At the Juncture of Mathematics, Statistics, Social Sciences, and Society: The Census Bureau Use Case

Director Robert L. Santos' Lecture at the Joint Mathematics Meeting

January 9, 2023, Boston, MA

- Good morning, everyone.
- It's great to be here with fellow statisticians and mathematicians in this wonderful city, even in the cold of winter.
- I thought I'd talk about something a little different from the usual fanfare at these meetings.
- It's a topic that we don't discuss maybe as much as we should.
- My topic is about how mathematical and social sciences come together and interact with society in the formation of statistical data solutions.
- Solutions to data and statistical problems aren't always a matter of optimizing squared error loss in a mathematical model.
- Social and even ethical factors need to be taken into account.
- Social factors can include different groups of stakeholders with competing needs based on their own use cases.
- Or laws that require confidentiality and protection of privacy.
- Or public trust that can affect the extent to which quality data can be obtained through solicitation, as we do in censuses and surveys.
- It's really quite fascinating when you dive into it.
- Specifically, I'll describe this in the context of the work we do at the U.S. Census Bureau.
- Although, the reality is that this topic is relevant to just about any operation responsible for generating data on people or businesses.
- But before beginning, let me note that I've just completed my first year as director of the Census Bureau.
- Thus far, as a statistician and avowed nonpolitician, the experience has been fascinating.
- Hey, just over a year ago I was a VP at a public policy research think tank. I was finishing up my term as president of the American Statistical Association.
- I was even gearing up to manage another year for the SXSW Photocrew side gig. And then, suddenly, I was confirmed by the senate.
- We packed up and moved from Austin, Texas, to DC, and I was sworn in on January 5 of last year.
- Many people might consider the director role to be particularly daunting at this point in our nation's history.
- I saw genuine opportunity to help the Census Bureau. In fact, my optimism grew after I arrived.
- I learned how the Census Bureau nimbly adapted its massive decennial census operation in the midst of challenges that included a global pandemic shutting down the nation just before Census Day—April 1, 2020.

- Subsequently, this 2020 experience helped motivate the Census Bureau to develop a robust transformation and modernization initiative.
- The initiative redefines how the Census Bureau collects, processes, and disseminates our statistical data products.
- I saw an opportunity to bolster that transformation.
- How? By diversifying how we innovate our systems, operations, policies, and even mathematical solutions to problems.
- That includes soliciting wide ranging feedback from both internal career staff and external stakeholders, partners, data users, and the public.
- So, I spent this last year laying the groundwork to create continuous, ongoing community engagement in its broadest sense.
- I and Census Bureau staff met with numerous scientific and government associations.
- We've conducted listening sessions with stakeholders representing both communities and data users.
- Internally, I met with career staff at all levels and in offices around the country.
- I went out on multiple field observations.
- I conducted media interviews and used blogs, videos, and other communications to reach out to stakeholders and the public.
- I met with tribal leaders from around the country.
- Throughout these engagements, we listened carefully and strengthened our ties.
- The experience was profound.
- I engaged with rural America and saw the struggles and joys of the lives of farmers and smalltown businesses.
- I visited inner-city neighborhoods and spoke to local community leaders and pastors who recognized the value of local statistical data.
- I witnessed the dignity of America's indigenous people and their honorable, indeed sacred way of life through living with nature.
- And I met with scholars and researchers from across the nation to understand their concerns.
- I learned that it takes a community-of-the-whole to maintain a fully functioning and successful federal statistical agency.
- We need to seek and act on feedback from stakeholders and partners external to the Census Bureau.
- This is highly related to my talk today.
- As I begin my second year in office, I believe that we at the Census Bureau are in a good place for continuing our transformation into a 21st century statistical agency.
- I'm optimistic and hopeful.
- OK, thanks for indulging me while I gave my reflections. I really wanted to share them with you.
- Now, let's switch to the topic of the day.
- Over our 120-year history, the Census Bureau has endured two global pandemics and two world wars.
- We've lived through the major societal impacts of an industrial revolution and our more recent technological renaissance.
- Most folks know that our principal function is to fulfil a constitutionally mandated decennial population census.

- But our mission is actually much broader. We provide authoritative, quality information on the nation's economy and its people.
- And we do so in vast quantities. In fact, we conduct three censuses—the other two being a census of governments and an economic census of businesses.
- Plus, we conduct over 130 surveys of people and businesses. They cover a wide range of topics.
- It's through these data collections where a confluence occurs that brings together mathematical, statistical, and social sciences, along with public perceptions and influence.
- Our solutions at some level must be vetted, understood, and accepted by society, by the public.
- And historically, the public has something to say about our products and our methods.
- We are obliged by our core values to take that into account.
- And after all, the Census Bureau fundamentally serves the public. I do, too!
- Now, our data and statistical products help us in myriad ways.
- They're used in governance such as advancing democracy via apportionment.
- They're used in evidence-building, policy-making, and program implementation.
- Commercially, they're used in product development, marketing, and economic development.
- They're used in public health and emergency planning. And, of course, they're used for primary research that seeks to understand who we are as a people, an economy and a nation.
- And I'm just scratching the surface . . . The scope of all these uses is vast.
- Let's now consider our data users in the broadest sense . . . to include everything from federal, state, and local government to community organizations, researchers, and so forth.
- These users not only inform what data are gathered, but they influence what methods are used.
- For instance, questions on the American Community Survey or ACS appear only if they're required by statute or judicial ruling.
- Most folks don't realize that.
- And most don't realize that data users influence the data we collect and release, subject to Title 13 confidentiality and OMB regulations.
- We listen to data users as part of our commitment to quality and scientific integrity.
- In fact, an essential part of data quality is the relevance of the data being collected.
- So, if the data aren't sufficiently useful to users, we'll definitely hear about it.
- Our commitments to quality, scientific integrity, and transparency are essential to building and maintaining public trust.
- And public trust is foundational to the success and utility of every Census Bureau data product.
- Without cooperation from public and commercial sectors, the data we collect could be subject to serious biases.
- It's funny how this is all connected, right?
- You see, our core values are connected to public trust, which is connected to participation in data collection, and that in turn is connected to data quality which affects perceptions of scientific integrity. We've come full circle.
- This is where mathematics, statistics, and the social sciences intersect with society.
- Societal changes should motivate us to assess the relevance of data elements that used to be important, but may no longer be.
- Or reexamine methods that have proven tried and true over time, but whose effectiveness is diminishing.

- Societal changes can alter people's perceptions of government and privacy, which impacts public trust and participation in data collection.
- So here's an example of a question on the ACS that was no longer relevant and where public sentiment played a role in its removal.
- For decades we collected data on whether a house had a flushing toilet.
- Well, public health and housing policy could be developed with these data.
- In fact, in the early 1950s, roughly a quarter of homes had no flush toilet.
- But by 2014, well over 99 percent of all homes were flushable, so to speak.
- But increasingly during the 2000s, the public expressed concerns about the intrusiveness of this "flushing" question.
- Some in Congress even chimed in. So, after some analysis and research, the Census Bureau removed the question in 2015.
- Next, here's an example of adding to ACS content. Consider the world wide web.
- The internet was publicly available in the 1990s and became popular in the early 2000s.
- Passage of the 2008 Broadband Improvement Act mandated the addition of computer and internet use questions to the ACS.
- These data are now instrumental in the implementation of the Infrastructure Investment and Jobs Act to make broadband access ubiquitous in the nation.
- This is an instance where technological advances led to societal changes in behavior which in turn sparked the need for data collection.
- And today, OMB is in the midst of revising its 1997 standards for collecting race and ethnicity.
- This is in response to an increasingly diverse nation and the recognized need to capture more granularity in race and ethnicity, as well as their mixtures.
- More broadly, because of our nation's social and economic evolution, there's an ongoing need to review the items we measure and how we measure them.
- Things we measured before may no longer be relevant.
- Things we measure now may require different methods or even data sources to measure accurately.
- We also need to adjust for shifting societal attitudes on privacy and trust in government overall.
- But there's another aspect that I mentioned earlier that I want to dive into:
  - Solutions to data and statistical problems aren't always a matter of mathematical optimization. Social and even ethical factors need to be taken into account.
- I'll illustrate this more directly with a couple of projects we do at the Census Bureau.
- I'll start with our Population Estimates Program.
- Under this program, population estimates are produced annually for the nation, states, metro areas, counties, and other geographies.
- The program is incredibly important.
- Over the course of a decade, these projections directly or indirectly influence the allocation of trillions of dollars of federal funds to states and communities.
- They're also used to calibrate our population surveys, including the American Community Survey.
- So, survey estimates—like the number of children in poverty—rely critically on these projections.
- Between 2010 and 2020, we used a straightforward approach to produce our annual population projections; it was called the cohort-component method.
- At a high level, the calculations started with the most recent decennial census count.

- That marked the known population at the beginning of a decade.
- Then, we added annual births, subtracted deaths, and added net migration each year to generate a set of projections for that vintage year. It's pretty simple, right?
- Well, the method worked well this last decade.
- But when it was time to do our 2021 projections, we faced a logistical problem: the 2020 Census data were not yet available.
- The pandemic had delayed the release of the starting point—the base population projection at the beginning of the decade—to which births, deaths, and migration occurring in 2021 would then be added.
- Well, necessity is the mother of invention, so our wonderful demographers and statisticians got together to create an alternative approach.
- We formed a new 2020 population base by combining three sources of data:
  - First, the state and county totals from the already-released 2020 Census redistricting file.
  - Second, the Vintage 2020 annual population estimates which used the 2010 decennial census counts as a starting point.
  - And third, our 2020 Demographic Analysis that was developed from administrative data such as vital records, international migration, and Medicare records.
- This method was called a blended-base approach. I won't describe how the data were combined for the sake of brevity.
- The point is that three different sources of data were brought together to form the 2020 base population.
- And that's what we used to generate our Vintage 2021 population estimates.
- But, listen, the story doesn't end there. In fact, it's only the beginning.
- To understand the significance of this new approach, let me step back and mention the some of the work we did to assess the quality of the 2020 census.
- Now, it's usual practice for us to examine the accuracy of our decennial censuses.
- We mainly do it to inform planning of the next decennial census. There are two specific studies I want to mention.
- The first is a demographic analysis that used only administrative data to calculate the total national population.
- It provides national population estimates by age, sex, race, and Hispanic origin.
- The second study is the Post Enumeration Survey or PES.
- This survey was conducted soon after the end of decennial collection as a comparison group to actual census returns.
- It was designed to inform who we missed, who we counted more than once, and who we counted correctly.
- Now, as I've said on many occasions, no census or survey is ever perfect.
- Historical error patterns persist, and the Demographic Analysis and PES provide two specific findings that are relevant, here.
- The first is that the Demographic Analysis showed that the 2020 census undercounted children under the age of five by about 5 percent.
- The second finding is from the PES.
- It showed that certain demographic subpopulations were again undercounted, such as Hispanics at 5 percent and Blacks at just over 3 percent.
- So, why are these important to our annual population projections?

- Well, remember that the 2021 population projections start with the 2020 census population counts.
- Once the blended-base 2020 population count was generated, we noticed that the undercount of children detected in the 2020 census was somewhat mitigated.
- Our new blended-base approach had leveraged unique data sources to achieve a higher level of accuracy.
- And indeed, the Demographic Analysis was one of the three blended sources used to create the 2020 base population counts.
- Seeing this result sparked some innovative, creative thinking. The lightbulbs flashed.
- Are there ways to use other data sources to generate an even better base population count that mitigates racial and ethnic miscounts detected in the PES, for instance?
- The Base Evaluation and Research Team also known as BERT was formally established and is currently exploring the possibilities.
- The team is examining the feasibility of more broadly using coverage measures from the PES and Demographic Analysis as well as administrative records to inform improvements in the 2020 base population counts.
- That's why the blended base solution for our Vintage 2021 population estimates represent a beginning, not an end.
- This example illustrates nicely the juncture between mathematics, statistics, and social sciences with society.
- Stakeholders; communities; and federal, state, and local government voiced concerns regarding the 2020 census impact on federal funding in light of the PES and Demographic Analysis findings.
- Even though both the 2010 and 2020 censuses provided accurate counts of the total population, both detected inaccuracies of specific subpopulation counts, as I discussed earlier.
- The use of the blended base approach begins a journey on how we might better address these concerns.
- It illustrates how the juncture of statistics, social science and society can influence solutions to really challenging statistical problems in a federal statistical agency.
- And the solutions aren't always mathematical or even methodological. I'll discuss this in the next example.
- Let's turn to the use of differential privacy in some of our statistical data products.
- As many in the audience know, the Census Bureau is bound by Title 13 of US Code that governs the protection of confidential information.
- For our 2020 Census data products we've adopted a new disclosure avoidance methodology involving a solution called differential privacy.
- It's fair to say this approach doesn't enjoy a consensus among our stakeholders.
- Yet, it's a methodology that most would agree is much more scientifically rigorous than the methods used in the last decennial.
- It addresses our 21st century confidentiality threats which over time will only grow in sophistication.
- Differential privacy represents a formal method that quantifies the risk of disclosure, unlike our other methods.
- However, it does come at a cost of diminishing some of the utility that data users have enjoyed.
- Differential privacy is also very complex and computationally demanding.

- It operates by adding noise to census tabulation cells, and it does so in a measured and a transparent way by adding uncertainty at a fixed, controllable level.
- It can meet the dual objectives of disclosure avoidance and data utility.
- However, no usable, quantifiable measure of data utility currently exists.
- Utility is currently determined by offering data users the opportunity to provide feedback.
- It's done through an iterative approach involving prototype data sets released to the public for review and feedback.
- That allows a calibration of sorts to reach the best balance between utility and disclosure risk.
- And perhaps that's why controversy remains.
- Stakeholders are anxious that too much information is being compromised by way of noise infusion, especially at lower levels of geography.
- There is also the issue of the many uses of the data.
- What might be acceptable for one set of data users—say public health policy makers at the national level—can be unacceptable to others, say community-based organizations or urban planners or tribal communities.
- Yet we have a mathematical solution that is quite elegant.
- It effectively addresses disclosure threats from society.
- Unfortunately, there's no practical, objective measure of data utility, perhaps because uses of census data are limited only by one's imagination.
- As a consequence, data utility is being empirically assessed through a collaborative review process that engages the data user community multiple times.
- Our iterative user-engagement process did guide us to a solution.
- And detailed 2020 census data will be published later this year.
- Indeed, this has been a learning process. And we are still learning.
- So, at the end of the day, our use of differential privacy of the 2020 Census offers a very interesting illustration of the juncture between mathematics, statistics, social sciences, and society.
- All were needed to reach a solution.
- Well, those were the stories I chose for you this morning.
- The stories have no ending, because most of the work we do involves the ongoing, continuous intersection of math, statistics, the social sciences, and society.
- New chapters will undoubtedly unfold.
- I can sum up my conclusion with these few remarks:
    - Sometimes a mathematically optimal solution is not workable for reasons having nothing to do with the theory.
    - You may find the best solution, but if it is not socialized or communicated well with stakeholders and the public, the solution may not be acceptable.
    - In the leadership world, I've always said that you can have a correct, spot-on solution to a problem, but at the end of the day it's all about the implementation process.
- As a policy researcher I've seen time and again when an exceptional social program failed because it was implemented without taking human nature fully into account.
- And then sometimes it can all come together.
- Social science techniques such as community based participatory research or just community engagement can be used form a methodological solution that works for a broad, diverse community.

- We at the Census Bureau increasingly understand these nuances and are actively planning and implementing solutions with both internal and external engagement in mind.
- The effectiveness of our approach is bolstered through the adoption of the principles of diversity, equity, and inclusion.
- These principles can be a catalyst for accelerating the development, implementation, and acceptance of solutions.
- And they can ensure that stakeholders and the public more generally have opportunities for dialogue and input to final solutions.
- Thank you so much for this opportunity to address you today. It was an honor.