



Differential Privacy and the Accuracy of County-Level Net Migration Estimates

Richelle L. Winkler¹ · Jaclyn L. Butler² · Katherine J. Curtis³ · David Egan-Robertson³

Received: 5 March 2021 / Accepted: 18 June 2021 / Published online: 8 July 2021
© The Author(s) 2021

Abstract

Each decade since the 1950s, demographers have generated high-quality net migration estimates by age, sex, and race for US counties using decennial census data as starting and ending populations. The estimates have been downloaded tens of thousands of times and widely used for planning, diverse applications, and research. Census 2020 should allow the series to extend through the 2010–2020 decade. The accuracy of new estimates, however, could be challenged by differentially private (DP) disclosure avoidance techniques in Census 2020 data products. This research brief estimates the impact of DP implementation on the accuracy of county-level net migration estimates. Using differentially private Census 2010 demonstration data, we construct a hypothetical set of DP migration estimates for 2000–2010 and compare them to published estimates, using common accuracy metrics and spatial analysis. Findings show that based on demonstration data released in 2020, net migration estimates by five-year age groups would only be accurate enough for use in about half of counties. Inaccuracies are larger in counties with populations less than 50,000, among age groups 65 and over, and among Hispanics. These problems are not fully resolved by grouping into broader age groups. Moreover, errors tend to cluster spatially in some regions of the country. Ultimately, the ability to generate accurate net migration estimates at the same level of detail as in the past will depend on the Census Bureau's allocation of the privacy loss budget.

Keywords Net migration · Differential privacy · Age · Error metrics · Spatial analysis · Planning

✉ Richelle L. Winkler
rwinkler@mtu.edu

¹ Michigan Technological University, 217 Academic Office Building, 1400 Townsend Dr., Houghton, MI 49931, USA

² Pennsylvania State University, State College, PA, USA

³ Applied Population Laboratory, University of Wisconsin, Madison, WI, USA

Introduction

The United States Census Bureau will implement a formal differential privacy disclosure avoidance system to protect respondent privacy for the first time in publishing Census 2020 data. Historically, respondent confidentiality was protected through various disclosure avoidance techniques, including suppression, noise infusion, and data swapping (Lauger et al., 2014; McKenna, 2018). The census data that users know and depend on, even decennial “counts,” have for decades had some level of information reduction and perturbation included to protect privacy. Given recent technological advancements and the proliferation of data from third party sources, the Bureau determined after Census 2010 that a modernized system of disclosure avoidance is necessary to avoid the potential for attackers to reconstruct Census data and identify respondents (Abowd, 2019). The Bureau committed to publishing 2020 Census data using a new “formally private system built on a set of differentially private algorithms” (US Census Bureau, 2018), known as “differential privacy” (DP).

Differential privacy attempts to balance data accuracy with privacy, recognizing trade-offs between the two principles. It is based on a mathematical system of quantifying privacy risk and building a synthetic dataset that, while maintaining data structure, introduces formal perturbations to the raw data where the amount of privacy leaked can be quantified and set so as not to exceed a certain threshold. This threshold – referred to as epsilon, or the Privacy Loss Budget (PLB) – for Census 2020 is to be set by the Data Stewardship Executive Policy Committee in 2021. The committee allocates a specific amount of the privacy loss budget for each variable-by-geography combination, so that combinations with greater allocations of this budget will be more accurate than those with less budget. PLB for the PL 94-171 redistricting data file was set June 9, 2021. A decision on the PLB for the Demographic and Housing Characteristics (DHC file), which includes age-specific data necessary for net migration estimates, will come later.

Differentially private approaches are based in a growing body of mathematical theory and research but have not been applied previously at the scale of the US decennial census. Implementing a DP approach at this scale is ground-breaking work, the research and development for which the Census Bureau is leading (US Census Bureau, 2018). The level of complexity involved in applying such a technique, and determining all of the associated allocations of privacy loss budget to appropriately balance accuracy and privacy for the wide variety of data use cases associated with the US Census, is a monumental challenge.

The Census Bureau is currently working to determine how to allocate the privacy loss budget among various uses. They are doing so by reviewing and prioritizing existing “use cases” – uses of census data for specific purposes including political representation and redistricting, funding allocation, regulatory functions and meeting legal mandates, research, and planning. In an effort to engage with the user community, build transparency, and evaluate the impact of differential privacy applications for various uses, the Bureau has (as of May 20, 2021) published five versions of differentially private 2010 demonstration data that can be

compared to the published Census 2010 Summary File 1 results. The first four versions maintained the same PLB, but improvements to processing algorithms have continually been made to reduce error (US Census Bureau, 2021a). The April 2021 version included an increased PLB to $\epsilon = 12.2$, which is almost three times larger than prior demonstration products and shows considerable increases in data accuracy (US Census Bureau, 2021b).

The purpose of this research brief is to evaluate the impacts of differential privacy application on one important and widely used “use case”: *Net Migration Estimates by Age, Sex, Race, and Hispanic-Origin for US Counties*.¹ For each decade since the 1950s, demographers have generated high-quality net migration estimates by age, sex, and race for US counties using decennial census data as the starting and ending populations. Before the series can be extended to 2010–2020, it is critical to know the accuracy of the differentially private Census 2020 data and, correspondingly, what level of accuracy we can expect in the net migration estimates. In estimating these impacts, we aim to (1) inform decision making in the US Census Bureau for setting and allocating the PLB in the DHC file; (2) evaluate the level of detail (age-race-ethnic breakdown) that can be provided accurately in the 2010–2020 net migration estimates; (3) draw attention to the potential for DP errors to cluster in space; and (4) raise awareness among data users more broadly who are trying to understand the potential impact of DP implementation on data quality for various uses.

Net Migration Estimates for US Counties

As noted above, demographers have generated net migration estimates by age, sex, and various race and ethnicity categories for US counties from the 1950s through the 2000s (Bowles et al., 1965 and 1977; Fuguitt et al., 1993; Voss et al., 2004; White et al., 1992; Winkler et al., 2013a). Currently, our research team is working to extend this series by producing a similar dataset for the 2010–2020 decade. The resulting series will be a longitudinal dataset of 70 years of age-specific net migration for all US counties. The existing data are available for download through ICPSR and are served to the public on an interactive website at www.netmigration.wisc.edu where users can generate custom maps and charts, and download data. The series has been widely used among demographers studying patterns of migration, applied demographers generating population estimates and projections, and planners and analysts in the business community. The 2000–2010 data/documentation has been downloaded from ICPSR over 700 times since their release in 2013 (ICPSR, 2020), and the interactive website has had over 200,000 visits including over 10,500 data downloads.

Estimates for the 2010–2020 decade, as in prior decades, are being derived using a forward cohort residual method, subtracting a measure of natural increase from

¹ We use the word “Hispanic” following the language used in the decennial census questionnaire and in Census Bureau reporting, to refer to people of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin regardless of race.

population change over the period to generate net migrants. For past decades, this process has produced highly accurate and reliable data given that they are generated using high-quality data that avoid sampling error, including enumerations from the decennial censuses conducted at the start and end of each decade and administrative records of US births and deaths from the National Center for Health Statistics. As such, the 2010–2020 estimates rely on Census 2020 for a count of the population at the end of the decade in each county. These final populations will be compared to an expected population, which the team is constructing by aging the starting population observed at Census 2010 forward ten years, adding in new births, and subtracting out deaths. In this approach, the difference between the final Census 2020 population and the expected 2020 population must be due to net migration. The central concern is that if differential privacy substantially alters the published Census 2020 counts at the county level, then the net migration estimates will be affected and their accuracy diminished.

The net migration data are unique in that they offer age-specific estimates of net migration at a relatively small scale (the county) for the entire country. Migration is highly selective by age, and age patterns of migration vary tremendously across space (Johnson & Winkler, 2015). Age-specific migration's spatial distribution influences the age structure of communities, with significant implications for population aging, child health, education, and economic well-being. Selective migration by age, sex, and race/ethnicity alters counties' population composition, having significant implications for health service provisioning and staffing, service infrastructure development, economic and labor market conditions, fertility, and mortality. Similar-quality data are not available from other sources. Migration estimates derived from the American Community Survey are the closest, but those estimates cannot be broken down by detailed demographic characteristics at the county level given the relatively small sample size and related margins of error.

Data/Methods

This study evaluates the impact of differential privacy on the accuracy of net migration estimates by age and Hispanic-origin for US counties. We take a counterfactual approach, comparing the net migration estimates published for the 2000–2010 decade with a hypothetical set of DP net migration estimates produced using differentially private Census 2010 data. In our approach, we compare the published estimates to what they would have been had the end-of-decade population been subjected to differential privacy. Given that this is the same scenario currently facing the Census 2020 counts, our approach should offer a realistic estimate of the impacts of differential privacy on the anticipated 2010–2020 net migration estimates.

In our analysis, we address the following questions:

1. How different (on average) are DP versions of net migration estimates than the originally published net migration estimates by five-year age group?
2. To what extent are Hispanic net migration estimates affected by DP application?

3. In how many counties are the DP estimates fit for use versus not fit for use, and how is fitness distributed by population size and across space?
4. To what extent does DP preserve the age signature in a county's net migration estimates?
5. Would collapsing estimates into broader age groups (~20 years vs. 5 years) make DP estimates fit for use?

Data

There are three data sources required for comparison. First, we use published county-level net migration estimates by age, sex, race, and Hispanic origin for the 2000–2010 decade (Winkler et al., 2013a). We compare these original estimates to similar ones constructed using two sets of differentially private Census 2010 demonstration data released by the US Census Bureau in 2019 and 2020. The first demonstration data were released in Fall 2019, known as Sprint 1. The Bureau released the second data metrics in May 2020 (Sprint 2) and then re-released them as privacy-protected microdata in July 2020 (Van Riper et al., 2020). Although the Bureau released additional DP demonstration data in September and November 2020 and in April 2021, these most recent releases did not include age breakdowns, other than over/under 18. Consequently, we are not able to evaluate the most recent releases in terms of how DP would impact age-specific net migration estimates given the finer age granularity of the migration estimates.

Processing

We organized DP demonstration data into the same age groups as the original net migration estimates, applying the same process that was used to structure the 2010 final populations when the team created the 2000–2010 net migration estimates. Although the estimates have been published with more detailed race/ethnic and sex breakdowns, we have chosen to focus first on total net migration by five-year age group and then to examine data on the Hispanic population by five-year age group. This approach enables us to gain a sense of how race/ethnic breakdowns could be impacted by DP implementation. In order to maintain consistency with the 2010 migration estimates' final population, the DP data were adjusted for net undercount and overcount. Drawing on findings from the US Census Bureau's 2010 Demographic Analysis and Census Coverage Measurement studies, Winkler et al. (2013b) derived adjustment percentages specific to each sex and five-year age group for the migration estimates' race and ethnicity categories. We apply those same adjustments to the DP data in our analysis. Finally, we exclude from this analysis counties that experienced boundary changes between 2000 and 2010 ($n=109$, including Virginia independent cities), yielding a total of 3133 valid observations.

This harmonization generates a DP final 2010 population analogous to the final 2010 net migration estimates' population. We calculate a DP version of the estimate of net migrants by substituting the DP value for the Summary File 1-based 2010 population count. We similarly construct a DP version of the net migration *rate* (per

100 expected population). The expected population for 2010 is the same for the original version and the DP version, as this value relies on the starting population from Census 2000 and natural increase observed over the 2000–2010 decade.

The original formulae for calculating net migrants and net migration rates, where M = net migrants and R = net migration rate:

$$M = \text{final 2010 population} - \text{expected 2010 population},$$

$$R = (M/\text{expected 2010 population}) * 100.$$

The differentially private formulae for calculating net migrants and net migration rates, where DPM = differentially private net migrants and DPR = differentially private net migration rate:

$$DPM = DP \text{ final 2010 population} - \text{expected 2010 population},$$

$$DPR = (DPM/\text{expected 2010 population}) * 100.$$

Metrics for Comparison

To compare the hypothetical DP 2000–2010 net migration estimates to the actual ones, we use a series of metrics that are commonly used to evaluate error in population estimates and projections. These are summarized below and are similar to those the Census Bureau is using to compare differentially private Census 2010 data to released Summary File 1 Census 2010 data (US Census Bureau, 2020).

Median Absolute Error (MAE). Measures the “average” numerical difference between net migration estimates and the DP version of the same estimates. Calculated as $\text{Abs}(DPM - M)$ for each of the 3133 counties, then take the median. Or, for rates, $\text{Abs}(DPR - R)$ for each county, then take the median.

Median Absolute Percent Error (MAPE). Measures the “average” relative difference between net migration estimates and the DP version of the same estimates. Calculated as $[\text{Abs}(DPM - M)/M]$ for each of the 3133 counties, then take the median.

Median Algebraic Percent Error (MALPE). Identifies systematic bias by showing the average direction of error (+ or -) and provides an alternative relative measure of error. The MALPE is generally closer to zero than the MAPE, because no absolute value is taken, and so positive and negative errors cancel out one another. Calculated as $(\sum((DPM - M)/M))/N * 100$, where N = number of observations (3133 counties).

We chose to present the median versions, rather than the mean, because means are heavily skewed by a small number of outliers with extreme differences, particularly when it comes to relative error. These three measures – MAE, MAPE, and MALPE – allow us to answer our first two research questions about the difference (on average) between DP versions of net migration estimates and the originally published estimates, and to summarize differences by age and Hispanic origin.

The outlying cases of extreme differences (glossed over by the “average” metrics) are also important, because some counties could end up with wildly inaccurate net migration estimates and, consequently, problematic analytic conclusions and policy decisions. To answer our third research question concerning fitness of use, we identify where DPRs are particularly inaccurate in several ways. First, we report the number of counties where the difference between published and DP net migration rates is <5, 5–9, 10–14, 15–19, and 20 or more by five-year age group. Because these are rates per 100 expected population, a difference of more than 5 percentage points in a rate could lead to meaningful misinterpretations and presents a challenge to the data’s fitness of use. To ease interpretation, we use a difference in the net migration rate of >5 as an indicator that the DP version is not fit for use. For example, in Mercer County, Ohio (population 40,814) the expected population at age group 65–69 is 1,576. The published estimate of net migrants age 65–69 is 117 (NMR=7). The DP estimate is 193 net migrants (DPR=18), such that the difference in rates is 11 percentage points and, thus, not fit for use in many applications, such as making population estimates or projections.

Next, we focus on error distributions by county population size for two select age groups: age 25–29 and age 65–69. Young adult migration is important for planning and community development, and migration propensities are highest among young adults. Correspondingly, we see DPRs performing relatively well at age 25–29. Retirees (age 65–69) are another important age group for migration that exhibit very different spatial patterns in comparison to young adults (Johnson & Winkler, 2015). We investigate interquartile and 90th percentile ranges of the absolute error between originally published NMRs and DPRs for these two age groups to offer a sense of the range of error by county population size. Then, we map counties by absolute error in NMRs for these age groups and run local indicators of spatial autocorrelation (LISA statistics) to explore the possibility of spatial clustering in error.

Rather than focusing on any one age group, many net migration estimates’ data users are interested in each county’s specific age pattern of migration, referred to as a county’s “net migration signature” (Johnson & Winkler, 2015; Johnson et al., 2005). Our fourth research question concerns how DP could impact counties’ signature pattern of age-specific migration. We evaluate this possibility by calculating a correlation coefficient (Pearson’s R) between each county’s series of age-specific migration estimates as originally published in comparison to the DP version. We then review summary statistics of these correlations.

Combined, these analyses reveal concerns with using DP Census 2020 data to generate net migration estimates by five-year age group. Therefore, we explore our fifth research question concerning one possible remedy: collapsing age groups into broader categories and recalculating metrics to assess the extent to which aggregation by age could improve data fitness of use.

Table 1 Difference between DP estimates and published net migration estimates

Age	<i>Total population</i>				<i>Hispanic population</i>			
	No. of net migrants			NMR	No. of net migrants			NMR
	MAE	MAPE (%)	MALPE (%)		MAE	MAPE (%)	MALPE (%)	
0–4	88	100	–30.1	4.8	27	77.6	–27.9	30.3
5–9	65	42.2	–5.2	3.5	24	57.1	–8.6	38
10–14	60	44.4	0	3.3	21	83.3	–1.6	31.6
15–19	71	42.5	0.2	3.5	22	75	–5.2	37.5
20–24	68	14.6	–0.7	3.5	23	64.3	–16.1	43.2
25–29	74	20.3	0.9	3.5	25	72.1	–22.2	43
30–34	71	31.2	–2	4.3	23	91.9	–20	37.7
35–39	67	40.3	–2.5	4	20	100	–22.9	39.8
40–44	63	51	–0.7	3.4	19	123	–20	40.1
45–49	65	55.7	–0.5	3	18	143.1	–28.5	40.5
50–54	67	58.8	4.5	3	15	166.7	–18.2	43.5
55–59	69	63.3	0.6	3.4	13	200	–22.8	50
60–64	70	53.8	7.7	4.1	10	207.1	–4.9	60
65–69	63	55.6	5.3	4.9	9	250	–27.6	73.5
70–74	57	80	6.1	5.5	7	300	–23.1	84.8
75–79	53	118.5	8.8	6.6	5	300	–33.3	100
80–84	50	120.6	6.2	8.2	4	366.7	–35.8	100
85+	84	89.7	–22.5	12.3	6	266.7	–60	100
Total	1029	4.63	–0.1	0.4	51	19	–6.3	5.7

MAE, MAPE, and MALPE are calculated as medians. DP estimates rely on Sprint 2 Demonstration Microdata, released July 2020

Results

Differences Between Published and DP Net Migration Estimates

The differences between the DP version of net migrants and the original estimates are substantial. Table 1 shows median absolute errors (MAE), median absolute percent errors (MAPE), and median algebraic percent errors (MALPE) by five-year age group and Hispanic origin using the Sprint 2 demonstration microdata release from July 2020. Median differences in the total number of net migrants range from 50 at age 80–84 to 88 net migrants at age 0–4. In relative terms, MAPEs range from 20% at ages 25–29 where migration is more common (larger denominator) to over 100% at the youngest and oldest ages. MALPE values indicate that the median direction of error varies by age group, but that DP estimates tend to underestimate net migrants at the youngest and oldest ages. Median percentage point errors in net migration rates generally range from 3 to 5 per 100 for most age groups and are higher at older ages.

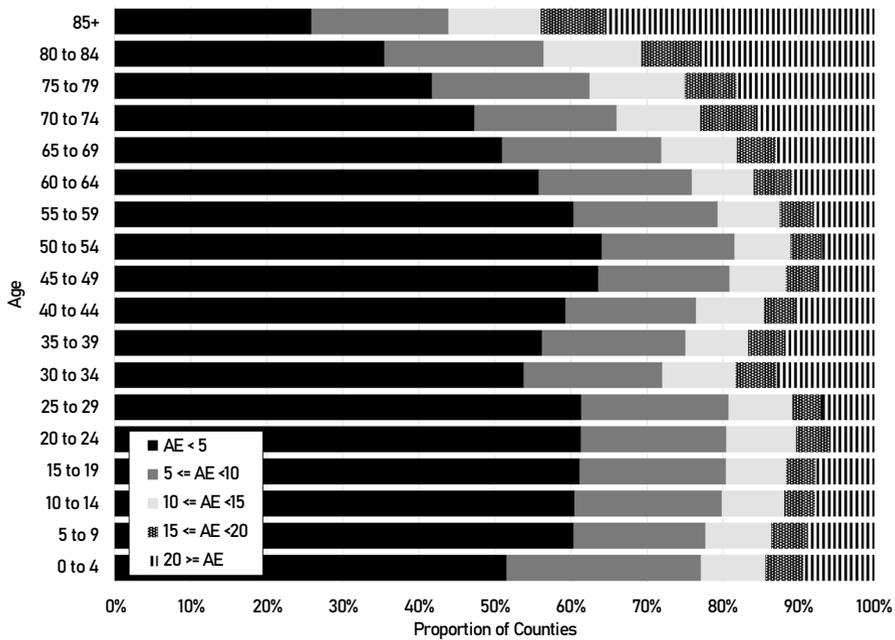


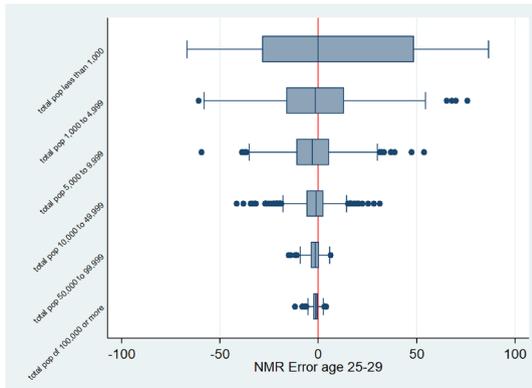
Fig. 1 Distribution of absolute error in net migration rates (DPR—NMR) by age group

Effect on Hispanic Estimates

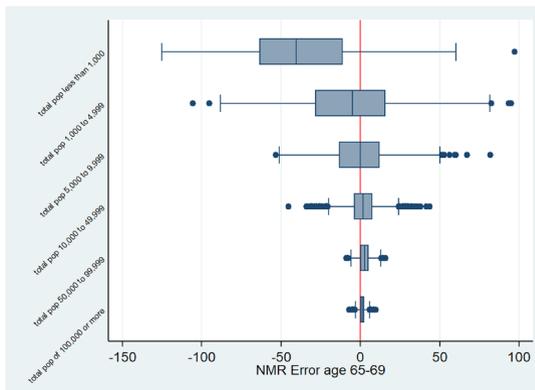
Our results show that errors are considerably higher for Hispanic migration estimates. Median relative errors on the number of net migrants by five-year age groups range from 57% to over 300%. NMR errors average between 30 and 50 percentage points for most age groups and are larger at older ages where the number of Hispanics is much smaller. MALPE values are consistently negative for Hispanics across age groups, indicating a negative bias. This means that the differentially private net migrant estimates are consistently lower than estimates from the original data version. Overall, the findings demonstrate that the most recent DP demonstration data broken down by five-year age group and Hispanic are not fit for use.

Number and Types of Counties Most Affected

Beyond these “average” differences, it is important to recognize the number and types of counties where errors are extreme. These are the cases in which applying differential privacy could have the greatest impact on data interpretation and related decision making. Figure 1 shows the distribution of absolute errors in net migration rates across counties by five-year age group. About half of all counties have DP rates that may not be fit for use, with errors greater than 5 percentage points for five-year age groups. About 25% of counties show errors greater than

a Age 25–29

Note: Five counties with NMR errors greater than 100 are excluded from this chart.

b Age 65–69

Note: Eight counties with NMR errors greater than 100 are excluded from this chart.

Fig. 2 Distribution of error in net migration rate (DPR–NMR) by county population size. **a** Age 25–29. Five counties with NMR errors greater than 100 are excluded from this chart. **b** Age 65–69. Eight counties with NMR errors greater than 100 are excluded from this chart

10 percentage points, and about 10% of counties have errors of more than 20 percentage points, indicating that the original rate is 20 per 100 population different from the DP version and could lead to seriously erroneous interpretations. For example, Iron County, Wisconsin (population 5916 in 2010) had a rate error of 47 percentage points at age group 25–29. This county saw a net *out*-migration of 197 young adults according to the originally published estimates (NMR = -47 per 100), while the DP version estimates a net *in*-migration of 1 young adult (NMR = 0.25 per 100). Even more counties have questionable or unfit DP rates

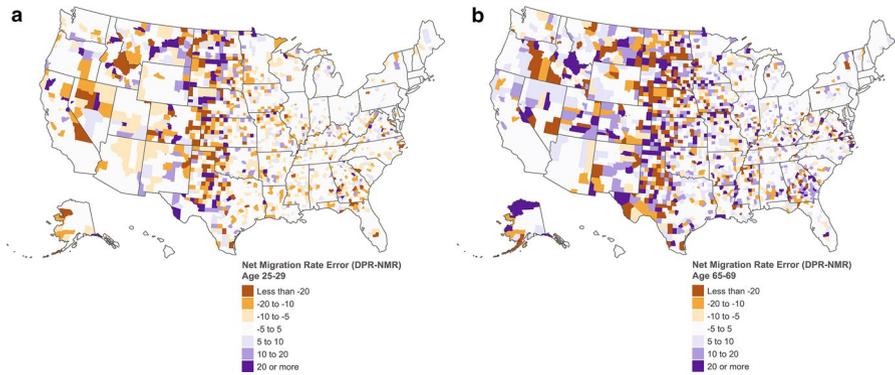


Fig. 3 Spatial distribution of error in net migration rate (DPR-NMR) by county. **A** Age 25–29. **B** Age 65–69

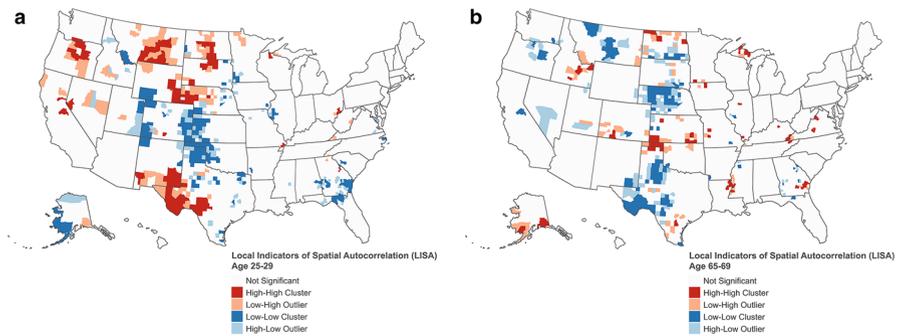


Fig. 4 Spatial clustering of error in net migration rates. **A** Age 25–29. **B** Age 65–69. LISA statistics calculated by authors using inverse distance-squared weights matrix. Statistical significance is based on a corrected 95 percent confidence level (Sensitivity analysis with alternative weights matrices (first-order Queen contiguity and inverse distance) revealed similar results.)

at ages 65 and over, and errors get worse as age increases so that almost 40% of counties have errors over 20 percentage points at age group 85 plus.

Figure 2 illustrates the distribution of net migration rate errors by county population size at ages 25–29 and ages 65–69. For all county size groups at ages 25–29, median errors are near zero, suggesting that errors cancel one another out and there is no systematic difference by county size between DP and original versions of net migration estimates. In contrast, for ages 65–69, DP rates generally underestimate net migration in smaller population counties and overestimate net migration in larger population counties. For both age groups, interquartile ranges for errors are small for counties with a population of 50,000 or more, and in these higher population counties, there are few outliers with extreme errors. However, several counties with smaller populations, even those in the 10,000 to 49,000 range, have errors that call into question the accuracy and usability of the DP rates.

Figure 3 shows the spatial distribution of these errors in net migration rates for ages 25–29 and 65–69, and Fig. 4 shows the propensity for the error to cluster in space using LISA statistics. Among both 25–29 and 65–69 year olds, extreme rates are most common in more rural areas, particularly across the Great Plains. Errors for both age groups are especially clustered in the Great Plains, with groups of counties that both significantly underestimate and overestimate net migration rates across the region. High-high clusters are places where DP rates overestimate net migration for those counties and their neighbors, whether underestimating the rate that people are leaving or overestimating the rate that people are moving in. Low-low clusters are counties where DP rates underestimate net migration for those counties and their neighbors (i.e., counties where DP overestimates the rate of people leaving and/or underestimates those moving in). Low-high and High-low outliers are counties with low error relative to their neighbors.

Our results demonstrate systematic spatial clustering in the impact of differential privacy. Global tests of spatial autocorrelation (Moran's i) indicate that errors are not statistically significantly autocorrelated (age 25–29, $i=0.004$; age 65–69, $i=0.012$). Although the lack of global spatial autocorrelation is encouraging, local indicators of association show significant spatial clustering in some regions and demonstrate an uneven bias in estimates across the nation (see Fig. 4). Among age group 25–29, DP rates overestimate net migration in areas of both the northern and southern Great Plains but underestimate net migration in the central Great Plains (Kansas, Oklahoma, eastern Colorado). DPRs also underestimate net migration rates in parts of the Intermountain West, and in parts of southern Georgia and the Florida panhandle, but overestimate net migration in parts of rural Oregon and California. Among age group 65–69, it is nearly the opposite, with DPR underestimating net migration in the northern and southern Great Plains and overestimating net migration in the central Great Plains, the Mississippi Delta, southeast Georgia, some parts of Appalachia, and Michigan's western Upper Peninsula.

Comparing Net Migration Signatures

County migration signatures show unique patterns of in-/out-migration rates that are correlated across age groups. For instance, central cities tend to attract young adults but lose middle-aged adults, children, and retirees, while suburban counties often attract middle-aged adults and children but lose young adults; and rural retirement destination counties attract retirees but lose young adults (Johnson & Winkler, 2015). We analyze the extent to which the DP migration estimates accurately reflect the correlation in migration patterns across age groups.

Across all counties, the average correlation coefficient (Pearson's R) comparing original net migration age signatures by five-year age group to the DP versions we constructed was relatively high at 0.80, but with considerable deviations. The signatures we compared included age groups 0–4 through 80–84 and excluded the terminal age group (85+) due to inconsistencies and uncertainties in net migration estimates at the most advanced ages. Our findings show that 42% of counties ($n=1329$) had an $r>0.9$, indicating strong agreement in the signatures. However, 15 counties

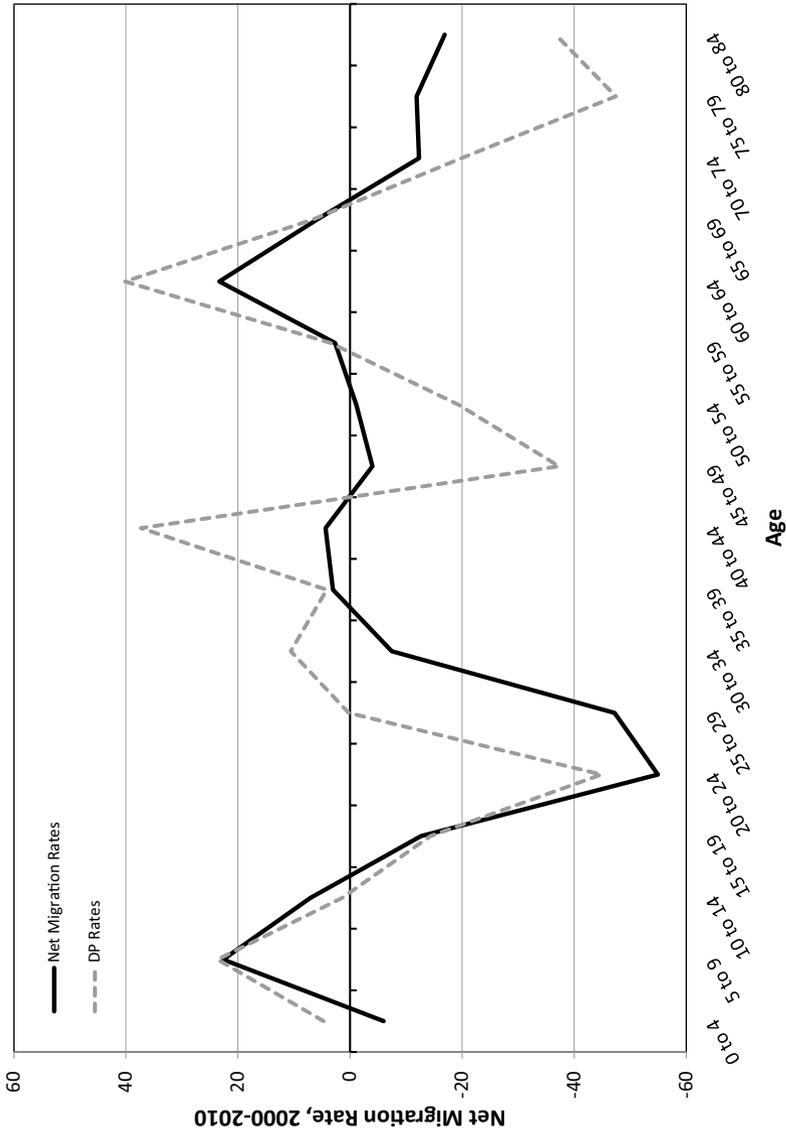


Fig. 5 Net migration signature comparison- Iron County, WI

(0.5 percent) had a negative correlation, which was unexpected and suggests profound disagreement between the DP and NMR signatures. An additional 304 counties (10 percent) had $r < 0.5$, indicating little agreement between the DP and NMR signatures. As in prior analyses, larger population counties performed better than smaller population counties. Figure 5 shows an example for Iron County, Wisconsin (discussed previously), where $r = 0.61$. While the DP signature is generally similar to the original net migration signature, the DP version exhibits higher peaks and valleys and a more erratic migration pattern.

Collapsing Across Age Groups

The error observed in the information presented thus far calls into question the fitness for use of net migration estimates by five-year age group. We investigate how fit might be improved by collapsing estimates into broader age groups. While migration is highly selective by age, it may be worth losing some granularity in order to provide more accurate estimates. The www.netmigration.wisc.edu website, which serves data from prior decades, includes interactive maps using age groups 0–14, 15–24, 25–29, 30–54, 55–74, and 75 plus. These broader groupings combine ages that tend to have similar place-based migration patterns.

Table 2 shows differences between DPRs and published rates by county population size, noting the median absolute error in percentage points and the percent of counties with absolute rate errors over 5 per 100 (unfit for use, as we have defined here). Although error is generally reduced, it remains problematic for smaller population counties and for some age groups, especially older age groups. For counties overall, median absolute errors range from 1.5 percentage points for age group 30–54 to 5.8 for age 75 plus. At the oldest ages, rates would be unfit for use in 54% of counties. Even for age group 30–54, 18% of counties still show rates unfit for use. In counties with populations less than 10,000 ($n = 676$), the majority of counties show DPRs that would be unfit for use. For example, 51% of counties with populations between 5000 and 9999 would be unfit for use for age group 55–74, which is a critical age for understanding retirement migration patterns in rural America. Even counties with populations greater than 50,000 or 100,000 show considerable proportions with unfit rates at age group 75 plus (37% and 13%, respectively).

Conclusions

Overall, our findings indicate a substantial and concerning impact of differential privacy on county-level net migration estimates. Median differences in the number of net migrants by five-year age group attributable to DP range from 50 to 88. Relative differences are also high, with median absolute percent errors ranging from 15% (for age group 20–24) to over 100% at ages 75 and over. Percentage point differences in net migration rates are particularly concerning. According to our estimates, about half of all counties would have net migration rates with errors over 5 percentage points and, thus, would not be fit for use at any given five-year age group. Moreover,

Table 2 Net migration rate errors for broader age groups by county population size

Age group	Total		Pop < 5000		5000 to 9999		10,000 to 49,999		50,000 to 99,999		100,000+	
	MAE	% with AE > 5	MAE	% with AE > 5	MAE	% with AE > 5	MAE	% with AE > 5	MAE	% with AE > 5	MAE	% with AE > 5
0-14	1.8	12.7%	7.1	61.5%	3.1	29.1%	1.8	6.8%	1.5	0.0%	1.6	0.0%
15-24	2.4	27.7%	11.3	77.0%	6.4	61.6%	2.8	25.3%	1.4	4.7%	0.8	0.7%
25-29	3.5	47.9%	15.8	89.3%	8.8	81.3%	4.3	55.2%	2.3	21.1%	1.5	3.2%
30-54	1.5	17.5%	10.2	73.5%	4.1	42.1%	1.8	10.8%	0.7	1.0%	0.4	0.0%
55-74	2.6	26.2%	10.8	73.2%	5.1	51.2%	2.9	25.6%	2.2	5.7%	1.1	0.5%
75 plus	5.8	54.3%	26.8	91.1%	16.7	83.1%	6.6	59.9%	3.8	36.8%	1.7	13.1%

MAEs are calculated as medians. DP estimates rely on Sprint 2 Demonstration Microdata, released July 2020. County Population Size at Census 2010

88% of counties have at least one five-year age group where data would not be fit for use. Errors are especially high for populations aged 65 and over and for the Hispanic population. The findings indicate that age breakdowns by Hispanic origin at the county level may not be feasible under current DP implementation plans.

Most concerning are the cases where we see extreme differences, numerically and relatively, when comparing the DP version of estimates to the originally published data (about 10% of counties). Such errors would seriously challenge the accuracy of the 70-year net migration data series, undermining their utility. Moreover, large errors (as in the Iron County, WI example) could lead to critically erroneous decision making and planning with detrimental implications for counties and their residents. The relative lack of consistency in the net migration signatures raises additional concern, with about 37% of counties showing a less than 0.8 correlation between the DP version and published rates.

We find that inaccuracies are greatest in smaller population counties. Yet, errors are sizeable in many counties with populations ranging from 10,000 to 50,000. Collapsing into broader age groups reduces error, but this strategy does not eliminate or adequately address concerns. Altogether our findings are similar to Hauer and Santos-Lozado's evaluation using age-specific DP data for calculating county-level covid mortality (Hauer and Santos-Lozado, 2021)

We also find evidence of systematic error by age, race, and geography. DP estimates tend to underestimate net migrants at the youngest and oldest ages, and among Hispanics at almost all ages. Moreover, there is evidence of systemic spatial clustering in some regions. The negative implications of drawing extraneous interpretations based on inaccurate data are exacerbated by DP results systematically under- or overestimating net migration in entire regions of the country. For example, if DP versions of net migration rates overestimate net migration at age group 25–29 across the northern Great Plains, the extent of young adult out-migration, and associated planning responses, from that region may be unclear. Similarly, if net migration among those ages 65–69 in the Mississippi Delta is overestimated, it may appear that these counties are attracting more retirement migrants than they actually are, potentially leading to extraneous investments in services or community and economic development strategies. In addition to applied consequences, erroneous estimates will challenge scholars' abilities to empirically assess the causes and consequences of age-, sex-, and race-specific migration patterns in the contemporary United States. To our knowledge, there has been little attention paid to the potential for problematic spatial clustering in the application of DP more broadly. Finding evidence of the problem here serves as a warning signal that it could apply to other census data applications as well. Future analyses of various DP applications should closely study the potential for spatial clustering of error.

In this analysis, we have only had the opportunity to compare Sprint 1 and Sprint 2 versions of the DP application to Census 2010. Various users have discovered substantial differences between the DP estimates in Sprint 1 and the published 2010 data, and raised serious concerns at a National Academies workshop in December 2019 (National Academies, 2019) as well as through feedback provided to the Census Bureau. Our analysis of the Sprint 2 version indicates limited improvement in comparison to Sprint 1. The Bureau reports further improvements

to algorithms in more recent releases but, regrettably, these releases do not include the five-year age groups necessary for the net migration estimates.

Ultimately, the results shown here could change dramatically, depending on final decisions the Bureau's Data Stewardship Executive Policy (DSEP) committee makes about how to implement DP, including where to set the privacy loss budget (PLB) and how to allocate it across variables and geographies. The Bureau's recent indication that they will release redistricting data with an $\epsilon = 19.6$, which indicates a PLB almost five times greater than was used in the demonstration data evaluated here (US Census Bureau, 2021b) suggests that the errors observed in this study may overestimate the ultimate impact of DP on the net migration estimates program. Therefore, we might think of the estimates we present as a worst case scenario for data accuracy. Still, the Bureau has not yet indicated how they will handle the PLB for the Demographic and Housing Characteristics (DHC) file, which includes the age, sex, and race/ethnicity breakdowns that will inform net migration estimates as well as various other uses. The fact that the Bureau found they need to raise the epsilon for the PL file to 19.6, in order to meet accuracy standards for redistricting in the PL 94-171 data release, suggests that an even larger epsilon will be required to produce reasonably accurate data in the DHC file with much more detailed data. Yet, as the epsilon increases, privacy diminishes. This will likely be a source of tension as DSEP considers how to implement DP in the DHC file.

Acknowledgements We thank the Integrated Public Use Microdata Series (IPUMS) and the National Historical Geographic Information System (NHGIS) teams for sharing the 2010 demonstration products in a format accessible for the data user community. This paper was presented at the Population Association of America's Applied Demography Conference on February 3, 2021 in a session that was moderated by David Van Riper. We thank David and the other conference participants (including Bill O'Hare) for their feedback and questions during that session. David Van Riper also provided code for processing privacy-protected microdata files. Ken Johnson and John Cromartie contributed to discussions around the impact of differential privacy on the net migration estimates series, and we thank all of our colleagues in the W4001 Multistate Research Group who have contributed to our understanding of differential privacy and its implications. Participating in the Census Scientific Advisory Committee's Differential Privacy Working Group and the presentations offered there by Census Bureau experts were also important for developing greater understanding of differential privacy. Caitlin Bourbeau at the UW-Madison's Applied Population Laboratory designed the maps reported in this paper. Finally, we wish to thank Mathew Hauer for his consultation on working with the privacy-protected microdata file.

Author contributions Winkler conceptualized the project, contributed to data organization and data analysis, drafted the original manuscript, and made edits/revisions. Butler served as chief data analyst, including data organization and analysis, and made edits to the manuscript. Curtis serves as principal investigator on the broader related project, consulted on conceptualization and analysis throughout, and made edits to the manuscript. Egan-Robertson consulted on conceptualization and analysis, and made edits to the manuscript.

Funding This research is supported by the Eunice Kennedy Shriver National Institute of Child and Human Development (NICHD), grant number 1R03HD100772-01 and by a cooperative agreement with the USDA Economic Research Service (agreement 58-6000-9-0043). It is further supported by the University of Wisconsin's Center for Demography and Ecology (funded by a Population Research Infrastructure center grant P2C HD047873 from NICHD), the Wisconsin Agricultural Experiment Station, the W4001 Multistate Research Project funded by Western Association of Agricultural Experiment Directors (WAAESD), and the Population Research Institute at Pennsylvania State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice

Kennedy Shriver National Institute of Child Health and Human Development, the National Institutes of Health, the United States Department of Agriculture, or any other organization.

Data availability The data upon which this analysis is based are publicly available. County-level Net Migration Estimates by Age, Sex, Race, and Hispanic-Origin 2000–2010 are available at www.netmigration.wisc.edu and are archived at the Inter-University Consortium for Political and Social Research (ICPSR). The US Census Bureau released differentially private 2010 demonstration data with regular updates published on the blog “2020 Disclosure Avoidance System Updates.” We primarily relied on the tabulated demonstration data published for public access by IPUMS NHGIS based on the privacy-protected microdata files which were publicly released by the US Census Bureau in July 2020.

Code availability Data were processed and analyzed using Stata Statistical software. Code (.do files) is available upon author request.

Declarations

Conflict of interest Winkler serves on the Census Scientific Advisory Committee of the US Census Bureau. Her efforts on this paper are in her role as a professor at Michigan Technological University and do not necessarily represent the views of the CSAC or of the Census Bureau.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abowd, J. (2019, February 16). Staring down the database reconstruction theorem. Presentation to the American Association for the Advancement of Science Annual Meeting Saturday. <http://www2.census.gov/programs-surveys/decennial/2020/resources/presentations-publications/2019-02-16-abowd-db-reconstruction.pdf>. Accessed 3 Jan 2021.
- Bowles, G. K., Tarver, J. D., Beale, C. L., & Lee, E. S. Net migration of the population by age, sex, and race, 1950–1970. (1965). Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR08493.v2>. Accessed 5 Sept 2016.
- Fuguitt, G. V., Beale, C. L., & Voss, P. R. County-specific net migration estimates, 1980–1990 [United States]. (1993). Inter-university Consortium for Political and Social Research [distributor]. 2010-04-02. <https://doi.org/10.3886/ICPSR26761.v1>. Accessed 4 Feb 2010
- Hauer, M. E., & Santos-Lozada, A. R. (2021). Differential privacy in the 2020 Census will distort COVID-19 rates. *Socius*, 7, 2378023121994014.
- ICPSR. (2020). <http://pcms.icpsr.umich.edu/pcms/reports/studies/34638/utilization>. Accessed 22 Sept 2020.
- Johnson, K. M., Voss, P. R., Hammer, R. B., Fuguitt, G. V., & McNiven, S. (2005). Temporal and spatial variation in age-specific net migration in the United States. *Demography*, 42(4), 791–812.
- Johnson, K. M., & Winkler, R. L. (2015). Migration signatures across the decades: Net migration by age in U.S. counties, 1950–2010. *Demographic Research*, 32, 1065–1080.
- Lauger, A., Wisniewski, B., & McKenna, L. (2014). Disclosure avoidance techniques at the U.S. Census Bureau: Current practices and research. RESEARCH REPORT SERIES (disclosure avoidance

- #2014-02). US Census Bureau. http://www.census.gov/srd/CDAR/cdar2014-02_Discpl_Avoid_Techniques.pdf
- McKenna, L. (2018, November). Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing. US Department of Commerce Report. <http://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf>
- National Academies of Sciences, Engineering and Medicine. (2019, December 11–12). Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations. Materials. http://www.sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518. Accessed 23 Sept 2020.
- Winkler, R., Johnson, K., Cheng, C., Voss, P., & Curtis, K. J. (2013). County-specific net migration by five-year age groups, hispanic origin, race and sex: 2000–2010. Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR34638.v1>. Accessed 9 May 2013.
- US Census Bureau. (2018). Disclosure avoidance. <https://www.census.gov/topics/research/disclosure-avoidance.html>. Accessed 21 Sept 2020
- US Census Bureau. (2020). Data metrics for 2020 Disclosure avoidance. <http://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-25-data-metrics-2020-da.pdf>. Accessed 21 Sept 2020.
- US Census Bureau. (2021a). Fine-tuning the disclosure avoidance system to ensure accuracy. 2020 Disclosure avoidance system updates blog. Published 2/3/2021. <http://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>. Accessed 4 Feb 2020.
- US Census Bureau. (2021b). Census Bureau Sets Key Parameters to Protect Privacy in 2020 Census Results. 2020 Disclosure avoidance system updates blog. Published 6/9/2021. <https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html>. Accessed 5 July 2021.
- Van Riper, D., Kugler, T., & Schroeder, J. (2020). IPUMS NHGIS Privacy-Protected Census Demonstration Data, version 20200527 [Database]. Minneapolis, MN: IPUMS. <http://www.nhgis.org/privacy-protected-demonstration-data>. Accessed 1 Nov 2020.
- Voss, P.R., McNiven, S., Hammer, R.B., Johnson, K.M., & Fuguitt, G.V. (2004). County-specific net migration by five-year age groups, hispanic origin, race, and sex, 1990–2000: [United States]. Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR04171.v1>. Accessed 23 May 2005.
- White, M. J., Mueser, P., & Tierney, J. P. (1992). Net migration of the population of the United States by age, race and sex, 1970–1980. Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR08697.v1>. Accessed 16 Feb 2011.
- Winkler, R. L., Johnson, K. M., Cheng, C., Voss, P. R., & Curtis, K. J. (2013a). County-specific net migration by five-year age groups, hispanic origin, race and sex: 2000–2010. Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR34638.v1>. Accessed 5 Sept 2013.
- Winkler, R.L., Johnson, K. M., Cheng, C., Voss, P. R., & Curtis, K. J. (2013b). County-specific net migration by five-year age groups, Hispanic origin, race and sex 2000–2010. CDE Working Paper No. 2013-04. Center for Demography and Ecology, University of Wisconsin-Madison. Madison, WI.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.