Harvard Data Science Review • Issue 2.2, Spring 2020

Error, Uncertainty, and the Shifting Ground of Census Data

Dan Bouk¹

¹Department of History, Colgate University, Hamilton, New York, United States of America

Published on: May 26, 2020

DOI: <https://doi.org/10.1162/99608f92.962cb309>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editors’ Note: *The U.S. Census has long been treated as a source of ‘ground truth,’ enabling analysis of population dynamics, income distributions, and immigration patterns. For this Mining the Past article, historian Dan Bouk analyses the struggles to create reliable census data by showing how paying attention to efforts to address ‘uncertainty’ and ‘error’ over time enrich our understanding of data-in-the-making.*

Keywords: census, data, nonsampling error, uncertainty, history

Uncertainty lurks, to some degree, in every data set. Statisticians have developed a whole host of measures of variability, from the standard deviation, to hypothesis testing’s Type I and Type II errors, to measures of sampling error. But the problem of uncertainty is much bigger than error. It’s the problem that arises when we attempt to fit a complicated, contested, contingent universe into the clean columns of a spreadsheet. It is the sort of problem that every data user should consider, one that can be studied and managed, if never fully solved.

The profound uncertainty silently roiling the world’s databases broke into the headlines earlier this year. On February 14, 2020, the *Washington Post* published [an expose](#) revealing that Google Maps presented users in some countries with one map and one set of national borders, while users in other countries saw a different set of borders. Users in India, for example, learned from Google that the long-disputed region of Kashmir was in fact fully and clearly part of India. Americans or Pakistanis viewing that same region saw, in contrast, a dotted line indicating Pakistan’s claims to the region and the contested status of the border. Asked about these differences, the director of product management at Google Maps answered, “Our goal is always to provide the most comprehensive and accurate map possible based on ground truth” (Bensinger, 2020). But this was a dodge. The problem with Google’s solution was not one of the truth of its measurements, but rather a problem of communicating inescapable (in this case, political) uncertainties.

The truth is that the ground is always shifting beneath every data set; instability is one of the few constants. Over the last hundred years, statisticians have developed a variety of methods and techniques to measure and express what they do not know. Yet some deep uncertainties have proven persistent, stubbornly resistant to what historian Theodora Dryer calls statisticians’ “uncertainty work” (Dryer, 2019). Few institutions have worked as diligently over the last century as the U.S. Census Bureau to tame uncertainty, and with great success. Yet some uncertainty refuses to be contained. Unstable borders haunted census statisticians in 1920, just as they haunt Google data scientists today. The story of the Census Bureau in the early and mid-twentieth century demonstrates both the power of methods for describing error and also their limits.

So far as I can tell, while Google allows its users to [report errors](#) in its maps, it does not communicate any measures of error to the public. The U.S. decennial census presents a stark (and praiseworthy) contrast. After each count, the Census Bureau releases estimated measures of ‘net undercount’ (the calculated difference between the actual population of the U.S. and the number counted); of the undercounts for five major racial

and ethnic groups; of the number of excess enumerations due to duplication; of the total number of omissions (or persons entirely missed in the count); and of the rate of ‘imputation,’ or the number of persons whose characteristics had to be guessed at by an algorithm (U.S. Census Bureau, 2012). These error-reporting systems have grown out of the Bureau’s longstanding commitment to methodological transparency and the influence of twentieth-century political forces.

Census Bureau officials began dreaming of more robust, quantitative measures of error during the 1940 census, when a new generation of Ph.D.-bearing social scientists took the reins of census operations. One of those Ph.D.s, Calvert Dedrick, advocated for a study of census completeness in March 1940 on the eve of the decennial count. The minutes of the Census Advisory Committee summarize his proposal this way: “We will have to make every effort in the field to insure 100 percent completeness of the schedules. There should be a field validation as to completeness following the census, he continued, going particularly into certain groups: the very young children, infants, and possibly certain migrant or transient groups” (Census Advisory Committee, 1940). That didn’t happen. Instead, the chance coincidence of a selective service enumeration of young men in 1940 shortly after the census count allowed for a later statistical study of underenumeration, one that revealed 3% of the total population to have been missed while an appalling 15% of African Americans had not been counted (Price, 1947). In subsequent years, projections based on rates of birth, death, and immigration joined post-enumeration sample surveys to form a systematic error-measurement system for the decennial census. Civil Rights movements gave political heft to those undercount surveys in the 1960s and have since inspired controversial, legally contested methods to improve or correct census coverage (Anderson and Fienberg, 1999).¹ At the same time, scholars have pointed out that the racial and ethnic categories used in census counts and error reporting also have messy histories too easily hidden behind the straight and sure lines of a data table. In fact, the census does not just report figures by race, but has throughout its history *produced* race by creating, defining, and giving substance to racial categories and ideas of racial difference (Anderson, 2015; Nobles, 2000; Prewitt, 2013; Schor, 2017; Thompson 2016).

While the Census Bureau models some exemplary error-reporting practices, it necessarily struggles with the problem of harder-to-quantify uncertainties. The economist Charles F. Manski recently criticized official statistics—from GDP to unemployment rates to Census Bureau income figures—for their failure to quantify “nonsampling error,” a hodge-podge category that includes missing data, data subject to revision over time, and categorical uncertainties (Manski, 2015). Back in 1940, the minutes of the Census Advisory Committee meeting explain that along with calling for a coverage study, Calvert Dedrick wanted an “other type of study, which he felt was more important [...] on the accuracy and the meaning of our ‘unknowns’ or refusals, and on the answers themselves”(Census Advisory Committee, 1940). While that study did not happen, Census Bureau statisticians later led the way in developing what has come to be known as the “Total Survey Error” framework, a much more expansive method for measuring or estimating nonsampling error in surveys (Groves & Lyberg, 2010).

The problem of uncertainties extending beyond ‘error’ shines out in any deep dive into the murky waters of original census data—the handwritten reports collected by an army of census takers who were once employed to enumerate every American.² The census category ‘birthplace’ illustrates the data’s limitations very well and provides an important context for thinking about the difficult task of judging data quality. When first gathered, answers to birthplace questions played a significant role in deciding who could and who could not immigrate to the United States during a period of strict immigration restriction. Today, they still shape what people know about their families and genealogies and how they make sense of their identities. The story of birthplace data is one of unstable borders, of the trouble of presenting both ground truth and foundational uncertainties, of the sorts of problems that got Google in trouble earlier this year.

The published figures from the 1920 census present birthplace as an established fact, as physical ground truth. We can learn, for instance, that exactly 124,727 people were born in France and 1,915,864 were born in Germany. There are no error bars, no hints of doubt about these almost comically precise figures. Published tables indicated that another 5,344,128 native-born white Americans had at least one parent born in Germany while 208,951 had at least one parent who had been born in France (U.S. Census Bureau, 1922, pg. II:897). A table limited only to foreign-born individuals also lists a sub-category beneath France, which indicates 34,321 Americans born in Alsace-Lorraine, a region on France’s eastern border (U.S. Census Bureau, 1922, pg. II:693).

It took a heroic statistical effort to get these figures when so much of the world was shaken by war and fractured by nationalisms. Between the years 1914 and 1919, the ground literally quaked as battling armies scarred the surface with trenches, bombs, and mortar. Blood flowed across borders while the borders themselves moved over highly contested ground. World War tends to have such effects: destabilizing or annihilating some boundaries, while hardening or extending others. The line that separated France from Germany shifted during the war, until Alsace-Lorraine fell back under French control for the first time since the end of the Franco-Prussian War in 1871.

Across the Atlantic Ocean, in Washington D.C., the census’s chief statistician convened the Census Advisory Committee to work out the knotty enumeration problems that surfaced when borders shifted. They sat down together to figure out what to do about the fact that Europe’s borders were moving. Joseph Hill, the statistician, set out the problem before them: what was the proper nation to enter for a person born in Alsace-Lorraine? When the enumerated individual was born, their birthplace was probably part of Germany; by the time the enumerator came around in 1920 that same birthplace was undoubtedly part of France. What was the ‘right’ answer?

Hill asked for guidance and together they decided on a rule. Those born in Alsace-Lorraine would all be French. It didn’t matter if an individuals’ entire life played out under a Kaiser’s reign. It didn’t matter where each person’s national allegiances or identity lay. Officially, each was French. The enumerator instructions for 1920 didn’t state this rule outright. Instead, it said that for certain nations, enumerators should report a region

or city instead of a nation: “If a person says he was born in Austria-Hungary, Germany, Russia, or Turkey as they were before the war, enter the name of the Province (State or Region) in which born, as *Alsace-Lorraine*, *Bohemia*, *Bavaria*, *German or Russian Poland*, *Croatia*, *Galicia*, *Finland*, *Slovakland*, etc.; or the name of the city or town in which born, as *Berlin*, *Prague*, *Vienna*, etc.” (U.S. Census Bureau, 1922, pg. II:1383). After the count, the Census Bureau could then use this finer grained detail to assign people to their proper places. This plan for limiting uncertainty in a later processing phase depended on first collecting more and different kinds of data.

But Hill knew the nature of such statistical problems too well to anticipate that this solution would actually *work*. Before they even made a decision, he told the assembled economists and statisticians in the advisory committee: “any rule we may adopt on this question is likely to be more or less uncertain of application—more or less of a dead letter.” He continued: “If a man or his wife or his landlady or whoever it may be to whom the question is addressed, says that he was born in Germany, for instance, it is a safe gamble that in 99 cases out of a hundred the enumerator will write down the answer without any further inquiry.” There was bound to be “a large element of uncertainty as to the extent to which the rule is observed or applied,” and yet, he said, “it is necessary to have a rule,” for the sake of the enumerator and in case anyone should ask (Hill, 1919).

The Census Bureau instructed enumerators not to write “Germany” for a person’s birthplace. Yet they did anyway: over 1.4 million times, against only half a million German-Americans who were identified by some other more specific birthplace (Ruggles, et. al., 2017). Hill had been right—the rule was pretty much a dead letter.

On a million doorsteps, when the data were collected, individuals made their own assertions about who they were, enumerators made judgments of their own, and communities exerted norms that shaped how people were reported and classified. Ambiguities abounded in the assembled data—ambiguities that are [catnip to the historian](#), offering a million tiny glances into the ways enumerators and those they counted negotiated an identity that could fit in the Bureau’s blank forms.

Back in Washington, D.C., those ambiguities presented a problem for statisticians and other data users. Joseph Hill and his colleagues had to face the fact that some of the 5.3 million Americans that their census listed as being of German ‘stock’ were probably actually, according to the new census rules, French. They had to accept too that some of those 5.3 million Americans would, regardless, understand themselves to come from Germany, or that in some cases the enumerator would consider them to be from Germany, and in many such cases, the Census Bureau would never know there was any confusion.

Still, Hill and his colleagues printed precise figures for birthplace, as if those figures weren’t run through with epistemological faults and fissures. Soon after, in 1924, Congress passed a sweeping immigration restriction bill built around census data on people’s ‘national origins,’ on the places they or their ancestors had been born. As a result, the statisticians’ excision of uncertainty took on enormous political and practical significance.

Future immigration quotas limiting the number of migrants to the U.S. would be set in proportion to the population of 1920s Americans born in each country or (for native-born Americans) in proportion to the foreign nation from which those people were descended. The 1920 census data played a significant role (and so did Joseph Hill) in making those consequential—and controversial—calculations (U.S. House, 1927, p. 3; Ngai, 1999, p. 67). The process lent the sheen of numerical objectivity to a system engineered to exclude Asians and Africans, while slowing the migration of southern and eastern Europeans nearly to a halt (Ngai, 1999). It was not the first time census data had been employed to justify the marginalization of some groups—indeed the questions about birthplace had served for decades to rationalize nativist alarm—and it would not be the last, as has been revealed by the recent (unsuccessful) effort to add to the 2020 census a citizenship question that could have discouraged migrant communities from participating in the count or could supply data to support the exclusion of noncitizens in the processes that allocate political power and federal funds.

Joseph Hill talked in passing about ‘uncertainty’ in 1920. The theorization of uncertainty followed soon after in two different fields: first in finance, and then in physics. In 1921, the economist Frank H. Knight defined risk as “*measurable* uncertainty” and true uncertainty as everything else—investigating true uncertainty, he claimed, would allow profit to be finally understood (Knight, 1921, p. 20). In 1927, Werner Heisenberg published his now-famous paper on his uncertainty principle, which worked out the mathematical relationship limiting the precision with which an observer could simultaneously measure the position and momentum of an electron (Barad, 2007, p. 116). Uncertainty in data science can find inspiration in either field. Finance invites us to frame error, like risk, as a measurable form of uncertainty, while acknowledging that much of the interpretation of a data set surely lies in wrestling with that which goes unmeasured. Physics invites us to think in new ways about how the observations that create a data set inevitably collapse a much messier reality, one of overlapping, fluid, contested borders, or identities, or what-have-you.

So, today, even if Google Maps more readily reported error measures, such reporting would still fall far short of an accurate characterization of its data. A Google official’s actual appeal to ‘ground truth’ missed the mark even farther. The term ‘ground truth’ seems to have been borrowed from the armed forces where, according to sociologist Phaedra Daipha, it referred to “data collected on the ground, where the action is” (Daipha, 2015, p. 95). From the military, the term made its way into meteorology where it described the observations and measurements that forecasters rely on to describe and predict complex weather conditions (Daipha, 2015). In recent years, machine learning practitioners have picked up on ‘ground truth’ as a way of describing the observational data against which models and predictions will be judged. When this journal [convened a symposium](#) of data scientists to discuss the census last October, participants sometimes used the term to refer to census data, which was also shown to be the baseline for many political, legal, medical, and social scientific investigations. ‘Ground truth’ must not be mistaken for truth, though, and every data scientist must keep a steady eye on error and, beyond error, uncertainty.

Disclosure Statement

Dan Bouk has no financial or non-financial disclosures to share for this article.

References

- Anderson, M. (2015). *The American census: A social history*. New Haven: Yale University Press.
- Anderson, M. & Fienberg, S. (1999). *Who counts?: The politics of census-taking in contemporary America*. New York: Russell Sage Foundation.
- Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Durham, NC: Duke University Press.
- Bensinger, G. (2020, February 14). Google redraws the borders on maps depending on who's looking. *The Washington Post*. <https://www.washingtonpost.com/technology/2020/02/14/google-maps-political-borders/>
- Census Advisory Committee. (1940). *Minutes of census advisory committee, 29-30 March 1940*. Entry 148 Minutes of Meetings, Correspondence, & Reports, Gen. Records, 1919-1949, Record Group 29 (Box 76, Folder Advisory Committee Meeting March 29 and 30, 1940). National Archives, Washington, D.C.
- Daipha, P. (2015). *Masters of uncertainty: Weather forecasting and the quest for ground truth*. Chicago: University of Chicago Press.
- Dryer, T. (2019). *Designing certainty: The rise of algorithmic computing in an age of anxiety 1920–1970* (Unpublished doctoral dissertation). University of California, San Diego.
- Groves, R. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879. <https://doi.org/10.1093/poq/nfq065>
- Hacker, J. (2013). New estimates of census coverage in the United States, 1850–1930. *Social Science History*, 37(1), 71–101. <https://doi.org/10.1215/01455532-1958172>
- Hill, J. (1919, February 5). *Memorandum for the director*. Entry 148 Minutes of Meetings, Correspondence, & Reports, Gen. Records, 1919–1949, Record Group 29 (Box 71, Folder Advisory Committee 1919-20). National Archives, Washington, D.C.
- Knight, F. (1921) *Risk, uncertainty, and profit*. New York: Houghton Mifflin Company. <https://hdl.handle.net/2027/uc1.b3276574>
- Manski, C. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. *Journal of Economic Literature* 53(3), 631-653. <https://doi.org/10.1257/JEL.53.3.631>

- Nobles, M. (2000). *Shades of citizenship: Race and the census in modern politics*. Stanford, CA: Stanford University Press.
- Ngai, M. (1999). The architecture of race in American immigration law: A reexamination of the Immigration Act of 1924. *Journal of American History*, 86(1), 67-92. <https://doi.org/10.2307/2567407>
- Prewitt, K. (2013). *What is your race?: The census and our flawed efforts to classify Americans*. Princeton, NJ: Princeton University Press.
- Price, D. (1947). A check on underenumeration in the 1940 census. *American Sociological Review*, 12(1), 44–49. <https://doi.org/10.2307/2086489>
- Ruggles, S., Genadek, K., Goeken, R., Grover, J. & Sobek, M. (2017) *Integrated public use microdata series: Version 7.0*. Minneapolis, MN: University of Minnesota. <https://doi.org/10.18128/D010.V7.0>
- Schor, P. (2017). *Counting Americans: How the US census classified the nation*. Translated by Lys Ann Weiss. New York: Oxford University Press.
- Steckel, R. (1991). The quality of census data for historical inquiry: A research agenda. *Social Science History*, 15(4), 579–599. <https://doi.org/10.1017/S0145553200021313>
- Thompson, D. (2016). *The schematic state: Race, transnationalism, and the politics of the census*. Cambridge: Cambridge University Press.
- U.S. Census Bureau. (1922). *Fourteenth census of the United States taken in the year 1920*. Washington, D.C.: Government Printing Office. <https://hdl.handle.net/2027/uc1.31175019423774>
- U.S. Census Bureau. (2012). *Census coverage measurement estimation report*. https://www.census.gov/coverage_measurement/pdfs/g01.pdf
- U.S. House. Committee on Immigration and Naturalization (1927) *National Origins Provision Immigration Act of 1924*. <https://hdl.handle.net/2027/mdp.39015020459098>

©2020 Dan Bouk. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.

Footnotes

1. Demographers and historians have also looked backward to estimate rates of historical undercounts. (Steckel, 1991; Hacker, 2013) [↵](#)

2. The Census Bureau's confidentiality requirements prevent any individual data from being made public or used by any other part of the government until 72 years have passed. But once that actuarial lifetime has passed by, we can dig into the data that was collected, and not only whatever the Bureau tabulated decades earlier. This is a boon to genealogical researchers, but also to anyone interested in learning more about how a data set can (and can't) reflect a larger truth. [↔](#)