TEXAS CENSUS INSTITUTE

Undercounting and Overcounting U.S. Counties' Population

A Determinants-Side Approach and its Application to Texas

Francisco A. Castellanos-Sosa
The University of Texas at Austin
Texas Census Institute

December 2022
RR22-001

Undercounting and Overcounting U.S. Counties' Population

A Determinants-Side Approach and its Application to Texas

Francisco A. Castellanos-Sosa [A]

Abstract

The 2020 U.S. Census undercounted population in six states and overcounted in eight. On top of that, the substate undercounting and overcounting estimates will not be officially estimated by the U.S. Census Bureau due to sampling size limitations. This report presents a practical alternative to estimate undercounting and overcounting at the county level, using a proportionally weighted index with theoretical undercounting determinants and applies it to Texas' case. Findings suggest its counties' undercounting estimates are primarily present in the metropolitan areas and main counties along the U.S.-Mexico border and that counties' share of people in younger age groups, and Hispanic categories, is related to higher undercounting. Similarly, the Census self-response rate via the internet is related to undercounting. On the other hand, the share of people in older age groups and white categories is correlated with less undercounting. Moreover, the Census self-response rate under traditional methods—such as phone and mail—is related to less undercounting.

[A] Castellanos-Sosa: Texas Census Institute (francisco@texascensus.org).

Acknowledgments: The author appreciates the insightful support provided by Lloyd B. Potter, Monica Cruz, Mary Campbell, and Shannon Cavanagh.

Notes: This is an updated version of the report as of August 2023. Author's contact information and minor content details were updated.

1. Introduction

The U.S. Census Bureau is the institution in charge of carrying out the Census every ten years as it is mandated in the Constitution of the United States of America (U.S. Constitution, art. I, § 2). Counting every single person and locating them in the right place, whatsoever, is a challenging task. Despite deploying more than $14 billion to its implementation, the latest 2020 U.S. Census undercounted population in six states and overcounted in eight (U.S. Census Bureau, 2022a; U.S. GAO, 2021).

Undercounting and overcounting estimates can be obtained via the Post-Enumeration Survey (PES) and the Demographic Analysis (DA). The PES allows us to identify whether the counting in a state is significantly different from the original counting. On the other hand, for the entire country, the DA uses current and historical vital records, data on international migration, and Medicare records to produce national estimates of the population by age, sex, DA race categories, and Hispanic origin and compare them to those of the Census. Nevertheless, county- or city-level estimates are not provided by the 2020 PES because of its limited sample size to hold appropriate assumptions at substate levels (U.S. Census Bureau, 2021b).

To fill this gap, we propose a methodology and apply it to Texas to approximate the undercounting, or overcounting, of the population at the county level. This methodology is intended to provide a starting point in this respect and open a healthy discussion about further improvements and extensions. To do so, an examination of the plausible determinants of undercounting and overcounting (hereon referred to as U&O) is first presented. After that, proxy variables to measure the determinants are identified. Then, each county's share of U&O is approximated based on an equally-weighted index.

2. *A synthesized theoretical framework*

There is a vast difference between identifying the characteristics of those undercounted when having the data and estimating the undercounting with no direct data at hand. The former is performed by the U.S. Census Bureau's Demographic Analysis. The latter is the purpose of this report, and it is primarily based on identifying the undercounting by observing plausible determinants of why people are being undercounted. In this regard, undercounting in social science surveys has been largely studied, but scholarly work has no consensual framework yet (Clogg et al., 1989; de la Puente, 1995; King & Magnuson, 1995; Martin & de la Puente, 1993; O'Hare, 2019; Tourangeau & Plewes, 2013; West & Fein, 1990).

The literature around undercounting can be synthesized in a three-dimensional space, in which U&O is a function dependent on three main types of variables: *personal*, *geographical*, and *Census features*. At the same time, they could overlap. Moreover, each dimension might be formed by different factors. For instance, the *personal* dimension might embrace aspects related to a) social capital and b) social exchange. The *geographical* dimension might account for c) physical easiness-to-reach people in large agglomerations and d) accuracy in the Master Address File (MAF) records. The *Census features* dimension considers aspects related to the census implementation, such as e) marketing strategies or f) interviewer/technological accessibility. Table 1 groups the theory around these factors and explains how each factor might be related to a more accurate Census counting.

Table 1
Synthesized theoretical framework of undercounting and overcounting

Dimension	Factor	Reduces undercounting through…	Scholarly work sorted by year
1) Personal	a) Social Capital	…social trust and cooperative attitude	Putnam (1995); Heyneman, (2000); Letki (2006); Brick and Williams (2013)
	b) Social Exchange	…non-economic exchanges of intangible social satisfaction	Homans (1958); Dillman et al. (2009)
2) Geographical	c) Easiness-to-reach	…easy access to people intended to be approached	Martin and de la Puente (1993); Martin (2007)
	d) Accuracy in MAF	…including all households for accurate planning and implementation of the Census	Mahler (1993); Kissam (2017); Kissam et al. (2018)
3) Census features	e) Marketing strategies	…encouragement of people to participate	West and Fein (1990); Bates (2017)
	f) Int/tech accessibility	…easing people's participation	West and Robinson (1999); Sinclair et al. (2012); Olson et al. (2021)

Note: A deeper examination of surveys' nonresponse and undercounts in the U.S. Census is presented by Tourangeau and Plewes (2013) and (O'Hare, 2019).

The synthesized framework shown above provides different mechanisms through which each factor is associated with lower undercounting. Then, identifying a set of variables for these factors would let us have a comparison measure of them and the dimensions across counties. After that, we estimate an equally-weighted index with these variables to proxy the number of people undercounted in each county since the U&O data for the 2020 U.S. Census is available only for states.

3. Matching theory to data

3.1. Data

Each factor can be approximated by using indicators that capture the essence of each of them. In this regard, the data is gathered from several sources (see Table 2).

Table 2
Data Sources for the synthesized framework on undercounting and overcounting

Dimension	Factor	Variable	Data Source
1) Personal	a) Social Capital	i) Cohesiveness by clustering (%)	Chetty et al. (2022a) and Chetty et al. (2022b)
	b) Social Exchange	ii) Volunteering (%)	Chetty et al. (2022a) and Chetty et al. (2022b)
2) Geographical	c) Easiness-to-reach	iii) Population density (hundreds of people per km^2)	U.S. Census Bureau (2020)
		iv) Population share (%)	U.S. Census Bureau (2020)
	d) Accuracy in MAF	v) Addresses unable to be geocoded in the county (%)	U.S. Census Bureau (2021a)
3) Census features	e) Marketing Strategies	vi) ACS 5-year nonresponse rate by refusal (%)	U.S. Census Bureau (2022b)
	f) Int/tech accessibility	vii) ACS 5-year nonresponse rate by other than refusal (%)	U.S. Census Bureau (2022b)

Note: The latest data available is used for each variable.

We aim to capture the essence of the Social Capital factor with a measure of social trust and cooperative attitude. With that in purpose, we lean toward the cohesiveness approach of Chetty et al. (2022a) and Chetty et al. (2022b). They define Cohesiveness as "*The degree to which friendship networks are clustered into cliques and whether friendships tend to be supported by mutual friends*". Then, they measure Cohesiveness by clustering as the average fraction of an individual's friend pairs who are also friends with each other. Theoretically, Social Exchange is envisioned as those non-economic exchanges of intangible social satisfaction across individuals. Chetty et al. (2022a) and Chetty et al. (2022b) also measure volunteering in quantifying Civic engagement. We take their Volunteering variable as it captures "*the percentage of Facebook users who are members of a group which is predicted to be about 'volunteering' or 'activism' based on group title and other group characteristics*".

The Geographical dimension is composed of the Easiness-to-reach and Accuracy in MAF factors. We approximate the Easiness-to-reach dimension by using population density and population share each county represents in the state. Higher levels of population density are assumed to impose difficulties for the U.S. Census to be accurate. Similarly, large population counties would impose difficulty in counting all individuals. Another geography-related characteristic is the Accuracy of the Master Address File to identify every housing unit accurately. To measure this factor, we use the share of housing units unable to be geocoded by the U.S. Census Local Update of Census Addresses (LUCA) from the U.S. Census Bureau (2021).

The third dimension considers Census features and embraces the Marketing Strategies and Interviewer/technological accessibility factors. The first is measured using the share of the ACS 5-year housing unit nonresponse by refusal. It is expected, therefore, that higher refusal levels would increase undercounting. The second factor in this dimension is measured with the share of the ACS 5-year housing unit nonresponse by other reason than refusal U.S. Census Bureau (2022b).

Each of the variables used here is assumed to explain the level of undercounting or overcounting in a one-way relationship. In other words, when a variable increases, it is expected to either increase the likelihood of being undercounted or decrease it, but not both. This report will present the county-level estimates for Texas. Therefore, since Texas presented an undercounting, the following sections of this report will focus on the estimation of undercounting at the county level. Table 3 presents a one-way relationship for each of the variables with undercounting.

Table 3
Expected effect of each variable in undercounting

Variable	Expected Effect
i) Cohesiveness by clustering (%)	Less undercounting
ii) Volunteering (%)	Less undercounting
iii) Population density (hundreds of people per km^2)	More undercounting
iv) Population share (%)	More undercounting
v) Addresses unable to be geocoded in the county (%)	More undercounting
vi) ACS 5-year nonresponse rate by refusal (%)	More undercounting
vii) ACS 5-year nonresponse rate by other than refusal (%)	More undercounting

Note: The effect is the type of change expected regarding undercounting when each variable increases. This relationship was reviewed and determined by the Texas Census Institute Advisory Board.

3.2. Summary statistics

Cohesiveness by clustering and Volunteering have a "Less undercounting" relationship with undercounting. Then, they are modified to associate them directly with undercounting. Since they are percentages—on a scale from 0 to 100—a natural way to modify them is by using the distance of each of them to the maximum value. Population density is here expressed in hundreds of people per km^2, which makes them have the highest of 11.45 people per km^2. Since this number lies between the traditional 0 to 100 scale, we proceed with no further adjustments. The remaining variables are expressed in percentages of what they are intended to measure. Table 4 presents the summary statistics of the main variables.

Table 4
Summary statistics of main variables

Variable	Obs.	Mean	Std. Dev	Min.	Max.
i) Cohesiveness by clustering (%, distance to 100)	245	88.95	1.76	79.89	92.30
ii) Volunteering (%, distance to 100)	245	93.3	3.96	69.13	98.89
iii) Population density (hundreds of people per km^2)	254	0.45	1.32	0.00	11.45
iv) Population share (%)	254	0.39	1.42	0.00	16.51
v) Addresses unable to be geocoded in the county (%)	254	3.97	3.51	0.00	32.88
vi) ACS 5-year nonresponse rate by refusal (%)	254	6.73	4.40	0.20	36.90
vii) ACS 5-year nonresponse rate by other than refusal (%)	254	8.99	4.83	0.20	32.80

Note: There is no information for the following nine counties regarding Cohesiveness by clustering and Volunteering: Borden, Hartley, Kenedy, Kent, King, Loving, Motley, Roberts, and Terrel. The minimum values of Population density and Population share are not zero but a minimal number.

As expected, the population density and share variables are related, with a correlation coefficient of 0.9356. However, this relationship does not impose any statistical problem since both are used to measure the same Easiness-to-reach factor. The rest of the variables do not present a high correlation among them. Suggesting our selection of variables, factors, and dimensions is statistically appropriate and do not impose a substantial weight on any set of variables by double-counting them. Table 5 shows the correlation matrix of the variables used here.

Table 5
Pairwise correlation of main variables

Variable	i)	ii)	iii)	iv)	v)	vi)	vii)
i) Cohesiveness by clustering (%, distance to 100)	1.000						
ii) Volunteering (%, distance to 100)	-0.039	1.000					
iii) Population density (hundreds of people per km^2)	0.425	0.089	1.000				
iv) Population share (%)	0.371	0.096	0.936	1.000			
v) Addresses unable to be geocoded in the county (%)	0.266	0.149	0.112	0.085	1.000		
vi) ACS 5-year nonresponse rate by refusal (%)	0.008	0.040	0.005	0.020	0.018	1.000	
vii) ACS 5-year nonresponse rate by other than refusal (%)	-0.005	0.094	-0.046	-0.019	0.253	0.273	1.000

Note: There is no information for the following nine counties regarding Cohesiveness by clustering and Volunteering: Borden, Hartley, Kenedy, Kent, King, Loving, Motley, Roberts, and Terrel.

4. Methodology

4.1. Measuring a county-level index for undercounting and overcounting

The main goal of building an index for the counties is to use it to know how much of the state-level undercounting can be attributed to each county. Whatsoever, the total measure is a net value that might contain both overcounting and undercounting at the county level. Therefore, we take the 90% confidence interval of the official state-level undercounting as lower and upper boundaries. The proportionally-weighted index proposed here would capture only undercounting for the Texas case, for which the undercounting is -1.92% with a 90% confidence interval between -3.27% and -0.57%. Due to the focus on the Texas case of undercounting, the following methodological approach takes positive terms to express undercounting and negative for overcounting. The county-level index CLI for county c in state s at year t is estimated as shown in Equation 1.[1]

$$CLI_{c,s,t} = \frac{\sum_d D_{d,c,s,t}}{3} \tag{1}$$

Where $D_{d,c,s,t}$ is the dimension subindex for each of the dimensions $d = (personal, geographical, census\ feeature)$ for county c in state s at year t. It is divided by three since it is an equally-weighted index to avoid over-or under-weighting across factors. The subindex $D_{d,c,s,t}$ is estimated as in Equation 2.

$$D_{d,c,s,t} = \frac{\sum_f F_{f,d,c,s,t}}{2} \tag{2}$$

Where $F_{f,c,s,t}$ is the average of the variables in each factor belonging to each dimension for county c in state s at year t (social capital and social exchange for the personal dimension; easiness-to-reach and accuracy in MAF for the geographical dimension; and marketing strategies and int/tech accessibility for the Census features dimension). It is divided by two since it is an equally-weighted index to avoid over-or under-weighting across factors. The original variables are standardized by dividing the difference of each value with respect to the mean by the

[1] We acknowledge that this is a starting point and would appreciate it if the reader takes any interpretation with cautious, according to the assumptions described along this report.

variable's standard deviation. This way, the original variables are first used as standard deviation units to the state mean.

4.2. Estimation of the undercounting and overcounting

The $CLI_{c,s,t}$ is therefore summarizing the dimensions, which are, at the same time, embracing its factors. Then, Equation 1 and Equation 2 can be summarized as in Equation 3.

$$CLI_{c,s,t} = \frac{\sum_d \sum_f F_{f,d,c,s,t}}{6} \tag{3}$$

$CLI_{c,s,t}$ is an average of the mean-standardized version of the variables. Then, we adjust its distribution to match the official state-level undercounting. The adjustment is performed by dividing the $CLI_{c,s,t}$ by the county-level index's maximum (or minimum) value when it has positive (or negative) values. This allows a dispersion between -1 and 1, which is multiplied by the absolute figures of the difference between the upper and lower bounds of the 90% confidence interval and the mean undercount. Let us call the adjusted index $\widehat{CLI}_{c,s,t}$. Therefore, the undercount for county c in state s at year t is calculated as in Equation 4.

$$undercount_{c,s,t} = \left(0.0192 + \widehat{CLI}_{c,s,t}\right) \times population_{c,s,t} \tag{4}$$

Where $undercount_{c,s,t}$ represents the population being undercounted when positive and overcounted when negative.

5. Results

The average county-level undercounting share is 1.52%, with a minimum of 0.46% and a maximum of 2.64%. In terms of the number of undercounted people, we found counties have an undercount of 6 to 117,073, with an average undercount of 2,237 people by county. Table 6 presents the main summary statistics.

Table 6
Summary statistics of undercounting across Texas' counties by data handling method

Variable	Obs.	Mean	Std. Dev	Min.	Max.
Undercounting, %	245	0.015	0.003	0.005	0.026
Undercounting, population	245	2,237	9,445	6.000	117,073

Note: There is no information for the following nine counties regarding Cohesiveness by clustering and Volunteering: Borden, Hartley, Kenedy, Kent, King, Loving, Motley, Roberts, and Terrel.

Figure 1 presents the distribution for the county-level undercounting share and (the log of) undercounting estimates in Panel a) and Panel b), respectively. Both panels show that our estimates are not skewed. It is important to emphasize that the mean and 90% confidence interval from the official U.S. Census Bureau undercounting state-level estimates are used first to estimate our undercounting share measure. Then, the distribution of our estimates might be considered moderate—or conservative—since the real undercounting at the county level might go out of the 90% confidence interval.

a) b)

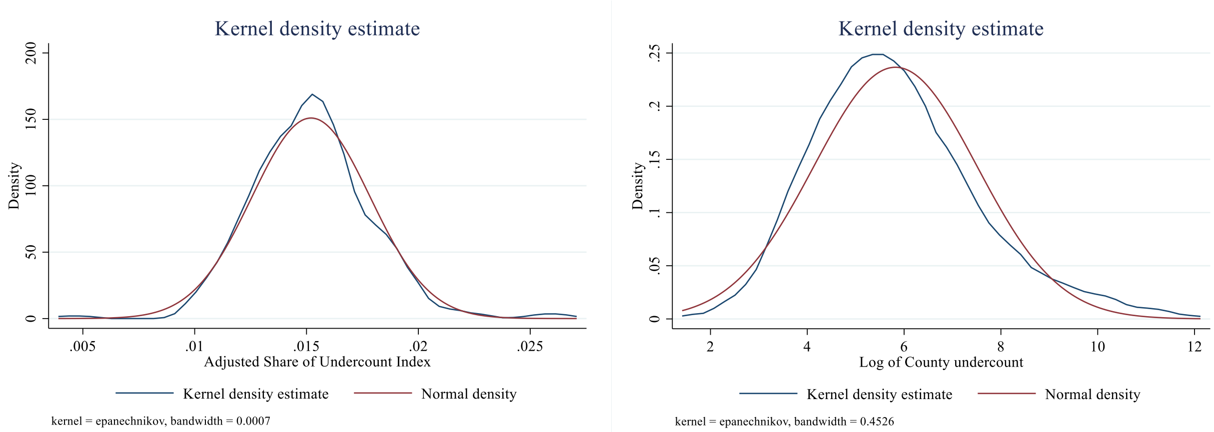

Figure 1 Distribution of the undercounting share and undercounting in Texas' counties.

Figure 2 presents the geographical distribution of undercounting across the 245 counties with available data for all the variables. Panels a) and b) in Figure 2 present each county's undercounting percentage and total values, respectively. The maps present seven bins for their colors. At first sight, Panel a) in Figure 2 depicts darker colors in densely populated areas (such as those of Austin, Dallas, Houston, and San Antonio) and the U.S.-Mexico border. This suggests that the intensity of undercounting (expressed in percentage terms) is higher in those areas. On the other hand, when studying the counties' undercounting estimates, Panel b) presents a different story. As intuitively expected, Panel b) in Figure 2 shows how undercounting is higher in volume terms in highly populated counties and just a few counties on the U.S.-Mexico border.

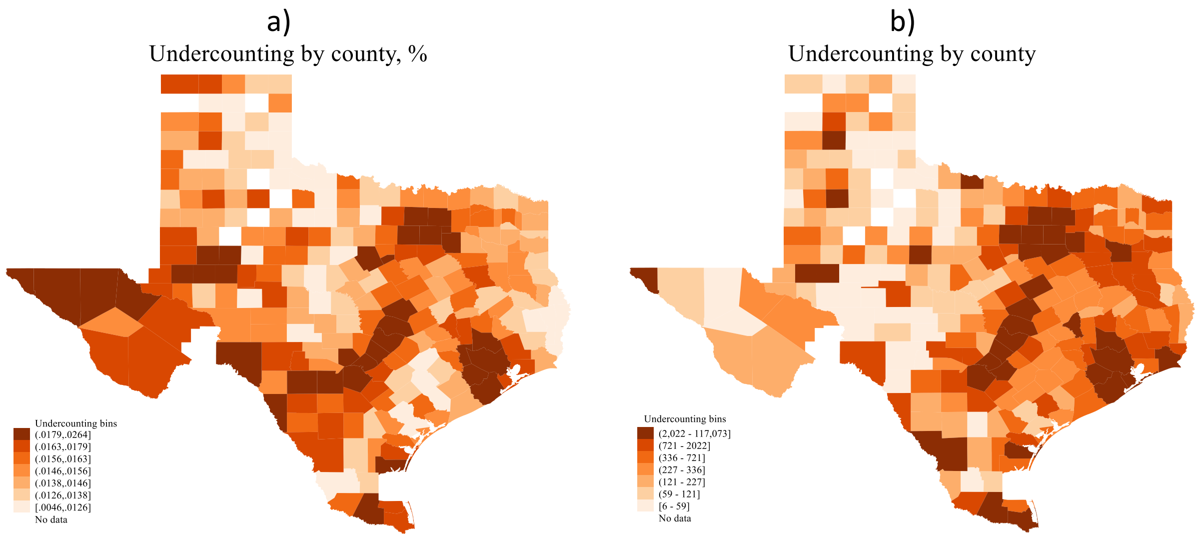

Figure 2 Geographical dispersion of the undercounting share and undercounting in Texas.

Table 7 presents the names and values of the top and bottom 20 counties in terms of undercounting. This table confirms the intense undercounting in the Austin (Travis, Williamson, Bell, and McLennan counties), Dallas (Dallas, Tarrant, Collin, and Denton), Houston (Harris, Fort Bend, Montgomery, Brazoria, and Galveston counties, and San Antonio (Bexar County) areas, and in those counties located in the U.S.-Mexico border—which also have a high undercounting in terms of population.[2] For instance, some of the counties in the U.S-Mexico border presenting darker color in both maps are El Paso, Hidalgo, Cameron, and Webb counties (where famous border cities such as El Paso, McAllen, Laredo, and Brownsville are located, respectively). In summary, from the Top-20 undercounted counties, 14 are part of large metropolitan areas, and 4 are on the U.S.-Mexico border. The other two counties, Lubbock and Webb, are located in the north and south of Texas, respectively.

[2] The entire list of counties and their undercounting measures is presented as Appendix.

The conjunct analysis of the maps in Figure 2 and lists in Table 7 allows us to observe why we should not rely only on one of the maps or on one of the extremes of the list when counties are ranked. For example, Culberson County (the third county from left to right in the maps) has remarkably different colors in the two panels, and Table 7 helps us clarify why. Culberson county has a relatively high share of undercounting (1.80%), but only 40 people were undercounted.

Table 7
Top and bottom 20 undercounted Texas counties.

County	Pop.	Und.	% Und.	County	Pop.	Und.	% Und.
a) Top 20				*b) Bottom 20*			
Harris	4,602,523	117,073	2.54%	Hardeman	3,952	42	1.06%
Dallas	2,586,552	58,165	2.25%	Hall	3,074	41	1.32%
Tarrant	2,019,977	42,047	2.08%	Culberson	2,241	40	1.80%
Bexar	1,925,865	40,404	2.10%	Cochran	2,904	39	1.34%
Travis	1,203,166	23,270	1.93%	Dickens	2,216	36	1.63%
Collin	944,350	17,791	1.88%	Collingsworth	2,996	35	1.18%
Hidalgo	849,389	16,250	1.91%	Edwards	2,055	35	1.70%
El Paso	837,654	16,132	1.93%	Jeff Davis	2,234	35	1.55%
Denton	807,047	14,963	1.85%	Oldham	2,090	32	1.51%
Fort Bend	739,342	14,827	2.01%	Glasscock	1,430	31	2.20%
Montgomery	554,445	10,706	1.93%	Menard	2,123	30	1.42%
Williamson	527,057	10,053	1.91%	Armstrong	1,916	25	1.29%
Cameron	421,750	7,445	1.77%	Stonewall	1,385	20	1.47%
Brazoria	353,999	6,414	1.81%	Briscoe	1,546	20	1.31%
Bell	342,236	6,394	1.87%	Sterling	1,141	20	1.72%
Nueces	360,486	6,390	1.77%	Cottle	1,623	19	1.18%
Galveston	327,089	5,613	1.72%	Irion	1,524	18	1.21%
Lubbock	301,454	4,961	1.65%	Throckmorton	1,567	18	1.12%
Webb	272,053	4,864	1.79%	McMullen	662	11	1.61%
McLennan	248,429	3,936	1.58%	Foard	1,408	6	0.46%

Note: The data is ranked by the undercounting (Und.) values.

5.1. County-level correlations

This subsection provides an overview of the correlation of our estimates to its original theoretical determinants and relevant socioeconomic variables. The former analysis will help us identify whether our estimates are disproportionately accounted for, regardless of their proportional weights and if some of the variables exert a significant role in explaining undercounting. The latter analysis might help us understand the social and economic features surrounding undercounting.

As a starting point, the correlation between the undercounting share and the seven original variables lies between 0.45 and 0.59. The similar correlation of the seven variables to the share of undercounting provides evidence in favor of the robustness of our approach in using variables almost equally crucial in determining undercounting. However, interpreting the correlation coefficients between the estimated undercounted people by county and the seven original variables must be taken with caution since the estimate of undercounted people is the product of counties' population and the share of undercounting. Therefore, the estimated undercounted population will automatically correlate with the counties' population share (0.99) and population density (0.90). Interestingly, the estimate of undercounted people is not related to four of the other five variables (volunteering, addresses unable to be geocoded in the county, ACS 5-year nonresponse rate by refusal, and ACS 5-year nonresponse rate by other than refusal) with correlation coefficients from -0.01 to 0.09; but slightly—if something—related to the Cohesiveness by clustering variable, with a correlation coefficient of 0.33.

The methodological approach presented is limited to the availability of reliable data at the county level for each of the seven theoretical determinants. Therefore, to further assess the relationship of the county-level undercounting to demographic characteristics and the Census implementation. Correlation coefficients are estimated for different population categories according to age, race, and the Census response method (see Table 8).

Table 8
Correlation of the share of undercounting with relevant variables in Texas.

Category	Corr. Coef.	Category	Corr. Coef.
a) Age Group		*b) Race*	
Below 5	0.1309	White	-0.1802
5-9	0.1660	Black	-0.0018
10-14	0.1770	Asian	0.3970
15-19	0.1944	Hispanic	0.2566
20-24	0.2408	Cuban	0.1618
25-34	0.3349	Mexican	0.2344
35-44	0.3070	Puerto Rican	0.2420
45-54	0.1437	Other origin	0.2619
55-59	-0.2680		
60-64	-0.2391	*c) Census Self-Response Rate*	
65-74	-0.4013		
75-84	-0.4566	Internet	0.4719
85+	-0.4543	Phone and mail	-0.5522

Note: The Age Group and Race categories are obtained from the 5-year ACS Demographic and Housing Estimates. The Phone and mail category of the Census Self-Response Rate is the result of subtracting the internet from the Overall Self-Response Rate. Each category represents the share of the total county-level population for the indicated population.

Panel a) in Table 8 shows that the share of people in groups below 54 years old is positively associated with higher undercounting, and groups above 55 years and older are negatively associated with undercounting. These findings might reflect the relevance and participation given to the Census by older groups of people. Similarly, this might arise since younger population groups are more likely to be part of the labor force and not to be available to respond to the Census or to be counted appropriately. The population groups below 54 years present an inverse-U relationship to undercounting, with its maximum correlation estimate for those 25-34 years old (0.3349). On the other hand, the negative relationship to undercounting increases as population groups get older, with its maximum in the 75-84 group (-0.4566). These results are robust to those presented by the Census Demographic Analysis in which younger groups are associated with undercounting, and older groups above are inversely related to it (Jensen & Kennel, 2022).

Regarding racial categories, the share of white people is negatively related to the counties' undercounting share, with a correlation coefficient of -0.1802. Suggesting that white people might be less likely to be undercounted. The share of the Black population in the counties is technically not related to undercounting, with a correlation estimate of -0.0018. On the other hand, the Asian and Hispanic population's shares are positively associated with a higher undercounting share, with a correlation coefficient of 0.3970 and 0.2566, respectively. When studying the Hispanic population, the share of those not from Puerto Rico, Cuba, or Mexico is associated with a higher counties' undercounting share (0.2619)—closely followed by Puerto Rico and Mexico, with correlation coefficients of 0.2420 and 0.2340. The correlation between our undercounting shares and racial groups coincides with those of the Census Post-Enumeration Survey, in which the Hispanic population is associated with higher undercounting, and the white population has the opposite relationship (Jensen & Kennel, 2022).

Our estimates can also be compared to the self-response rates of the 2020 U.S. Census. In this regard, it is important to signal that the last Census was the first one in which it was implemented via the Internet (Bates, 2017). Our county-level undercounting share has a strong and positive correlation coefficient to the share of people that self-responded via the internet (0.4719) and a strong and negative value with the share that self-responded via traditional methods, such as telephone and mail (-0.5522). While these findings unveil some plausible risks from implementing the Census online, we encourage the reader to take this with caution since a causal statement should not arise from this analysis. Instead, we encourage future research lines to study the causal mechanisms driving the undercounting and overcounting in the United States.

6. Concluding remarks

In this report, we propose a practical methodology to estimate Census undercounting at the county level and present its main results for Texas and Texas' population groups—categorized by age, race, and Census self-response method. To do so, we account for personal, geographical, and Census features dimensions to first build a theory-based model with determinants of undercounting. Then, we estimate a proportionally-weighted index to allocate counties along the 90% confidence interval of the state-level undercounting provided by the Census.

Texas' estimates suggest intense undercounting—in terms of undercounting share—occurs in the Austin (Travis, Williamson, Bell, and McLennan counties), Dallas (Dallas, Tarrant, Collin, and Denton), Houston (Harris, Fort Bend, Montgomery, Brazoria, and Galveston counties, and San Antonio (Bexar County) areas. Moreover, undercounting is observed in those counties located on the U.S.-Mexico border (El Paso, Hidalgo, Cameron, and Webb County, where El Paso, McAllen, Laredo, and Brownsville are located)—which also have a high undercounting in terms of population.

The county-level dynamics across age groups and race categories suggest this approach is robust to the overall dynamics found by the U.S. Census Demographic Analysis and Post-Enumeration Survey. Our analysis suggests that the share of the population in younger groups is associated with higher undercounting and that the share of older groups is inversely related to undercounting. We also find that the counties' share of white people is inversely associated with undercounting, and the share of the Hispanic population is associated with higher levels of undercounting. Moreover, we identified a positive relationship between the counties' Census self-response rates via the internet and our estimates of the share of undercounting, which might be the result of undercounting occurring in counties where the access to the internet is limited or just of the lack of strong participation of people via the internet. Our theory-based approach aims to be a cornerstone in the alternative estimation of undercounting and overcounting. More research is recommended to obtain a comprehensive understanding of undercounting and overcounting.

References

Bates, N. (2017). The Morris Hansen Lecture: Hard-to-Survey Populations and the U.S. Census: Making Use of Social Marketing Campaigns. *Journal of Official Statistics*, *33*(4), 873–885. https://doi.org/10.1515/jos-2017-0040

Brick, J. M., & Williams, D. (2013). Explaining Rising Nonresponse Rates in Cross-Sectional Surveys. *Annals of the American Academy of Political and Social Science*, *645*(1), 36–59. https://doi.org/10.1177/0002716212456834

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., … Wernerfelt, N. (2022a). Social Capital I: Measurement and Associations with Economic Mobility. *Nature*, *608*(7921), 108–121. https://doi.org/10.1038/s41586-022-04996-4

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., … Wernerfelt, N. (2022b). Social

Capital II: Determinants of Economic Connectedness. *Nature*, *608*(7921), 122–134. https://doi.org/10.1038/s41586-022-04997-3

Clogg, C. C., Massagli, M. P., & Eliason, S. R. (1989). Population Undercount and Social Science Research. *Social Indicators Research*, *21*(6), 559–598.

de la Puente, M. (1995). *Using ethnography to explain why people are missed or erroneously included by the Census: Evidence from small area ethnographic studies* (No. SM95-16; Census Working Papers).

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research*, *38*(1), 1–18. https://doi.org/10.1016/j.ssresearch.2008.03.007

Heyneman, S. P. (2000). From the Party/State to Multiethnic Democracy: Education and Social Cohesion in Europe and Central Asia. *Educational Evaluation and Policy Analysis*, *22*(2), 173–191.

Homans, G. C. (1958). Social Behavior as Exchange. *American Journal of Sociology*, *63*(6), 597–606.

Jensen, E. B., & Kennel, T. (2022). *Who Was Undercounted, Overcounted in the 2020 Census? Detailed Coverage Estimates for the 2020 Census Released Today* (America Counts: Stories Behind the Numbers).

King, M. L., & Magnuson, D. L. (1995). Perspectives on Historical U. S. Census Undercounts. *Social Science History*, *19*(4), 455–466.

Kissam, Ed, Quezada, C., & Intili, J. A. (2018). Community-based canvassing to improve the U.S. Census Bureau's Master Address File: California's experience in LUCA 2018. *Statistical Journal of the IAOS*, *34*(4), 605–619. https://doi.org/10.3233/SJI-180480

Kissam, Edward. (2017). Differential undercount of Mexican immigrant families in the U.S. Census. *Statistical Journal of the IAOS*, *33*(3), 797–816. https://doi.org/10.3233/SJI-170388

Letki, N. (2006). Investigating the roots of civic morality: Trust, social capital, and institutional performance. *Political Behavior*, *28*(4), 305–325. https://doi.org/10.1007/s11109-006-9013-6

Mahler, S. (1993). *Alternative Enumeration of Undocumented Salvadorans on Long Island* (No. EV93-26; Census Working Papers).

Martin, E. (2007). Strength of Attachment: Survey Coverage of People with Tenuous Ties to Residences. *Demography*, *44*(2), 427–440.

Martin, E., & de la Puente, M. (1993). *Research on sources of undercoverage within households* (No. SM93-03; Census Working Papers).

O'Hare, W. P. (2019). *Differential Undercounts in the U.S. Census*. http://link.springer.com/10.1007/978-3-030-10973-8

Olson, K., Smyth, J. D., Horwitz, R., Keeter, S., Lesser, V., Marken, S., Mathiowetz, N. A., Mccarthy, J. S., O'Brien, E., Opsomer, J. D., Steiger, D., Sterrett, D., Su, J., Suzer-Gurtekin, Z. T., Turakhia, C., & Wagner, J. (2021). Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR Task Force Report. *Journal of Survey Statistics and Methodology*, *9*(3), 381–411. https://doi.org/10.1093/jssam/smz062

Putnam, R. D. (1995). Bowling Alone: America's Declining Social Capital. *Journal of Democracy*, *6*(1), 65–78.

Sinclair, M., Otoole, J., Malawaraarachchi, M., & Leder, K. (2012). Comparison of response rates and cost-effectiveness for a community-based survey: Postal, internet and telephone modes with generic or personalised recruitment approaches. *BMC Medical Research Methodology*, *12*(1), 1. https://doi.org/10.1186/1471-2288-12-132

Tourangeau, R., & Plewes, T. J. (2013). Nonresponse in social science surveys: A research agenda. In R. Tourangeau & T. J. Plewes (Eds.), *Nonresponse in Social Science Surveys: A Research Agenda*. The National Academies Press. https://doi.org/10.17226/18293

U.S. Census Bureau. (2020). *Average Household Size and Population Density - County*.

https://covid19.census.gov/datasets/USCensus::average-household-size-and-population-density-county/about

U.S. Census Bureau. (2021a). *Local Update of Census Addresses (LUCA) Operation*. https://www.census.gov/programs-surveys/decennial-census/about/luca.html

U.S. Census Bureau. (2021b). *The Post-Enumeration Survey: Measuring Coverage Error*. https://www.census.gov/newsroom/blogs/random-samplings/2021/12/post-enumeration-measuring-coverage-error.html

U.S. Census Bureau. (2022a). *Census Bureau Today Releases 2020 Census Undercount, Overcount Rates by State*. America Counts: Stories Behind the Numbers.

U.S. Census Bureau. (2022b). *Housing unit response and nonresponse rates with reasons for noninterviews*. Explore Census Data. https://data.census.gov/cedsci/table?q=B98021&g=0400000US48,48%240500000&tid=ACSDT5Y2020.B98021

U.S. Constitution.

U.S. GAO. (2021). *2020 CENSUS: Innovations Helped with Implementation, but Bureau Can Do More to Realize Future Benefits* (No. 21–478; Issue June).

West, K. K., & Fein, D. J. (1990). Census Undercount: An Historical and Contemporary Sociological Issue. *Sociological Inquiry*, *60*(2), 127–141. https://doi.org/10.1111/j.1475-682X.1990.tb00134.x

West, K. K., & Robinson, J. G. (1999). *What Do We Know About The Undercount of Children?* (POP-WP039; Census Working Papers).

Appendix

Table A1
Texas' counties undercounting.

County	Pop.	Und.	% Und.	County	Pop.	Und.	% Und.
Harris	4,602,523	117,073	2.54%	Pecos	15,797	262	1.66%
Dallas	2,586,552	58,165	2.25%	Scurry	17,239	261	1.51%
Tarrant	2,019,977	42,047	2.08%	DeWitt	20,435	260	1.27%
Bexar	1,925,865	40,404	2.10%	Hutchinson	21,571	257	1.19%
Travis	1,203,166	23,270	1.93%	Young	18,114	252	1.39%
Collin	944,350	17,791	1.88%	Karnes	15,387	248	1.61%
Hidalgo	849,389	16,250	1.91%	Montague	19,409	247	1.27%
El Paso	837,654	16,132	1.93%	Lee	16,952	243	1.43%
Denton	807,047	14,963	1.85%	Tyler	21,496	235	1.09%
Fort Bend	739,342	14,827	2.01%	Nolan	14,966	234	1.57%
Montgomery	554,445	10,706	1.93%	Robertson	16,890	231	1.37%
Williamson	527,057	10,053	1.91%	Lavaca	19,941	231	1.16%
Cameron	421,750	7,445	1.77%	Moore	21,801	227	1.04%
Brazoria	353,999	6,414	1.81%	Trinity	14,569	227	1.56%
Bell	342,236	6,394	1.87%	Madison	14,128	205	1.45%
Nueces	360,486	6,390	1.77%	Zavala	12,131	200	1.65%
Galveston	327,089	5,613	1.72%	Comanche	13,495	196	1.45%
Lubbock	301,454	4,961	1.65%	Eastland	18,270	194	1.06%
Webb	272,053	4,864	1.79%	Dawson	12,964	190	1.47%
McLennan	248,429	3,936	1.58%	Ward	11,586	190	1.64%
Hays	204,150	3,818	1.87%	Morris	12,424	185	1.49%
Brazos	219,193	3,605	1.64%	Callahan	13,770	185	1.34%
Jefferson	255,210	3,514	1.38%	Lamb	13,262	183	1.38%
Midland	164,194	3,295	2.01%	Jackson	14,820	178	1.20%
Smith	225,015	3,205	1.42%	Blanco	11,279	178	1.58%
Ector	158,342	3,016	1.90%	Terry	12,615	175	1.39%
Ellis	168,838	2,870	1.70%	Rains	11,473	172	1.50%
Johnson	163,475	2,775	1.70%	Dimmit	10,663	172	1.61%
Guadalupe	155,137	2,696	1.74%	Somervell	8,743	168	1.92%
Comal	135,097	2,451	1.81%	Camp	12,813	167	1.30%
Randall	132,475	2,189	1.65%	Zapata	14,369	165	1.15%
Taylor	136,348	2,144	1.57%	Franklin	10,679	164	1.53%
Kaufman	118,910	2,131	1.79%	Live Oak	12,123	163	1.34%
Wichita	131,818	2,115	1.60%	Brewster	9,216	162	1.75%
Parker	129,802	2,072	1.60%	Wilbarger	12,906	161	1.25%
Grayson	128,560	1,972	1.53%	Red River	12,275	156	1.27%
Tom Green	117,466	1,960	1.67%	Parmer	9,852	155	1.57%
Potter	120,899	1,920	1.59%	Newton	14,057	152	1.08%
Gregg	123,494	1,854	1.50%	Duval	11,355	146	1.28%
Rockwall	93,642	1,758	1.88%	Marion	10,083	139	1.38%
Victoria	91,970	1,493	1.62%	Ochiltree	10,348	138	1.33%

Liberty	81,862	1,465	1.79%	Winkler	7,802	136	1.74%
Hunt	92,152	1,444	1.57%	Clay	10,387	135	1.30%
Bastrop	82,577	1,312	1.59%	Runnels	10,310	132	1.28%
Bowie	93,858	1,299	1.38%	Archer	8,789	132	1.50%
Henderson	80,460	1,294	1.61%	Sabine	10,458	129	1.23%
Coryell	75,389	1,183	1.57%	Presidio	7,123	125	1.75%
Walker	71,539	1,148	1.60%	LaSalle	7,409	121	1.63%
Angelina	87,607	1,144	1.31%	Dallam	7,243	121	1.67%
San Patricio	67,046	1,093	1.63%	Yoakum	8,571	121	1.41%
Wise	64,639	1,092	1.69%	San Augustine	8,327	118	1.42%
Maverick	57,970	1,048	1.81%	Stephens	9,372	118	1.26%
Starr	63,894	1,019	1.59%	Hamilton	8,269	111	1.34%
Nacogdoches	65,558	988	1.51%	Bailey	7,092	111	1.56%
Orange	84,047	970	1.15%	Martin	5,614	110	1.97%
Harrison	66,645	960	1.44%	Hudspeth	4,098	108	2.64%
Waller	49,987	949	1.90%	Jack	8,842	108	1.22%
Anderson	57,863	944	1.63%	McCulloch	8,098	102	1.26%
Medina	49,334	925	1.87%	Coleman	8,391	100	1.20%
Valverde	49,027	908	1.85%	Mitchell	8,558	100	1.17%
Hood	56,901	908	1.60%	Brooks	7,180	91	1.27%
Kendall	41,982	829	1.98%	Goliad	7,531	90	1.20%
Atascosa	48,828	825	1.69%	Crosby	5,861	86	1.47%
Rusk	53,595	803	1.50%	Hansford	5,547	85	1.53%
Wilson	48,198	800	1.66%	Castro	7,787	85	1.09%
Cherokee	51,903	791	1.52%	Lynn	5,808	85	1.46%
Erath	41,482	780	1.88%	Delta	5,215	83	1.60%
Van Zandt	54,368	755	1.39%	Swisher	7,484	83	1.11%
Hardin	56,379	741	1.32%	Garza	6,288	81	1.29%
Kerr	51,365	723	1.41%	San Saba	5,962	81	1.36%
Burnet	45,750	720	1.57%	Floyd	5,872	78	1.33%
Lamar	49,532	714	1.44%	Crane	4,839	76	1.58%
Chambers	40,292	693	1.72%	Refugio	7,236	75	1.04%
Navarro	48,583	680	1.40%	Childress	7,226	74	1.02%
Howard	36,667	663	1.81%	Haskell	5,809	73	1.26%
Caldwell	41,401	658	1.59%	Mills	4,902	68	1.39%
Jim Wells	41,192	635	1.54%	Wheeler	5,482	68	1.23%
Cooke	39,571	615	1.56%	Carson	6,032	65	1.07%
Polk	47,837	611	1.28%	Hemphill	4,061	62	1.54%
Wood	43,815	611	1.39%	Concho	4,233	62	1.47%
Upshur	40,769	597	1.46%	Sutton	3,865	60	1.56%
Wharton	41,551	593	1.43%	Mason	4,161	60	1.43%
Hopkins	36,240	573	1.58%	Knox	3,733	60	1.59%
Kleberg	31,425	567	1.80%	Fisher	3,883	59	1.52%
Hill	35,399	531	1.50%	Real	3,389	56	1.66%
Matagorda	36,743	529	1.44%	Reagan	3,752	56	1.49%
Washington	34,796	520	1.49%	Kinney	3,675	55	1.49%
Titus	32,730	515	1.57%	Kimble	4,408	55	1.24%
Bee	32,691	510	1.56%	Coke	3,275	53	1.63%

Fannin	34,175	509	1.49%	Crockett	3,633	53	1.47%
Uvalde	27,009	497	1.84%	Shackelford	3,311	52	1.58%
Brown	37,834	496	1.31%	Jim Hogg	5,282	52	0.99%
Grimes	27,630	489	1.77%	Sherman	3,058	52	1.70%
Hale	34,113	485	1.42%	Upton	3,634	50	1.37%
Austin	29,565	465	1.57%	Schleicher	3,061	47	1.55%
Cass	30,087	417	1.39%	Lipscomb	3,469	47	1.35%
San Jacinto	27,819	417	1.50%	Baylor	3,591	45	1.25%
Palo Pinto	28,317	409	1.45%	Donley	3,387	42	1.25%
Jasper	35,504	403	1.14%	Hardeman	3,952	42	1.06%
Milam	24,664	399	1.62%	Hall	3,074	41	1.32%
Willacy	21,754	388	1.78%	Culberson	2,241	40	1.80%
Aransas	24,763	387	1.56%	Cochran	2,904	39	1.34%
Gillespie	26,208	379	1.45%	Dickens	2,216	36	1.63%
Gaines	20,321	362	1.78%	Collingsworth	2,996	35	1.18%
Hockley	23,162	340	1.47%	Edwards	2,055	35	1.70%
Calhoun	21,807	332	1.52%	Jeff Davis	2,234	35	1.55%
Shelby	25,478	332	1.30%	Oldham	2,090	32	1.51%
Bandera	21,763	330	1.52%	Glasscock	1,430	31	2.20%
Jones	19,891	329	1.65%	Menard	2,123	30	1.42%
Limestone	23,515	329	1.40%	Armstrong	1,916	25	1.29%
Andrews	17,818	309	1.74%	Stonewall	1,385	20	1.47%
Gray	22,685	305	1.34%	Briscoe	1,546	20	1.31%
Frio	19,394	304	1.57%	Sterling	1,141	20	1.72%
Houston	22,955	303	1.32%	Cottle	1,623	19	1.18%
Panola	23,440	300	1.28%	Irion	1,524	18	1.21%
Fayette	25,066	297	1.18%	Throckmorton	1,567	18	1.12%
Freestone	19,709	290	1.47%	McMullen	662	11	1.61%
Lampasas	20,640	290	1.40%	Foard	1,408	6	0.46%
Bosque	18,122	285	1.57%	Borden	665		
Llano	20,640	284	1.38%	Hartley	5,767		
Reeves	15,125	283	1.87%	Kenedy	595		
Gonzales	20,667	278	1.35%	Kent	749		
Leon	17,098	277	1.62%	King	228		
Burleson	17,863	277	1.55%	Loving	102		
Colorado	21,022	275	1.31%	Motley	1,156		
Deaf Smith	18,899	269	1.43%	Roberts	885		
Falls	17,299	269	1.56%	Terrell	862		

Note: Nine counties with no information Cohesiveness by clustering and Volunteering are excluded from the estimation procedures: Borden, Hartley, Kenedy, Kent, King, Loving, Motley, Roberts, and Terrel.