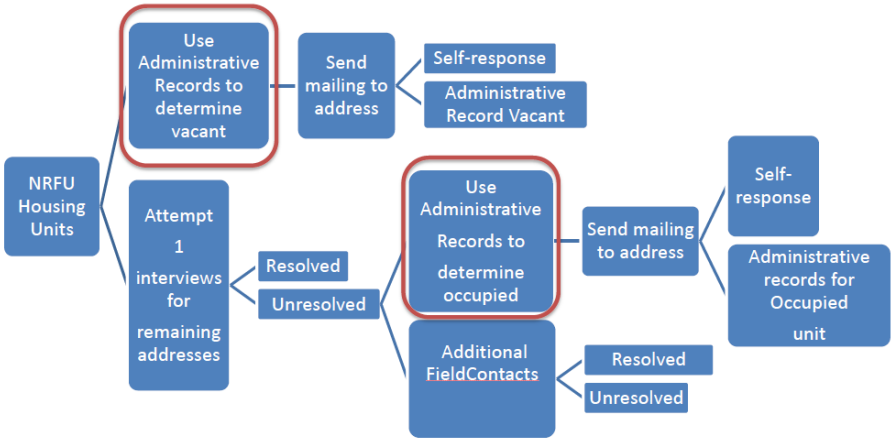

Fig. 1: Nonresponse Followup Contact Strategy for Administrative Record Cases

The remaining cases received an initial field visit. This visit allowed each case to be resolved in several ways. It was resolved by

- completing the interview with the household member,
- determining the address to be vacant, or
- determining the address was not a housing unit.

If nobody in the household was home, the enumerator left a notice of visit. This notice of visit included information that instructed persons in the household to respond

- by going online,
- dialing the questionnaire assistance number, or
- sending the paper questionnaire that they received earlier.

Cases determined to be occupied by AR only received the initial visit in the 2016 Census Test. While they received only the initial visit, an additional postcard mailing was sent to

the address. This postcard had information detailing how the household could still go online or dial the questionnaire assistance number to self-respond. As shown, there are several ways before and during NRFU that the Census Bureau is attempting to obtain and use self-responses before having to use AR determinations. The remainder of this paper focuses on coverage ramifications when applying the models to determine the AR Vacant and AR Occupied cases that are circled in Figure 1.

3. Administrative Records Simulation

To identify vacant units with AR, we developed a multinomial logit model which predicted the probability that an AR would have been enumerated as vacant during the 2010 Census. The dependent variable had three possible values for each AR address in the NRFU universe:

- occupied
- vacant, or
- delete (i.e., not a housing unit).

We defined a Euclidian vacant distance function for AR Vacant identification as:

$$d_{AR_{Vac}} = \sqrt{(1 - p_{vacant})^2 + \left(0 - p_{occupied}\right)^2}$$

The formula shows that cases with the smallest distance were those with the highest vacant probability and lowest occupied probability. Starting with the smallest vacant distance, AR Vacant cases were identified by allowing for increased vacant distance values up to a vacant threshold. This threshold was based on analysis of 2010 Census NRFU data. This modeling approach identified 5.108 million AR Vacant units nationally.

Two models were developed to identify AR Occupied units: a person-place model and a household (HH) composition model. The person-place model predicted the probability that an AR person would be enumerated at the sample address if fieldwork was conducted. The HH composition model predicted the probability that the sample address would have the same HH composition determined by NRFU fieldwork as its pre-identified AR HH composition. HH composition is defined by the number of adults in the unit and the absence or presence of children.

Similar to AR Vacant, we defined a Euclidian occupied distance function for AR Occupied identification as:

$$d_{AR\_Occ} = \sqrt{\left(1 - p_{person-place}\right)^2 + \left(1 - p_{HH\ composition}\right)^2}$$

The formula shows that cases with the smallest occupied distance were those where the person-place probability was closest one and the household composition probability was closest to one (i.e. the (1,1) point). Starting with the smallest occupied distance, AR Occupied cases were identified by allowing for increased occupied distance values up to an occupied threshold. This threshold was based on analysis of the 2010 Census NRFU data. This modeling approach identified 7.077 million AR Occupied units nationally.

Table 1 shows the distribution of cases identified as vacant and occupied by AR models and those for which no AR determination was made.

Table 1: NRFU Universe by AR Model Category

AR Model Category	Total	No Determination	AR Occupied	AR Vacant
N	49,817,252	37,632,033	7,077,460	5,107,759
Percent	100.0%	75.5%	14.2%	10.3%

3.1 Comparing AR Modeling Simulation to 2010 NRFU

To understand the possible error in the model, we compared AR enumerations to the 2010 Census enumerations. That is, how many AR Occupied cases were occupied during 2010 NRFU? Similarly, how many AR Vacant cases were vacant during 2010 NRFU? Table 2 shows four 2010 NRFU status outcomes:

- occupied (Occ),
- vacant (Vac),
- delete (Dele), or
- unresolved (Unres; i.e., status not resolved during NRFU operation).

Of the 5,107,759 cases identified as vacant by AR, about 10.3% were classified as occupied in the 2010 NRFU operation. Similarly, of the 7,077,460 AR Occupied cases, 7.9% were classified as vacant and 1.7% were deleted.

Table 2: NRFU Status Assigned Via Simulation versus 2010 NRFU Status

AR Model Category	Total	2010 NRFU Status				%			
		Occ	Vac	Dele	Unres	Occ	Vac	Dele	Unres
AR Vacant	5,107,759	525,644	4,012,480	528,462	41,173	10.3%	78.6%	10.3%	0.8%
AR Occupied	7,077,460	6,377,432	561,396	119,453	19,179	90.1%	7.9%	1.7%	0.3%

At the core of this paper is the idea that solely comparing possible AR modeling methods to previous 2010 Census results is insufficient because census results have errors. One might be tempted to conclude that the 525,644 units identified as vacant by AR but enumerated as occupied by NRFU in Table 2 are all misclassification errors attributed to the AR models. However, it is possible that all or some persons in these units may be erroneous enumerations or whole-person census imputations. Hence, the AR simulation may be more accurately viewed through the prism of the 2010 Census Coverage Measurement (CCM) program. To understand this, Section 3.2 integrates potential AR modeling methods with the results from the 2010 CCM. Note that this paper focuses on one specific simulation as a qualitative demonstration. However, the use of CCM to evaluate AR models has been extended to many simulations.

3.2 Comparing AR Modeling Simulation to Census Coverage Measurement

The 2010 CCM program evaluated coverage of the 2010 Census to aid in improving future censuses. The CCM measured the net coverage and components of census coverage of housing units and persons, excluding group quarters and persons residing in group quarters. The CCM sample design was a probability sample of 170,000 housing units. Remote areas of Alaska were out of scope for the CCM.

The general estimation approach for components of census coverage for persons fell into four categories:

- estimates of correct enumerations
- estimates of erroneous enumerations
- tabulations of whole-person census imputations
- estimates of omissions

Keller and Fox (2012) provided the 2010 components of census coverage for the national household population. Since a goal of the CCM process was to aid in improving future censuses, we show coverage properties of the AR simulation above to provide additional insight into the quality of the simulation.

Table 3 has separate estimation domains for AR No Determination, AR Occupied, and AR Vacant cases. It shows the components of census person coverage for the 7.077 million AR Occupied units, 5.108 million AR Vacant units, and the remaining No Determination units that the AR models indicated as insufficient to enumerate with AR. The first column shows the census count. The census count is then broken into rates of correct enumeration, erroneous enumeration by duplication, erroneous enumeration for other reasons, and whole-person imputation.

Table 3 shows that we enumerated 16.243 million persons in the 7.077 million housing units the simulation identified as occupied by AR. Among these enumerations, 91.6% were estimated to be correct enumerations, 2.2% of these enumerations were erroneous due to duplication, 0.6% of these enumerations were erroneous due to some other reason, and 5.7% of these enumerations were whole-person census imputations. It is clear that not every census enumeration in these units was correct. Central to the point of this paper, a lower AR simulation total in comparison to the Census 2010 total may result in greater census quality given that 0.450 million persons were enumerated in error in the 2010 Census. In addition, for 0.918 million persons, we had to impute each characteristic. In practice, if we were to call these units occupied and enumerate them from AR data, there would be no whole person imputations.

In the 2010 Census, we enumerated 0.987 million persons in 5.108 million housing units that were classified as vacant by AR models (see Table 3). However, not all these persons were correct enumerations. Among these enumerations, 70.7% were estimated to be correct enumerations, 8.5% of these enumerations were erroneous due to duplication, 1.5% of these enumerations were erroneous due to some other reason, and 19.3% of these enumerations were whole-person census imputations. In practice, if we were to call these units vacant from AR data, AR methods would omit 0.698 million correctly enumerated persons.

Table 3: Components of Census Coverage by AR Simulation Category

AR Status	Census Count (Thousands)	Correct (%)	Erroneous (%)		Whole-Person Imputations (%)
			Duplication	Other	
U.S. Total	300,703	94.7	2.8	0.5	2.0
	(0)	(<0.1)	(<0.1)	(<0.1)	(0)
AR No Determination	283,473	94.9	2.9	0.5	1.7
	(0)	(0.1)	(0.1)	(<0.1)	(0)
AR Occupied	16,243	91.6	2.2	0.6	5.7
	(0)	(0.2)	(0.2)	(0.1)	(0)
AR Vacant	987	70.7	8.5	1.5	19.3
	(0)	(1.1)	(1.1)	(0.4)	(0)

Standard errors are shown in parentheses below the estimate. See Imel et al. (2013) on how CCM standard errors were derived.
The 2010 Census count excludes persons in group quarters and persons in Remote Alaska.

Table 4 displays the national implications of using AR enumeration by integrating AR Occupied and AR Vacant enumeration. Note that this is not a perfect representation of census operations in the sense that we assume that all AR Occupied units get enumerated via AR because none of the NRFU units were resolved on the first contact. As seen in Section 2.1, the current plan is to conduct one in-person visit to these units and send an additional mailing if that visit is unsuccessful at resolving the case. Thus, there are multiple opportunities to obtain a census response rather than using the AR result. Second, these simulations do not account for the fact that had we used AR enumerations, subsequent count imputation results would have been altered due to the changes in the donor universe. That is, we replaced the Census 2010 enumerations with the AR enumerations, leaving the 2010 count imputation results as fixed.

Column (2) of Table 4 shows that, had we used AR to enumerate the AR Occupied units, we would not have enumerated the 16.243 million persons. Column (3) shows that, had we used AR to enumerate the AR Occupied units instead, we would have enumerated 16.757 million persons in these same units. Since no interviews were completed, characteristics would have to be taken from AR or imputed for sex, age, race, Hispanic origin, and relationship to householder.

The Census Bureau matches AR persons to the Social Security Numident file to obtain age and sex data. To identify a NRFU unit as occupied via AR models, it must have all ages filled for all persons in AR. In addition, sex is usually a non-missing characteristic because of its presence on the Numident. To identify race and Hispanic origin for persons enumerated in AR Occupied units, we used AR data from multiple sources. Ennis et al. (2015) explain how race and Hispanic origin are assigned to persons in AR data and previous census responses. Obtaining relationship to householder from administrative records is a subject of ongoing research. Hence, a possible advantage of AR Occupied enumeration is that it could potentially require less characteristic imputation. Czajka (2009) discusses directly substituting AR for survey data.

Column (4) shows the 0.987 million persons that we would not have enumerated in the AR Vacant units. Column (5) shows the simulation population when subtracting columns (2) and (4) and adding column (3) to the 2010 Census population (1). This total

represents aggregate effect on the population when using AR models to identify occupied and vacant units. Overall, the simulation results 300.230 million persons. In comparison, the 2010 Census had 300.703 million persons and the CCM had a population estimate of 300.667 million persons. Column (6) shows the undercount observed by the 2010 CCM (2010 CCM Undercount). Treating the CCM estimate as truth, this resulted in a 0.01% overcount as seen in column (6). That is,

$$CCM\ Prod\ UC = \frac{CCM\ Estimate - Census}{CCM\ Estimate} \times 100$$
$$= \frac{300{,}667{,}287 - 300{,}703{,}438}{300{,}667{,}287} \times 100\% = -0.01\%$$

Column (7) shows the undercount observed replacing by the 2010 Census count with the simulation population using AR (AR Simulation Undercount). Again, the CCM estimate is seen as truth, and a 0.15% undercount is seen in column (7). That is,

$$CCM\ Prod\ UC = \frac{CCM\ Estimate - Simulation}{CCM\ Estimate} \times 100\%$$
$$= \frac{300{,}667{,}287 - 300{,}230{,}304}{300{,}667{,}287} \times 100\% = 0.15\%$$

Hence, the net effect of using AR is a change from a point estimate of a 0.01% overcount by the 2010 census to a 0.15% undercount when applying the AR simulation. Both undercounts are within the 95% confidence interval on the 2010 CCM undercount standard error seen in the appendix.

Table 4: AR Simulation Results - National

Category	**(1)** 2010 Census Population	**(2)** 2010 Census People in AR Occupied Units **(Remove)**	**(3)** AR People in AR Occupied Units **(Add)**	**(4)** 2010 Census People in AR Vacant Units **(Remove)**	**(5)** AR Simulation Population	**(6)** 2010 CCM Undercount	**(7)** AR Simulation Undercount
National	300,703,438	16,242,893	16,757,022	987,263	300,230,304	-0.01%	0.15%

Standard errors for 2010 CCM undercount in column (6) shown in appendix.

Using the CCM estimates to understand the ramifications of AR extends to domains as well. Table 5 shows results similar to Table 4 broken out by age and sex groupings. For example using the point estimates, had we used AR enumeration instead of the Census, for 0-4 aged children CCM would have reported a 0.59% undercount as opposed to a 0.72% undercount reported in production. In other words, this simulation shows that using AR in this manner decreases the 0 to 4 undercount. On another note, the simulation shows that this AR enumeration scheme decreases the magnitude of the overcount for 50+ people. For example, the 0.32% overcount of 50+ males seen in the 2010 Census is reduced to a 0.14% undercount in the AR simulation. The 2.35% overcount of 50+ females seen in the 2010 Census is reduced to a 1.96% undercount in the AR simulation. This same idea has been applied over other estimation domains to check for possible ramifications of using AR. The point of this analysis is to get a macro-level understanding of the ramifications of AR usage on census coverage errors.

Table 5: AR Simulation Results – Age/Sex Groupings

Age and Sex Groupings	**(1)** 2010 Census Population	**(2)** 2010 Census People in AR Occupied Units **(Remove)**	**(3)** AR People in AR Occupied Units **(Add)**	**(4)** 2010 Census People in AR Vacant Units **(Remove)**	**(5)** AR Simulation Population	**(6)** 2010 CCM Undercount	**(7)** AR Simulation Undercount
0 to 4	20,157,618	1,437,036	1,510,588	47,885	20,183,285	0.72%	0.59%
5 to 9	20,314,652	1,568,249	1,691,485	38,379	20,399,509	-0.33%	-0.75%
10 to 17	33,429,889	2,126,752	2,258,973	52,605	33,509,505	-0.97%	-1.21%
18 to 29 Male	23,981,678	1,123,864	1,042,832	125,139	23,775,507	1.21%	2.06%
18 to 29 Female	23,912,124	1,179,771	1,186,339	109,435	23,809,257	-0.28%	0.15%
30 to 49 Male	40,256,193	2,738,787	2,894,928	150,524	40,261,810	3.57%	3.56%
30 to 49 Female	41,814,983	2,612,327	2,773,339	115,966	41,860,029	-0.42%	-0.53%
50+ Male	44,886,182	1,671,109	1,639,081	174,839	44,679,315	-0.32%	0.14%
50+ Female	51,950,119	1,784,998	1,759,457	172,491	51,752,087	-2.35%	-1.96%
Total	300,703,438	16,242,893	16,757,022	987,263	300,230,304	-0.01%	0.15%

Standard errors for 2010 CCM undercount in column (6) shown in appendix.

4. Conclusions

The potential use of AR represents a substantial change to census procedures. Simulations with AR data have been helpful for understanding positive and negative aspects of potential AR usage. To evaluate the use of AR, we compared simulation results back to the 2010 Census. However, only comparing various AR modeling methods to previous 2010 Census results is insufficient because census results have errors.

In this paper, we showed how a single simulation using AR to identify occupied and vacant units resulted in an increased national undercount while the undercount decreased for other sub-national domains. Other simulations using AR data have shown different results. In general, this research demonstrates how that the 2010 CCM can be another tool by which AR usage can be evaluated to see the ramifications for national and subnational domains.

5. References

Czajka, J. (2009). "Can Administrative Records Be Used to Reduce Nonresponse Bias?" The ANNALS of the American Academy of Political and Social Science January 2013 645: 171-184.

Ennis, S.R., Porter, S.R., Noon, J.M., and Zapata, E. (2015). "When Race and Hispanic Origin Reporting are Discrepant Across Administrative Records and Third Party Sources: Exploring Methods to Assign Responses." Center for Administrative Records Research and Applications Working Paper #2015-08. Washington, DC: U.S. Census Bureau.

Imel, L., Mule, V.T., Seiss, M., and Mulligan, J. (2013), "2010 Census Coverage Measurement Estimation Methods: Measures of Variation," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-06.

Keller, A., Fox, T., and Mule, V.T. (2016). "Analysis of Administrative Record Usage for Nonresponse Followup in the 2014 Census Test." U.S. Census Bureau.

Keller, A. and Fox, T. (2012), "2010 Census Coverage Measurement Estimation Report: Components of Census Coverage Results for the Household Population in the United States," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-04.

Morris, D.S., Keller, A., and Clark B. (2016). "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census," Statistical Journal of the International Association of Official Statistics, 32 (2016): 177-188.

Mule, V.T. and Keller, A.. (2014), "Using Administrative Records to Reduce Nonresponse Followup Operations," in JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 3601-3608.

U.S. Census Bureau 2015. *2020 Census Operational Plan*. Washington DC: Census Bureau. Available at: http://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/2020-oper-plan.pdf. (accessed March 2016).

Walejko, G., Keller, A., Dusch, G., and Miller, P.V. (2014). "2020 Research and Testing: 2013 Census Test Assessment." U.S. Census Bureau.

Walker, S., Winder, S., Jackson, G., and Heimel, S. (2012). "2010 Census Nonresponse Followup Operations Assessment," 2010 Census Planning Memoranda Series, No. 190, April 30, 2012.

6. Appendix

2010 CCM Undercount Standard Errors for Tables 4 and 5

Table	Category	2010 CCM Undercount Standard Error
4	National	0.14%
5	0 to 4	0.40%
	5 to 9	0.31%
	10 to 17	0.29%
	18 to 29 Male	0.45%
	18 to 29 Female	0.36%
	30 to 49 Male	0.20%
	30 to 49 Female	0.21%
	50+ Male	0.14%
	50+ Female	0.14%