Popul Res Policy Rev
https://doi.org/10.1007/s11113-018-9464-6

ORIGINAL RESEARCH

# Demography in the Big Data Revolution: Changing the Culture to Forge New Frontiers

**Stephanie A. Bohon**[1]

Received: 24 November 2017 / Accepted: 10 March 2018


**Abstract** Despite the widespread and rapidly growing popularity of Big Data, researchers have yet to agree on what the concept entails, what tools are still needed to best interrogate these data, whether or not Big Data's emergence represents a new academic field or simply a set of tools, and how much confidence we can place on results derived from Big Data. Despite these ambiguities, most would agree that Big Data and the methods for analyzing it represent a remarkable potential for advancing social science knowledge. In my Presidential address to the Southern Demographic Association, I argue that demographers have long collected and analyzed Big Data in a small way, by parsing out the points of information that we can manipulate with familiar models and restricting analyses to what typical computing systems can handle or restricted-access data disseminators will allow. In order to better interrogate the data we already have, we need to change the culture of demography to treat demographic microdata as Big. This includes shaping the definition of Big Data, changing how we conceptualize models, and re-evaluating how we silo confidential data.

**Keywords** Big Data · Population-generalizable data · Data security · Complex statistical modeling · Big data demography

This paper is a version of the Presidential Address given to the Southern Demographic Association, Athens, Georgia, October 13, 2016.

✉ Stephanie A. Bohon
sbohon@utk.edu

[1] Department of Sociology, University of Tennessee, 907 McClung Tower, Knoxville, TN 37996, USA

Published online: 28 March 2018

## Introduction

Demographers and other researchers have used Big Data analytics to study poverty eradication, promote sustainable agriculture, end hunger, and improve health (Murdoch and Detsky 2013; Pokhriyal et al. 2015; Vital Wave Consulting 2012; Waga and Rabah 2014). In recent years, cell phone records, Twitter tweets, Google search queries, night-light satellite images, and online prices at retail websites have been analyzed to obtain information in data-poor countries, capture real-time data, and obtain data more cheaply than what is usually available (Chen and Zhang 2014; Kitchin 2014a).

However, the data typically referred to as Big Data represent what Ruggles (2014) refers to as Big "shallow" Data. Indeed, the types of data frequently categorized as Big are *exhaust data*—data accidentally created for purposes not related to research. In fact, these types of data are so common that some US government analysts actually define Big Data as "non-sampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference" (Horrigan 2013). The fact that these data have questionable generalizability is problematic for those seeking to make true statements about people's conditions. Because of this limitation, demographers at the United Nations have recently called for a new data ecosystem which goes beyond exhaust data and encompasses the types of population-generalizable data that are the bases of good demographic analyses (HLG-PCCB 2016).

I agree with Ruggles that demographers have long collected and analyzed potentially Big *deep* Data—large datasets comprised population-generalizable data. Certainly, the entirety of coded US Census data would be one example; 50 years of the Panel Study of Income Dynamics would be another example. Currently, however, our methodological, statistical, and computer training as demographers have left us ill-prepared to tackle the types of problems that can be addressed with these kinds of Big Data. That is, we have Big Data, but we treat it in a small fashion. Even if we knew how to pull four decades of US Census data into a system with a large enough memory or enough processors, would we know what to do with it? Most of us would not. So parsing data into easily useable but small pieces is de rigueur, but doing so also prevents us from visualizing data in its entirety. What are we missing? Additionally, current security practices for handling confidential data such as geo-referenced person-level records limit the computing platforms on which data can be analyzed, which prevents us from building and analyzing really large predictive models, such as social networks with millions of nodes and ties. Here, I discuss demography's new frontier as it comprises advances in computing and the availability of new techniques for combining data. I also explore the ways that demography and demographers must change as we enter this new frontier.

## Revolutions in Science

A few years ago I attended a speech where a National Science Foundation division director speculated that the social sciences are currently where physics was 20 years ago with regard to advanced computing. I was initially stung by having my own work maligned as backward, but I could not entirely disagree with the conclusion that we are behind. More recently, publishing giant, Sage, released a white paper arguing that social science research is at a "turning point" with Big Data (Metzler et al. 2016). Specifically, the Sage-funded researchers argue that the availability of new types of data—Twitter tweets, remote sensing (see Fig. 1), Google searches—and the sheer volume of these data will require new research training, new analytical techniques, and new ontological orientations in the research process, all centered around advanced computing. I agree fully with all of this.

We are at a turning point. Those of us solidly in our mid- or later-career stages experienced such a turning point before, when the development, launch (in August 1991), and widespread adoption of the World Wide Web revolutionized business as usual in demography and other fields. When I started my graduate training in demography at Bowling Green State University in 1992, obtaining an advanced degree in population studies meant attending one of only a handful of programs at universities where Census data were available on data tapes. Writing a master's thesis meant hours in a computer lab writing job control language (JCL) which was processed by a large IBM mainframe computer housed in a remote building. After a proper waiting period, I trudged across campus in order to get a large, green printout

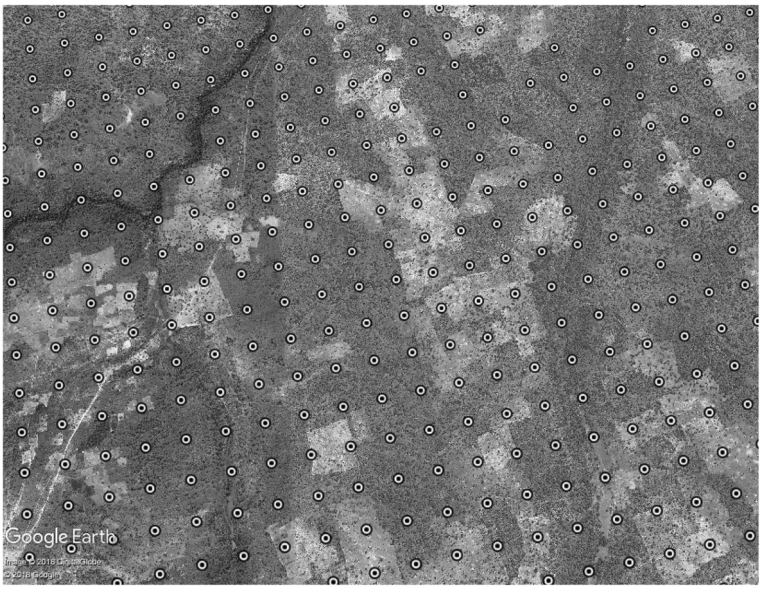


**Fig. 1** Remote sensing (satellite) images like this one of Mali, West Africa allow researchers to obtain a spatial sample that can be combined with data from the Demographic and Health Surveys (Grace et al. 2016; Grace and Nagle 2015)

that often simply read "error." Such was my master's thesis experience. But by the time I was writing my dissertation at Penn State just 3 years later, data could be obtained electronically on a UNIX system using a file transfer protocol, and, in seemingly no time, data were available on personal computers using compact disks. The JCL that I struggled so hard to learn as a first-year master's student was obsolete before I earned a doctoral degree. Before I reached tenure, data could easily be downloaded online. My experience was not unique, and many of today's social scientists had the same experience. Some even remember using punch cards.

With the introduction and expansion of the internet came the easy availability of volumes of disparate data; this changed the social science for good and for bad. Before the internet, demographers often spent their whole careers navigating one set of data. The data they used—which varied from scholar to scholar—were typically the ones they had relatively easy access to and, because of that, the ones they knew thoroughly. But after the internet was launched, this changed. Now a scholar could fill their computer with hundreds of datasets, and access to data was made even easier when the Inter-university Consortium of Political and Social Research at the University of Michigan launched their online platform and the Integrated Public Use Microdata Series (IPUMS) project at the University of Minnesota was established. These innovations opened up a world of data to demographers.

Unfortunately, the availability of a virtual cornucopia of data made it far too easy to do bad things with good data, and, I would argue, there was a short period of crisis in the late 1990s. Widespread data availability and easier-to-use statistical interfaces meant that, in minutes, a researcher could pull an unfamiliar dataset into their computer's memory, and, with some statistical software, he or she could point and click to results that were not necessarily reflective of social truth (McCoach and Adelson 2010). This ease of access to data led to the proliferation of research findings from complexly sampled data analyzed using improperly specified statistical methods (for some discussion of the publication of poor statistical analysis, see Austin and Stuart 2015; Bell et al. 2012). This problem made it necessary to develop and disseminate new statistical methods and refinements and expand the capabilities of statistical software. Many demographers will remember the considerable efforts that Carolina Population Center researchers made in the late 1990s to ensure that researchers did not analyze the Add Health dataset using what was then the current version of SPSS software which could not adjust standard errors distorted by cluster sampling (c.f. Chantala and Tabor 1999; Udry 2003).

Today, we no longer routinely treat all data as though it is randomly sampled and normally distributed. Today, we talk about primary sampling units. We utilize generalized linear models. We redeveloped our existing models to account for data hierarchies. We recognized the problem of spatial autocorrelation. So, the easy access to data made us change how we analyze data, and, because of it, we are all better, more sophisticated data users, better statisticians, and—I would argue—better demographers. So, I agree with Sage that the social sciences, including demography, are reaching a turning point, just as we reached one with the introduction of the internet. Just as we did in the nineties, we are about to see our work world transformed again.

## Data Culture in Demography

Although I agree that we are at a scientific turning point, I cannot entirely agree with the National Science Foundation director's assertion that demographers are 20 years behind the so-called "hard" scientists when it comes to the use of Big Data. We might be even further behind than that. I say this because Big Data means something different for Demography than it does for fields like Physics. In the physical sciences, Big Data entered the fields in a different way than it has for the social sciences. Physical and biological scientists began analyzing Big Data and developing Big Data tools because new, automated instrumentation such as genome sequencers allowed for the creation of massive volumes of the types of information that biologists, chemists, geneticists, and physicists were already analyzing (Hayden 2015). Thus, Big Data in the physical and biological sciences was and is defined by size, and these researchers' primary challenge was to scale up their models and increase their computing power to handle much larger samples (Tripathi et al. 2016).

In the social sciences, Big Data can be voluminous, such as data taken from remote sensing or social media. More importantly, regardless of size, Big social science Data often represent new forms of complex data and new ways of considering existing data that do not just require computers with larger memory. Our problems also require new ways of thinking about the data we have and new methods for analyzing and visualizing these data. We need to think about models without dependent variables or with several of them. We need to de-emphasize theory-driven modeling.[1] We need to rethink formal hypothesis testing. We need to redefine what we mean by a case. We cannot simply scale up; we have to change almost everything. Thus, Big Data demography is arguably a radical new area of demography, while Big Data genetics is mostly genetics with more cases.

A few years ago I attended the annual meetings of the Population Association of America accompanied by a computational chemist and a Big Data security expert from Oak Ridge National Laboratories. I tasked them with attending several paper sessions in order to observe the scope of demographic work and the ways in which demographers approach data. They watched everything with interest. When they returned, impressed by the content of our work and the rigor of our statistical methods, they summarized their observations of our data use in a way that I can paraphrase thusly: *It looks like what demographers do is take data and smash it into tiny bits. Then you pick up a few of the fragments that look interesting and analyze those fragments with complex statistics.*

Their description was neither flattering nor nuanced. Nor was it intended to be critical; these computational scientists were genuinely puzzled by what we were doing with our data. Some of this confusion derives from the differences in fields—in experimental fields, you usually only collect the data you need; in the social

[1] Although I do not discount the tremendous value of theory, I (and others) would argue that a major limitation to taking advantage of so many of the Big Data techniques like machine learning is that they are not designed for theory-driven modeling. As long as program officers at funding agencies and reviewers of journal articles demand theory-driven research, we will never be able to totally engage with Big Data in the way that it has been advanced in other disciplines.

sciences, we usually collect all the data we reasonably can. However, some of what these scientists were reacting to was the fact that we do not take large systems approaches. Thus, their description of our process is largely correct. One of the reasons that demographers love IPUMS is that it allows us to, very easily, select the parts of the American Community Survey (ACS) or other large data sets that we want to analyze, obtain a subsample of these data very quickly, and immediately attend to our analyses. For routine Big Data users, like astrophysicists, it is hard for them to understand why we're not looking at all of the Census data all at once. For the demographers, it has never occurred to us to do so.

Many demographers lack access to computing platforms that would allow us to pull a truly large data set into memory, nor do we have the programming skills to manipulate data on a supercomputer; but even if we did—if we could pull the entire ACS and even every wave of the ACS into memory all at once—then what? What would we do with such data? Part of the Big Data revolution in demography is answering that questions. The good news is that, to the extent that social scientists are entering the Big Data revolution, demographers appear to be leading the charge. Currently, demographers comprise about half of all Big Data social scientists (Metzler et al. 2016). I predict that, in another generation or two, all demographers will be Big Data scientists.

## New Directions for Demography

In order to radically transform the field, changes will occur. I make three assertions about how Demography, as a field, has to evolve if we are to embrace the Big Data revolution in the social sciences: (1) we need to define Big Data; (2) we need to embrace models that are not deductive; and (3) we need to rethink confidential data practices. Before I make these points, let me first acknowledge that there has already been a lot of work on the obstacles to Big Data use in the social sciences (e.g., Manovich 2011; Kitchin 2014b). These obstacles include the fact that most social scientists do not know how to use Python, Hadoop, SQL or other supercomputing tools (Davenport and Patil 2012), and although an increasing number of universities—including my own—are offering majors and advanced degrees in Data Science, the curriculum of these programs often centers on skills that do not translate well to the social sciences. We do not have access to some of what would be the most informative Big Data such as the Internal Revenue Service or US Social Security databases. We are not routinely trained in data mining or graph theory (Dinov 2016). And not all of us have supercomputers at our disposal or would know what to do with one if we had one. These obstacles are large, and I acknowledge them. At the same time, I argue that these obstacles are endemic to the social sciences across the board, of which Demography is but one area. Here I want to focus specifically on the challenges that are unique to Demography.

*First, we need to define Big Data*. Let me be clear that I am not lamenting the lack of a commonly agreed upon operational definition of Big Data, nor am I asserting that we need to determine an arbitrary volume that a data set must exceed in order to be considered Big. What I am arguing is that, at this point in time when

scientists have not yet completely converged on a widely accepted definition of Big Data, we need to ensure that whatever is ultimately determined to be the widely accepted definition of Big Data is a definition that advantages Demography or at least encompasses what Big Data mean to us.

My unabashedly selfish concern over definitions is motivated by the fact that federal funding agencies and private funders are already circulating requests for proposals for Big Data projects, and it is in my own best interests (and yours) to ensure that Demographers are not excluded from these resources. To avoid exclusion, we cannot allow Big Data to be defined by size alone, although size does matter.

Certainly, we live at a time when data on the human condition are becoming available at an exponential rate. In the United States, nearly four million birth records are generated each year (Martin et al. 2017). Combine these data with more than two million death records, millions of Census microdata records, Social Security records, Medicaid records, and more, and the number is staggering. Add to that what has been collected worldwide over more than two hundred years, and the number of available records on the human condition reaches the billions. Ruggles (2014) estimates that there will be at least six billion available public-release microdata records by 2018. This means that, today, we have so much data we can measure it by zetabytes.

Despite this enormous volume of data, I tend to agree with Gary King (2016) that the volume of Big Data is far less important than what you can do with it. King is Harvard's Director of the Institute for Quantitative Social Science who argues that data is nothing compared to a big algorithm:

> Throughout, we need to remember that for the most part, Big Data is not about the data. Data is easily obtainable and cheap, and more so every day. The analytics that turn piles of numbers into actionable insights is difficult, and more sophisticated every day. The advances in making data cheap have been extremely valuable but mostly automatic results of other events in society; the advanced in the statistical algorithms to process the data have been spectacular, and hard fought. Keeping the two straight is critical for understanding the Big Data revolution and for continuing the progress we can make as a result of it (vii).

This focus on Big Data techniques—statistical or geostatistical algorithms that require considerable computational power—rather than data volume that requires a great deal of memory is especially important to demographers, who are arguably less interested in a petabyte (or even zettabyte) of Facebook statuses than expanding the ways in which we can analyze population-generalizable data like censuses that do not fit some computer scientists' working definitions of Big. These population-generalizable data are Big *deep* Data (in contrast to what Ruggles calls Big "shallow" Data (2014, p. 295)). New techniques like linking satellite data to "small" microdata (which is already being done at the Minnesota Population Center) and improving simulations (see Moretti 2002) are where demographers can most usefully push the boundaries of science and knowledge.

Thus, Big Data should not be defined by the size of the raw data but by its complexity and the computing power needed to analyze it in sophisticated ways. I was reminded of this in a conversation with Mark Hayward when he was Director of the Population Research Center at the University of Texas. Hayward lamented that correctly executed statistical bootstrapping, an iterative process of repeatedly producing and averaging standard errors and other estimates, takes considerable time on even a state-of-the-art desktop computer. Hayward has long had an interest in healthy life expectancy—the expected years of good health at birth for a group in a society (Hayward et al. 2014). For multivariate estimation models, bootstrapping in the process of predicting healthy life expectancy for a single group using a computer that was not being used for any other task was taking Hayward and his colleagues almost a month.

Hayward readily admitted to me that a likely bottleneck in his computational process was coding. To the extent that most demographers are skilled in coding, we know enough about how to program in R, SAS, and Stata to produce desired results, but we need more skills to program complex algorithms that run efficiently. (Indeed, a review of Hayward's code by a computer scientist confirmed that rewriting the program would make it run faster by a factor of about 16.) But even if we write a bootstrapping script that is maximally efficient, properly executed bootstrapping needs a multi-core processor if processing time is an issue.[2] Otherwise, we are forced to place arbitrary and small parameters around the number of bootstrap iterations, because we do not want to wait days or weeks for results. We do what is "computationally practical" instead of statistically ideal (Pattengale et al. 2010).

For these reasons, demographers do not want Big Data to be defined by the size of the data set. Nor do we want Big Data to be defined as exhaust data, because the fact that exhaust data naturally occur makes their generalizability problematic. This was an important lesson learned when Google launched Google Flu Trends (GFT), claiming to be able to use Big Data analytics to accurately pinpoint and determine the size of influenza outbreaks as they occur by looking at internet search data (Butler 2008). In 2013, GFT was discontinued after it spectacularly and very publicly erred in its prediction by a margin of 140 percent due to what some analysts called "big data hubris" (Lazer et al. 2014). The inability of most Big Data analyses to be generalized to the larger population has prompted a call for a new data ecosystem which goes beyond exhaust data and encompasses the types of population-generalizable data that are the bases of good demographic analyses (c.f. Letouzé 2015). Ideally, as we move forward and come to terms with what Big Data are and what they can do for demography and demographic analyses, we want data to be defined as Big if they have big computational potential both in memory and parallel processing and if they can be used to make accurate inferences about the social world.

[2] According to Wilcox (2010), a minimum of 599 bootstraps is necessary; however, an exchange by statisticians on the online forum, Cross Validated, reveals that many statisticians consider 100,000 to 1000,000 iterations to be necessary, and decisions are made based on the number a researcher "can afford to wait for." See https://stats.stackexchange.com/questions/86040/rule-of-thumb-for-number-of-bootstrap-samples.

However, it is not enough to embrace increasingly complex algorithms; thus, *second, we need to embrace abductive models.* In research, abduction means moving from the inductive to the deductive, and even moving back and forth between the two logic processes (Bryant and Raja 2014; Crowder and Carbone 2017). To forge new frontiers in research in this way, we have to break free from our worship of *p*-values at the feet of statistical hypothesis testing.

Certainly, I am not the first to say this (c.f., Head et al. 2015). In an influential 2014 article in *Nature*, Nuzzo points out the myriad problems with using *p*-values as indicators of scientific validity, and there are many research questions to which the answers defy formal hypothesis testing. Take, for example, questions related to residential segregation (e.g., Fossett 2006; Iceland et al. 2002) which typically employ measurement techniques that are not statistically inferential. Indeed, Massey and Denton's (1988) influential analyses that establish five dimensions of residential segregation use some techniques without *p*-values to establish the dimensions. However despite the fact that we have long used some non-inferential statistics to examine residential segregation, we seem surprisingly resistant to it in other areas of inquiry. When I recently served as editor of *Population Research and Policy Review*, I was surprised how many reviewers would recommend rejecting a manuscript because of the lack of formal hypotheses in cases where they were not warranted, in my judgment. Work using techniques commonly seen in the physical sciences were almost guaranteed to be scathingly reviewed.

The focus on hypothesis testing is reinforced across our work environment. Requests for proposals from funding agencies often explicitly state that applicants should clearly state their hypotheses, and even the Southern Demographic Association, in their call for papers for competitions, requires students to state hypotheses. I would argue that such calls deter demographers from pursuing abductive work necessary to make new advances in our understanding of demography.

Of course there are analyses where hypothesis testing is warranted, but there are clearly research strategies where hypothesis testing is not the goal. In the past, I have naively assumed that proposal and manuscript reviewers can easily make these distinctions; yet my own experiences suggest otherwise. I have lost count of the number of times, sitting on research review panels, that a reviewer has objected that, "There are no hypotheses." There does not always need to be, and formal hypothesis testing does not guarantee that the science is valid. Big Data analytics will increasingly force us to see that.

As a field, we need to be open to beginning research without a search for a well-formulated relationship involving a single dependent variable. We need to be more open to data-driven approaches that allow us to look at the data holistically to see observable patterns and then begin to test what we think we see. We need to allow demographic researchers to engage in data-driven work without labeling their product "unscientific," for it is just such data-driven research that is the *de rigueur* practices of engineers and many physical scientists who engage with Big Data, and they have advanced their own fields using these methods. For example, from data-driven, abductive approaches, astrophysicists saw that some planets were out of place, which led them to question whether or not planets migrate (Gomes et al.

2005). This observation led to more research about the possibility of planetary migration. The establishment of this possibility—which some astrophysicists now see as the "natural result of planet formation" (Tsiganis et al. 2005, p. 459)—is the breakthrough that may explain why Mars does not have its expected volume.

I have seen abductive research in demography but not often. As part of his doctoral dissertation, Maples (2012) used supercomputers to calculate ethnic niches for every occupation in every metropolitan area in the United States for every single year of ACS data for every country of origin immigrant group for which there were sufficient data. He started with one objective: do the calculations and then look for patterns. It was a difficult proposal to defend to his committee, because the research questions he would pose and the hypotheses he would *eventually* test were unknown before entering the project. Ultimately, Maples was successful in allowing his mentors to let him try. From the output of literally millions of calculations, he was then able to uncover some interesting patterns. One important finding was how Hurricane Katrina dramatically expanded labor market diversity for Mexican workers in New Orleans. Another was how the Great Recession changed the landscape of work in the Chinese-dominated garment industry in San Francisco.

Despite the interesting and important findings, it is difficult to imagine how anyone could have sought external funding for Maples' original work, and research funding is often necessary for Big Data projects in order to pay for computing time. Reviewers are unlikely to take a chance on a project that starts with computing 300 million algorithms on Census employment data just to see the patterns. We are not ready to embrace the "trust me" proposal, but we will need to if we plan to push the boundaries of discovery at more than just a handful of big research institutions. Alternatively, we will have to entrust all Big Data research to demographers working in some government agencies and private firms and the very few whose universities allow unfettered access to a research supercomputer.

Moving to an abductive approach, especially as we begin to embrace advanced computing, also means that we will need to reexamine statistical clustering techniques, improve upon them, and agree to some standards. We will also need to make advances in simulation and develop new ways to conduct large network analyses. Undoubtedly, there are other advances we have not yet conceived. However, even if we advanced our models, the promise of these advances is currently limited because the results from these types of models do not easily lend themselves to examination in static print journals. Readers may need to visualize the movement of the data to fully understand the research, which means the results themselves will need to be available online. Although we certainly have this capability, we have created a financial model around online publishing that is exploitive, requiring authors or taxpayers to bear most of the fiduciary burden of disseminating science in this manner (Fuchs and Sandoval 2013). We do the work and pay publishers to profit from our labor. Until scholars revolt against this exploitation, online publishing options are limited to those with financial resources to use this format.

To illustrate how these issues of abductive work, non-inferential models, visualization, and publishing are compounding, consider an example from new destination migration. New destinations are US states, countries, or metropolitan

areas that have recently become settlement places for immigrants despite having no modern history of immigrant settlement (Singer 2004). Reading across the new destination literature (e.g., Lichter and Johnson 2009; Riosmena and Massey 2012), the clear subtext is that new destinations are inferior to established destinations as sites of new immigrant adjustment, because established destinations already have infrastructure in place to absorb new immigrants quickly and new destinations do not. Is this true? Are places that have had a long history of immigration better suited to absorb new immigrants than new destinations are?

The question is basic and foundational, but it is also a difficult question to answer, because it requires data that are hard to capture. How do you measure place-level infrastructure, and how do you determine which infrastructure is important? Does having access to an organization like a Latin American Cultural Center help Latino immigrants adjust to life in their new society, and could researchers identify such organizations across places given that they likely operate under a variety of names? Would immigrants from Asia also benefit from the presence of Chinese, Japanese, or Korean Societies? What other formal and informal infrastructures are important? The availability of English-language classes? Access to work? Legal processes for easily obtaining driver's licenses? Given these identification and measurement difficulties, I wondered if you could initially approach the problem in an inferential direction. In other words, if researchers look at the characteristics of relatively recent immigrants and limit analyses to those who seem to have been in the same place for their entire stay, could they see patterns of success in some places relative to others and then infer that those places have better, albeit elusive, infrastructure? It is possible that once the places where relatively recent immigrants are most successfully adapting are identified, the search for commonalities among places can more fruitfully begin.

To approach this kind of problem with the type of computing environment to which most of us have easiest access—a desktop or laptop computer—a researcher would first need to identify key predictors of immigrant adaptation success such as educational attainment (Perreira et al. 2006), US citizenship status (Ramakrishnan 2005), ability to speak English (Portes and Rumbaut 2006; Baek Choi and Thomas 2009), and so forth. Then the researcher would need to aggregate data on those key indicators by place, so that for each place, for example, they would have an average immigrant educational attainment score and an indicator of the percent of immigrants who can speak English well. Then the researcher would have to standardize those aggregate point estimates. To determine how many place clusters they have, the researcher would then run an algorithm—chosen from among a seemingly endless array of many different algorithms—to force an $n$ cluster solution. Once the number of clusters is determined (largely by best guess), a different algorithm may need to be used to determine which places fall into which clusters. In sum, at every step the researcher would be forced to make choices, and many of those choices would be made by simply guessing. The researcher would need to choose the measures, aggregation methods, standardization procedures, and algorithms.

Although some of these choices (such as the measures) are guided by the literature, there is little information to help us determine if one algorithm is superior

to another. There are a few tests, some visualization techniques, and some problematic solutions (like chaining) that are easy to identify. However, the information is limited, and the final result will still offer several plausible solutions. Thus, researchers often choose the algorithm whose solutions are most easy to interpret after trying a handful of common algorithms, producing results of unknown validity.

To date, demographers have done so little with cluster analysis techniques that we have not established a set of generally accepted guidelines for how to approach these methods with the types of population data we have. My experientially based opinion is that a consequence and a cause of so little exploration is that many demographers have never learned any statistical clustering techniques. Those who are trained in cluster analysis tend to view the techniques as having little value, since the results are largely dependent on somewhat arbitrary decisions made by the researcher, and we tend to be more comfortable with statistics for which we can test a null hypothesis.

Fortunately, advanced computing can significantly reduce the uncertainty in decision making, but the culture of skepticism regarding non-inferential methods will need to soften if we are to trust the results. Figure 2 illustrates the 2013 solution of a cluster analysis I conducted with the help of computational scientists at Oak Ridge National Lab. The Figure is a US map which shows metropolitan areas classified as good, moderate, or poor based on the level of adaptation success of Mexican immigrants who have been in the United States for just 5 years. Our solution was one in which—because we could run and test millions of solutions very quickly—we did not have to make a lot of blind decisions beforehand. We did not have to decide which indicators of "success" we wanted to use—we used all the


**Fig. 2** Data on Mexican immigrants from the 2013 American Community Survey 5-year estimates are analyzed to reveal the metropolitan area clustering of immigrants who have most successfully adapted to the United States

information the Census provides for our population of interest (in this case, Mexican immigrants in the United States for 5 years). We did not have to limit our analysis to one point in time; we could use time as a test of the stability of our models using each year of the ACS starting in 2005. We did not have to aggregate the data initially, so we could proceed with full information on each immigrant in each metropolitan area. And we did not have to choose between standardization procedures or algorithms; we could test them all and use new techniques to visualize solutions, and we could use machine learning to allow the computer to reject unworkable solutions and test between workable ones. Thus, we did not have to personally check the results of millions of solutions. Working from our findings, we have now begun to use more traditional statistical tests for differences across the places labeled *most*, *moderately*, and *least successful* places. This work is still preliminary, but findings, so far, suggest that places with more diversified labor market opportunities for immigrants are the residences of the most successfully adapted immigrants.

However, despite breakthroughs in reducing uncertainty, the problem of presentation remains. One of the more interesting parts of our analyses of how places facilitate immigrant adjustment is examining whether or not places get better at it as years go by. My computationally savvy colleagues at Oak Ridge helped me create an interactive electronic map where changes in the shading show places getting better or worse at absorbing immigrants over time. However, publishing such results in a print journal is problematic, as it would require allocating pages of print to maps and requiring readers to flip through the pages rapidly, evoking memories of some 1970s children's publications. It sounds like fun, but it is not practical. Publishing in an on-line only format is one solution, but that limits the types of journals where an article can be placed and is often cost-prohibitive. Many simulation studies also yield results with changes in color shading as new simulations are produced. The high cost to the researcher of publishing in color is also a challenge for these types of Big Data analyses.

Despite these limitations, I am optimistic that we will find solutions to our presentation problem soon, because the findings are too important. Ultimately, I believe that if we can examine data in their totality, looking across time and place, we should be able to see fairly clear and never-before-seen patterns in the human condition that should provide insights into the sources, mechanisms, and consequences that lead to economic disparities and other divides. In my conversations about Big Data with Hayward, he speculated that if we could look at the entirety of the Panel Study of Income Dynamics, all at once, we could probably explain the hollowing out of the middle class. I think he is correct, just as soon as we figure out how to do it and how to get it funded and published. Our conventional methods and thinking will not provide answers to some of the most pressing questions in demography.

*Third, and finally, we need to rethink confidential data practices.* If demographers are going to lead the Big Data revolution, we must rethink our security models for protecting restricted data. This may be the biggest cultural change that is required for demographers to make the biggest breakthroughs using Big *deep* Data.

Unlike much of the data from animal science, chemistry, physics, and engineering, social science data often involve human subjects, and great care must be taken to protect the people whose data are collected (Fiske and Hauser 2014). Restricted-access datasets usually provide information about key identifiers such as the respondent's geographic location, and these locational data are often of the most interest to social scientists, because we recognize that human activity has a socio-spatial component (Schwirian 1983; Martin 1996; Leung and Takeuchi 2011; Kwan 2012). It is also these data that add exponential complexity to existing models, and fully exploiting the complexities of place can exhaust the resources of personal computers and workstations. For example, geographic data can be polygons, which cannot be easily merged with microdata with or without advanced computing.

We have two primary methods of getting proprietary forms of geo-coded data into the hands of researchers, and both of these methods unintentionally restrict computationally intensive complex modeling. One method is to require that researchers themselves secure the computing environment where they will analyze the data, such as using a computer that is not networked. The other method is to work in a government-secured facility, especially Federal Statistical Research Data Centers (FSRDC).

To understand how both of these procedures are limiting, consider ongoing work conducted by demographers Elizabeth Fussell and Sara Curran who are interested in how people respond to and recover from extreme weather events like hurricanes and tornadoes (Fussell et al. 2017). We know that some people recovered from disasters like Hurricane Katrina better than others (Cutter et al. 2006), and we can assume that the elderly, people of color, and the poor will be least able to recover from future weather disasters. But we do not fully understand the process by which recovery occurs or not, despite our ability to make reasoned inferences.

Given the fact that most US counties will eventually be struck by a weather-related disaster (Greenough et al. 2001), a potentially useful method for better understanding the recovery process is to link restricted-access, geo-coded, longitudinal microdata to atmospheric data and satellite raster images. Currently, demographers can obtain ACS data linked to environmental data through IPUMS Terra (formerly Terra Populus; Minnesota Population Center 2016); IPUMS Terra is an excellent example of Big Data in demography. The same type of linking, but to confidential, geo-coded longitudinal data like Add Health or PSID, could be extremely valuable in work like Fussell and Curran's. However, common contractual agreements for obtaining and using geo-located microdata are to house those data on a computer disconnected from any network. Often, computers permitted for restricted data use must be housed in a locked room accessible only to the researchers permitted to use the data. Routine backups of data cannot be made, except to the computer housing the data.

However, one of the challenges of the project described here is that such a secure method for handling data use does not allow the researchers the computing power they need to link microdata to satellite raster images that detail climate conditions and local destruction. The complicated data linking is computationally possible, as IPUMS Terra demonstrates, but not in work station (i.e., laptop or desktop) environment. Advanced computing environments for academic research, such as the

NSF-sponsored National Institute for Computations Sciences, are typically multi-million dollar investments of the federal government that are designed for use by thousands of researchers. Taking them off-line for one research team, even for a short period, is not practical, nor is restricting access to a supercomputer to a single research team. Some centers are now building data enclaves to create a secure data environment within a supercomputer, but these enclaves are expensive, will take time to create and test, and new security procedures will need to be established and agreed upon between the major disseminators of confidential data. Likely funders for these projects, like the Office of Advanced Cyberinfrastructure at NSF, will need to be convinced that Demography has enough Big Data users to make such an investment worthwhile. This will take time.

The other way that demographers typically gain access to proprietary data is through FSRDCs, and the computing limitations in that system severely restrict research possibilities. The FSRDC system currently houses much of our restricted-access federal data, and the current trend is to move even more data into FSRDCs (Vilhuber 2016). All of those data are housed on a single system which is a relatively small computing environment by supercomputing standards. Additionally, linking one set of data to another in the FSRDC environment requires extensive permissions, and adding new computing processes is especially difficult. Put simply, getting all the necessary permissions to link FSRDC restricted data to satellite raster data would take years, but the current computing environment does not have the capability to make the linkages efficiently, anyway. Yet the types of data in the FSRDCs are exactly the types of complex data that could provide important new information if analyzed in a much larger environment. One solution is to allow FSRDC data to be analyzed in a data enclave on a much larger system housed elsewhere, but implementing this solution requires traversing complicated political waters, which will also be time consuming and not necessarily lead to success. Another solution is to get Congress to allocate funds for a much larger computing environment for the FSRDC, which is also unlikely.[3]

## Conclusions

My discussion has focused on three of the largest obstacles to engaging with Big Data that are unique to Demography. First, we must push back against those social scientists who want to define Big Data as exhaust data and focus on ensuring that Big Data includes population-generalizable microdata that can be approached with complex algorithms requiring advanced computing. This idea has already been

[3] This Presidential Address was given on the eve of the 2016 Presidential elections. Between the time of the address and this publication, it has become clear that the US Census Bureau is facing a budget crisis. Continuing Resolutions in Congress in 2016 and 2017 froze the overall federal budget at previous levels, which does not provide sufficient funding for the 2020 Census. The Trump administration has asked for additional funding, but the Census Project—a grassroots organization comprised demographers and other stakeholders—believes that the requested additional funds are insufficient for the task, even if given. Thus, the probability that the Census would allocate the funds necessary to upgrade the FSRDC computing environment seems even less likely now than in 2016.

advanced by Ruggles (2014), but other demographers must carry the message to funding agencies and key stakeholders in our work. Second, we must begin to think about data and statistical modeling in new ways, and we must be less conservative in our scrutiny of unfamiliar methods if we are to break new ground. Third, we need to develop and agree to better ways for sensitive data to be available to permitted researchers on computing platforms which have both large memories and multiple processors. We also need to develop cybersecurity policies and techniques that can be applied broadly across social science disciplines as we begin to think about moving data to large systems. Thus, an essential element to which we need to attend as we develop computational demography techniques is privacy-preserving infrastructure, including a statistical differential control (for privacy-protecting analysis) and compartmented computing (for restricting access to data) that will enable researchers to access and analyze data on advanced computing systems while maintaining the required protection of privacy.

Admittedly, I have focused on the problems and offered little in terms of solutions, because much of what is needed is cultural change, and that often takes time. It is also unlikely that a single solution to many of the problems I have indicated will work well across the entire spectrum of Demography, which is a field that always maintains a strong partnership between the academy, government, and business. Additionally, because of our heavy reliance on government agency data, we will continue to be constrained by politics over which we have little control, but with which we need more engagement.

As pessimistic as this sounds, there are also myriad points that suggest grounds for optimism. Demographers are already at the forefront of current advances in Big Data among social scientists. Demographers have long worked hand-in-hand with statisticians, and we are likely to continue to do so as we develop new techniques for approaching our data in a Big manner. We embraced the full possibilities of the internet in less than a decade, so we have already shown a remarkable ability to adapt. My hope is that, in 10 years, the concerns expressed here will appear quaint, as we will have long overcome the obstacles.

Although it is not my desire to trivialize, I cannot help but see the future of Big Data in Demography as a video game where the end goal is known and achievable, but many obstacles must be eliminated, one by one, to reach the goal. We will have to repeatedly try and fail to push the boundaries of science in Big Data before we have resolved all of the issues. These efforts will take a strong will by many. The task is ours, and we are ready for it.

## References

Austin, P. C., & Stuart, E. A. (2015). Moving toward best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine, 34*(28), 3661–3679.

Baek Choi, J., & Thomas, M. (2009). Predictive factors of acculturation attitudes and social support among Asian immigrants in the USA. *International Journal of Social Welfare, 18*(1), 76–84.

Bell, B. A., Onwuegbuzie, A. J., Ferron, J. M., Jiao, Q. G., Hibbard, S. T., & Kromrey, J. D. (2012). Use of design effects and sample weights in complex health survey data: a review of published articles using data from 3 commonly used adolescent health surveys. *American Journal of Public Health, 102*(7), 1399–1405.

Bryant, A., & Raja, U. (2014). In the realm of Big Data. *First Monday* 19(2). http://firstmonday.org/article/view/4991/3822. Accessed 17 Jan 2018.

Butler, D. (2008). Web data predict flu. *Nature, 456,* 287–288.

Chantala, K., & Tabor, J. (1999). Strategies to perform a design-based analysis using the Add Health data. Resource document. Carolina Population Center, University of North Carolina at Chapel Hill. http://www.cpc.unc.edu/projects/addhealth/documentation/guides/weight1.pdf. Accessed 17 Jan 2018.

Chen, C. L. P., & Zhang, C. (2014). Data-intensive applications, challenges, techniques an technologies: a survey on Big Data. *Information Sciences, 275,* 314–347.

Crowder, J. A., & Carbone, J. A. (2017). Abductive artificial intelligence learning models. In H. R. Arabnia, D. de la Fuente, E. B. Kozerenko, J. A. Olivas, & F. G. Tinetti (Eds.), *Proceedings of the 2017 International Conference on Artificial Intelligence* (pp. 90–96). Las Vegas: CSREA Press.

Cutter, S. L., Emrich, C. T., Mitchell, J. T., Boruff, B. J., Gall, M., Schmidtlein, M. C., et al. (2006). The long road home: race, class, and recovery from Hurricane Katrina. *Environment: Science and Policy for Sustainable Development, 4*(2), 8–20.

Davenport, T. H., & Patil, D. J. (2012). Data scientist—the sexiest job of the 21st century: meet the people who can coax treasure out of messy, unstructured data. *Harvard Business Review, 95*(5), 70–76.

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience, 5*(1), 1–15.

Fiske, S. T., & Hauser, R. M. (2014). Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences, 111*(38), 13675–13676.

Fossett, M. (2006). Ethnic preferences, social distance dynamics, and residential segregation: theoretical explorations using simulation analysis. *The Journal of Mathematical Sociology, 30*(3–4), 185–273.

Fuchs, C., & Sandoval, M. (2013). The diamond model of open access publishing: why policy makers, scholars, universities, libraries, labour unions and the publishing world need to take non-commercial, non-profit open access serious. *TripleC: Communication, Capitalism & Critique, 11*(2), 428–443.

Fussell, E., Curran, S. R., Dunbar, M. D., Babb, M. A., Thompson, L., & Meijer-Irons, J. (2017). Weather-related hazards and population change: a study of hurricanes and tropical storms in the United States, 1980-2012. *The Annals of the American Academy of Political and Social Science., 669*(1), 146–167.

Gomes, R., Levinson, H. F., Tsiganis, K., & Morbidelli, A. (2005). Origin of the cataclysmic late heavy bombardment period of the terrestrial plants. *Nature, 4353,* 466–469.

Grace, Kathryn, & Nagle, Nicholas. (2015). Using high resolution remotely sensed data to examine the relationship between agriculture and fertility in a pre-transitional setting: a case study of Mali. *The Professional Geographer, 67*(4), 641–654.

Grace, Kathryn, Nagle, Nicholas N., & Husak, Greg. (2016). Can small-scale agricultural production improve children's health? examining stunting vulnerability among very young children in Mali, West Africa. *Annals of the Association of American Geographers, 106*(3), 722–737.

Greenough, G., McGeehin, M., Bernard, S. M., Trtanj, J., Riad, J., & Engelberg, D. (2001). The potential impacts of climate variability and change on health impacts of extreme weather events in the United States". *Environmental Health Perspectives, 109*(Supp 2), 191–198.

Hayden, E. C. (2015). Genome researchers raise alarm over Big Data. *Nature: International Weekly Journal of Science.* http://www.nature.com/news/genome-researchers- raise-alarm-over-big-data-1.17912. Accessed 17 Jan 2018.

Hayward, M. D., Hummer, R. A., Chiu, C., Gonzalez-Gonzalez, C., & Wong, R. (2014). Does the Hispanic paradox in mortality extend to disability? *Population Research and Policy Review, 33,* 81–96.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology.* http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106. Accessed 17 Jan 2018.

HLG-PCCB (High-level group for partnership, coordination and capacity-building for statistics for the 20130 agenda for sustainable development). (2016). Global action plan for sustainable development

data. Report. https://unstats.un.org/sdgs/files/global-consultation-hlg-1/GAP_HLG-20161021.pdf. Accessed 17 Jan 2018.
Horrigan, M. W. (2013). Big data and official statistics. presentation for the international year of statistics. Bureau of Labor Statistics, Office of Prices and Living Conditions Washington, DC
Iceland, J., Weinberg, D. H., & Steinmetz, E. (2002). *Racial and ethnic residential segregation in the United States: 1980–2000*. Washington, DC: US Census Bureau, Series CENSR-3.
King, G. (2016). Preface: big data is not about the data. In R. Michael Alvarez (Ed.), *Computational social science: discovery and prediction*. Cambridge: Cambridge University Press.
Kitchin, R. (2014a). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1–12.
Kitchin, R. (2014b). *The data revolution: big data, open data, data infrastructures & their consequences*. Los Angeles: Sage.
Kwan, M. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers, 102*(5), 958–968.
Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203–1205.
Letouzé, E. (2015). Demography, meet big data; big data, meet demography: reflections on the data-rich future of population science. In Paper presented at the United Nations EGM on strengthening the demographic evidence base for the post-2015 development agenda. New York, October 5.
Leung, M., & Takeuchi, D. T. (2011). Race, place, and health. In L. M. Burton, P. Kemp, M. Leung, S. A. Matthews, & D. T. Takeuchi (Eds.), *Communities, neighborhoods, and health: expanding the boundaries of place* (pp. 73–88). New York: Springer.
Lichter, D. T., & Johnson, K. M. (2009). Immigrant gateways and Hispanic migration to new destinations. *International Migration Review, 43*(3), 496–518.
Manovich, L. (2011). Trending: the promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the Digital Humanities 2* (pp. 460–475). Minneapolis: University of Minnesota.
Maples, J. N. (2012). *Changes in US Ethnic Niches,* 2005-2010. Doctoral Dissertation, University of Tennessee. http://trace.tennessee.edu/socioetds/. Accessed 17 Jan 2018.
Martin, D. (1996). *Geographic information systems: socioeconomic applications*. New York: Routledge.
Martin, J. A., Hamilton, B. E., Osterman, M. J. K., Driscoll, A. K., & Matthews, T. J. (2017). Births: final data for 2015. *National Vital Statistics Reports, 66,* 1–70.
Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation". *Social Forces, 67*(2), 281–315.
McCoach, D. B., & Adelson, J. L. (2010). Dealing with dependence (Part I): understanding the effects of clustered data. *Gifted Child Quarterly, 54*(2), 152–155.
Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). *Who is doing computational social science? A white paper*. Sage Publishing. https://us.sagepub.com/sites/default/files/compsocsci.pdf. Accessed 17 Jan 2018.
Minnesota Population Center. (2016). *Terra populus: integrated data on population and environment: version 1*. Minneapolis: University of Minnesota.
Moretti, S. (2002). Computer simulations in sociology: what contribution? *Social Science Computer Review, 20*(1), 43–57.
Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of Big Data to health care. *JAMA, 309*(13), 1351–1352.
Nuzzo, R. (2014). Statistical errors: *p* values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature, 506,* 150–152.
Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E., & Stamatakis, A. (2010). How many bootstrap replicates are necessary?". *Journal of Computational Biology, 17*(3), 337–354.
Perreira, K. M., Harris, K. M., & Lee, D. (2006). Making it in America: high school completion by immigrant and native youth. *Demography, 43*(3), 511–536.
Pokhriyal, N., Dong, W., & Govindaraju, V. (2015). *Big data for improved diagnosis of poverty: a case study of Senegal*. Washington, DC: A report for the brookings institution africa in focus series.
Portes, A., & Rumbaut, R. G. (2006). *Immigrant America: a portrait*. Berkeley: University of California Press.
Ramakrishnan, S. K. (2005). *Democracy in Immigrant America: changing demographics and political participation*. Palo Alto: Stanford University Press.
Riosmena, F., & Massey, D. S. (2012). Pathways to El Norte: origins, destinations, and characteristics of Mexican migrants to the United States. *International Migration Review, 46*(1), 3–36.
Ruggles, S. (2014). Big microdata for population research. *Demography, 51*(1), 287–297.

Schwirian, K. P. (1983). Models of neighborhood change. *Annual Review of Sociology, 9,* 83–102.
Singer, A. (2004). *The rise of new immigrant gateways*. Washington, DC: Brookings Institution, Center on Urban and Metropolitan Policy.
Tripathi, R., Sharma, P., Chakraborty, P., & Varadwaj, P. K. (2016). Next-generation sequencing revolution through big data analytics. *Frontiers in Life Science, 9*(2), 119–149.
Tsiganis, K., Gomes, R., Morbidelli, A., & Levinson, H. F. (2005). Origin of the orbital architecture of the giant planets of the Solar system. *Nature, 435*(7041), 459–461.
Udry, J. R. (2003). *The national longitudinal study of adolescent health (Add Health), Wave 1, 1994*. Chapel Hill: Carolina Population Center, University of North Carolina.
Vilhuber, L. (2016). Census research nodes: a progress report. In *Presentation at the 2016 FSRDC Research Conference*. September 15. College Station, Texas.
Vital Wave Consulting. (2012). Big data, big impact: new possibilities for international development. A report for the World Economic Forum. Geneva, Switzerland.
Waga, D., & Rabah, K. (2014). Environmental conditions', big data management, and cloud computing analytics for sustainable agriculture. *World Journal of Computer Application and Technology, 2*(3), 73–81.
Wilcox, R. R. (2010). *Fundamentals of modern statistical methods: substantially improving power and accuracy*. New York: Springer.