The Potential for Big Data to Improve Neighborhood-Level Census Data

Seth E. Spielman

Abstract The promise of "big data" for those who study cities is that it offers new ways of understanding urban environments and processes. Big data exists within broader national data economies, these data economies have changed in ways that are both poorly understood by the average data consumer and of significant consequence for the application of data to urban problems. For example, high resolution demographic and economic data from the United States Census Bureau since 2010 has declined by some key measures of data quality. For some policy-relevant variables, like the number of children under 5 in poverty, the estimates are almost unusable. Of the 56,204 census tracts for which a childhood poverty estimate was available 40,941 had a margin of error greater than the estimate in the 2007–2011 American Community Survey (ACS) (72.8 % of tracts). For example, the ACS indicates that Census Tract 196 in Brooklyn, NY has 169 children under 5 in poverty ±174 children, suggesting somewhere between 0 and 343 children in the area live in poverty. While big data is exciting and novel, basic questions about American Cities are all but unanswerable in the current data economy. Here we highlight the potential for data fusion strategies, leveraging novel forms of big data and traditional federal surveys, to develop useable data that allows effective understanding of intra urban demographic and economic patterns. This paper outlines the methods used to construct neighborhood-level census data and suggests key points of technical intervention where "big" data might be used to improve the quality of neighborhood-level statistics.

Keywords Census • American Community Survey • Neighborhood data • Uncertainty • Data fusion

S.E. Spielman (✉)
University of Colorado, Boulder, CO, USA
e-mail: seth.spielman@colorado.edu

P. Thakuriah et al. (eds.), *Seeing Cities Through Big Data*, Springer Geography,
DOI 10.1007/978-3-319-40902-3_6

1 Introduction

The promise of "big data[1]" for those who study cities is that it offers new ways of understanding urban environments and their affect on human behavior. Big data lets one see urban dynamics at much higher spatial and temporal resolutions than more traditional sources of data, such as survey data collected by national statistical agencies. Some see the rise of big data as a revolutionary of mode of understanding cities, this "revolution" holds particular promise for academics because, as argued by Kitchin (2014), revolutions in science are often preceded by revolutions in measurement. That is, big data could give rise to something even bigger, a new science of cities. Others, such as Greenfield (2014) argue that real urban problems cannot be solved by data and are deeply skeptical of the potential for information technologies to have meaningful impacts on urban life. Here, we aim to contextualize the enthusiasm about urban big data within broader national data economies, particularly focusing on the US case. This paper argues that changes to national data infrastructures, particularly in the US, have led to decline of important sources of neighborhood-level demographic and economic data and that these changes complicate planning and policymaking, even in a big data economy. We argue that in spite of the shortcomings of big data, such as uncertainty over who or what is being measured (or not measured), it is possible to leverage these new forms of data to improve traditional survey based data from national statistical agencies.

The utility of data is contingent upon the context within which the data exist. For example, one might want to know the median income in an area and the crime rate in isolation these data are far less useful then they are in combination. Knowing that place is high crime might help guide policy. However, policy might be much more effectively targeted in one knew context within which crime was occurring. In a general sense we might refer to the context within which data exist as the "data economy." For those who work on urban problems, the data economy has changed in ways that are both poorly understood by the average data consumer and of consequence to the application of big data to urban problems. Traditional sources of information about cities in the US have recently changed in profound ways. We argue that these changes create potential, and problems, for the application of data to urban questions.

In particular, the data collected by the US Census Bureau has recently undergone a series of dramatic changes, some of these changes are a result of the gradual accrual of broader social changes and some have been abrupt, the result of changes to federal policy. The National Research Council (2013) document a gradual long term national trend of increases in the number of people who refuse to respond to public (and private) surveys. Geographic and demographic patterns in survey

[1] Defining big data is difficult, most existing definitions, include some multiple of V's (see Laney 2001). All are satisfactory for our purposes here. We use the term to distinguish between census/survey data which we see as "designed" measurement instruments and big data which we see as "accidental" measurement instruments.

non-response make it difficult for surveys to accurately describe populations and create the need for complex statistical adjustments to ensure that the estimates produced by the survey are representative of the target population. If for example, low income immigrants do not respond to official surveys they would be invisible to the data-centric urban analyst. More realistically, if they respond with a much lower frequency than the average person then they would appear much less prevalent than they actually are unless one accounts for their differential response rate when producing estimates. However, efforts to reduce bias due to non-response can add uncertainty to the final estimate creating large margins of error and complicating data use.

In fact, high levels of uncertainty now plague almost all fine resolution[2] urban data produced by the United States Census Bureau (USCB). Neighborhood-level data from the Census Bureau are terribly imprecise, for some policy-relevant variables, like the number of children in poverty, the estimates are almost unusable—of the 56,204 tracts for which a poverty estimate for children under 5 was available 40,941 had a margin of error greater than the estimate in the 2007–2011 ACS (72.8 % of tracts). For example, the ACS indicates that Census Tract 196 in Brooklyn, NY has 169 children under 5 in poverty $\pm$174 children, suggesting somewhere between 0 and 343 children in the area live in poverty. Users of survey data often face the situation in Table 1, which shows the ACS median income estimates for African-American households for a contiguous group of census tracts in Denver, Colorado. Income estimates range from around $21,000 to $60,000 (American Factfinder website accessed 7/15/2013). Without taking account of the margin of error, it would seem that Tract 41.06 had the highest income, however, when one accounts for the margin of error, the situation is much less clear—Tract 41.06 may be either the wealthiest or the poorest tract in the group.

The uncertainty in Table 1 is all but ignored by practicing planners, a voluntary online survey of 180 urban planners that we conducted during 2013 found that most planners (67 %) simply delete or ignore information about the quality of estimates, like the margin of error, when preparing maps and reports. This practice, according to planners is driven by the "demands" of their "consumers." That is, the audience for their maps and reports would have difficulty incorporating the margins of error into decision-making processes. This practice is further reinforced by federal agencies, which use only the tract level estimates to determine eligibility for certain programs (for example, see the eligibility guidelines for the Treasury's New Markets Tax Credit program). The problem with the margins of error is especially pronounced for the Census Transportation Planning Package (CTTP), a key input for transportation planning and travel demand models.

[2] We use the terms "fine" and "high" resolution to refer to census tract or smaller geographies, these data are commonly conceived of as "neighborhood-scale" data. We conceive of resolution in the spatial sense, higher/finer resolution means a smaller census tabulation unit. However, the geographic scale high resolution of census units is a function of population density.

Table 1 2006–2010 ACS estimates of African-American median household income in a selected group of proximal tracts in Denver County, Colorado

Tract number	African-American median household income	Margin of error
Census Tract 41.01	$28,864	$8650
Census Tract 41.02	$21,021	$4458
Census Tract 41.03	$43,021	$14,612
Census Tract 41.04	$36,092	$3685
Census Tract 41.06	$60,592	$68,846

The decline in the quality of neighborhood scale data in the United States began in 2010, the year the American Community Survey (ACS) replaced the long form of the United States decennial census as the principal source of high-resolution geographic information about the U.S. population. The ACS fundamentally changed the way data about American communities are collected and produced. The long form of the decennial census was a large-sample, low frequency national survey; the ACS is a high-frequency survey, constantly measuring the American population using small monthly samples. One of the primary challenges for users of the ACS is that the margins of error are on average 75 % larger than those of the corresponding 2000 long-form estimate (Alexander 2002; Starsinic 2005). This loss in precision was justified by the increase in timeliness of ACS estimates, which are released annually (compared to the once a decade long form). This tradeoff prompted Macdonald (2006) to call the ACS a "warm" (current) but "fuzzy" (imprecise) source of data. While there are clear advantages to working with "fresh" data, the ACS margins of error are so large that for many variables at the census tract and block group scales the estimates fail to meet even the loosest standards of data quality.

Many of the problems of the American Community Survey are rooted in data limitations. That is at critical stages in the creation of neighborhood-level estimates the census bureau lacks sufficient information and has to make assumptions and/or use data from a coarser level of aggregation (municipality or county). We argue that one of the major potential impacts of big data for the study of cities is the reduction of variance in more traditional forms demographic and economic information. To support this claim, we describe the construction of the ACS in some detail, with the hope that these details illuminate the potential for big data to improve federal and/or state statistical programs.

2 Understanding the American Community Survey

Like the decennial long form before it, the ACS is a sample survey. Unlike complete enumerations, sample surveys do not perfectly measure the characteristics of the population—two samples from the same population will yield different estimates. In the ACS, the margin of error for a given variable expresses a range

of values around the estimate within which the true value is expected to lie. The margin of error reflects the variability that could be expected if the survey were repeated with a different random sample of the same population. The statistic used to describe the magnitude of this variability is referred to as *standard error* (SE). Calculating standard errors for a complex survey like the ACS is not a trivial task, the USCB uses a procedure called Successive Differences Replication to produce variance estimates (Fay and Train 1995). The margins of error reported by the USCB with the ACS estimates are simply 1.645 times the standard errors.

One easy way to understand the ACS Margin of Error is to consider the simple case, in which errors are simply a function of the random nature of the sampling procedure. Such sampling error has two main causes, the first is the sample size—the larger the sample the smaller the standard error, intuitively more data about a population leads to less uncertainty about its true characteristics. The second main cause of sampling error is heterogeneity in the population being measured (Rao 2003). Consider two jars of U.S. coins, one contains U.S. pennies and the other contains a variety of coins from all over the world. If one randomly selected five coins from each jar, and used the average of these five to estimate the average value of the coins in each jar, then there would be more uncertainty about the average value in the jar that contained a diverse mixture of coins. If one took repeated random samples of five coins from each jar the result would always be the same for the jar of pennies but it would vary substantially in the diverse jar, this variation would create uncertainty about the true average value.[3] In addition, a larger handful of coins would reduce uncertainty about the value of coins in the jar. In the extreme case of a 100 % sample the uncertainty around the average value would be zero. What is important to realize is that in sample surveys the absolute number of samples is much more important than the relative proportion of people sampled, a 5 % sample of an area with a large population will provide a much better estimate than a 5 % sample of a small population. While the ACS is much more complicated than pulling coins from a jar, this analogy helps to understand the standard error of ACS estimates. Census Tracts (and other geographies) are like jars of coins. If a tract is like the jar of pennies, then the estimates will be more precise, whereas if a tract is like the jar of diverse coins or has a small population, then the estimate will be less precise.

While the simple example is illustrative of important concepts it overlooks the central challenge in conducting surveys; many people included in a sample will choose not to respond to the survey. While a group's odds of being included in the

[3] The Census Bureau generally is not actually estimating the "average" value, they are estimating the "total" value of coins in the jar. Repeatedly grabbing five coins and computing the average will over many samples get you a very precise estimate of the average value, but it will give you no information on the total value. To get the total value, you need a good estimate of the average AND a good estimate of the total number of coins in the jar. The loss of cotemporaneous population controls caused by decoupling the ACS from the Decennial enumeration means that the census does not have information about the number of coins in the jar. This is discussed in more details later.

ACS sample are proportional to its population size, different groups of people have different probabilities of responding. Only 65 % of the people contacted by the ACS actually complete the survey (in 2011, 2.13 million responses were collected from 3.27 million samples). Some groups are more likely to respond than others, this means that a response collected from a person in a hard to count group is worth more than a response from an easy to count group. Weighting each response controls for these differential response rates. In the ACS each completed survey is assigned a single weight through a complex procedure involving dozens of steps. The important point, as far as this paper is concerned, is that these weights are *estimated* and uncertainty about the appropriate weight to give each response is an important source of uncertainty in the published data.

3 Sampling

Before 1940, the census was a complete enumeration; each and every housing unit (HU) received the same questionnaire. By 1940 the census forms had become a long, complicated set of demographic and economic questions. In response, the questionnaire was split in 1940 into a short set of questions asked of 100 % of the population and an additional "long form" administered to a subset of the population. Originally, this long form was administered to a 5 % random sample, but in later years it was sent to one HU in six (Anderson et al. 2011). Before 1940 any error in the data could be attributed either to missing or double counting a HU, to incorrect transcription of a respondent's answer, or to intentional/unintentional errors by the respondent. After 1940, however, the adoption of statistical sampling introduced new sources of uncertainty for those questions on the long form.

Up until 2010 the sample based (long form) and the complete enumeration (short form) of the census were administered at the same time. In 2010 the ACS replaced the sample based long form. The American Community Survey constantly measures the population; it does not co-occur with a complete census. The lack of concurrent complete count population data from the short form is a key source of uncertainty in the ACS. Prior to the rise of the ACS, short form population counts could serve as controls for long-form based estimates. The decoupling of the sample from the complete enumeration accounts for 15–25 % of the difference in margin of error between the ACS and the decennial long form (Navarro 2012). Population controls are essential to the ACS sample weighting process, now population controls are only available for relatively large geographic areas such as municipalities and counties. This is a key data gap which as discussed later might be addressed with big data.

4 Spatial and Temporal Resolution of Census Estimates

Prior to the advent of sampling, the complete count census data could, in principle, be tabulated using any sort of geographic zone. Tract based census data has become a cornerstone of social science and policy making. the decennial census by the late twentieth century. However, users of the once a decade census were increasingly concerned about the timeliness of the data (Alexander 2002). A solution to this problem was developed by Leslie Kish, a statistician who developed the theory and methods for "rolling" surveys (Kish 1990).

Kish's basic idea was that a population could be divided into a series of non-overlapping annual or monthly groups called subframes. Each subframe would then be enumerated or sampled on a rolling basis. If each subframe were carefully constructed so as to be representative of the larger population, then the annual estimates would also be representative, and eventually, the entire population would be sampled. The strength of this rolling framework is its efficient use of surveys. The decennial census long form had to sample at a rate appropriate to make reasonable estimates for small geographic areas such as census tracts, which contain on average 4000 people. Therefore, citywide data released for a municipality of, say, one million people would be based on considerably more samples than necessary. Spreading the samples over time lets larger areas receive reasonable estimates annually, while smaller areas wait for more surveys to be collected. The rolling sample therefore increases the frequency of data on larger areas. The primary cost comes in the temporal blurring of data for smaller areas. The advent of sampling made census data for small geographic areas less precise. Since there are a finite number of samples in any geographic area, as tabulation zones become smaller sample sizes decline, making estimates more uncertain. The rise uncertainty is greater for small populations; for instance the effects of reducing a sample size from 200 to 100 is much greater than the effect of reducing a sample size from 20,000 to 10,000. The USCB counteracts this decline in sample size by pooling surveys in a given area over multiple years, thus diluting the temporal resolution of the estimates.

Rolling sampling is straightforward in the abstract. For example, suppose that there are $K = 5$ annual subframes, that the population in a tract is known ($N = 1000$), that the sampling rate is $r = 1/6$, and that the response rate is 100 %; then one would sample $n = N/(K*1/r)$ people per year. Over a 5-year period 1/6 of the population would be sampled and each returned survey would represent $w =$ (N/n)/K people, where w is the weight used to scale survey responses up to a population estimate. In this simple case, the weight assigned to each survey would be the same. For any individual attribute y, the tract level estimate would be $y_t = \Sigma w_i y_i$ (equation 1), a weighted summation of all i surveys collected in tract t. If the weights are further adjusted by ancillary population controls X, then the variance of the estimate is $\Sigma w_i^2 \mathrm{VAR}[y_i|X]$ (equation 2; Fuller 2011, assuming independence.). If the rolling sample consisting of long-form-type questions were administered simultaneously with a short form census, then all the parameters in our simple example (N,K, X) would be known.

However, in the ACS good population controls are not available for small areas (N and X are unknown) because, unlike the long form, the survey is not contemporaneous with the complete enumeration decennial census. Thus weights (w) for each response must be estimated and this is an important source of uncertainty in the ACS.

5 Weighting

In the ACS each completed survey is assigned a weight (w) that quantifies the number of persons in the total population that are represented by a sampled household/individual. For example, a survey completed by an Asian male earning $45,000 per year and assigned a weight of 50 would in the final tract-level estimates represent 50 Asian men and $2.25 million in aggregate income. The lack of demographically detailed population controls, and variations in response rate all necessitate a complex method to estimate w_i. The construction of ACS weights is described in the ACS technical manual (which runs hundreds of pages, U.S. Census Bureau 2009a). Individually these steps make sense but they are so numerous and technically complex that in the aggregate they make the ACS estimation process nearly impenetrable for even the most sophisticated data users. The cost of extensive tweaking of weights is more than just lack of transparency and complexity. Reducing bias by adjusting weights carries a cost. Any procedure that increases the variability in the survey weights also increases the uncertainty in tract-level estimates (Kish 2002). Embedded in this process is a trade-off between estimate accuracy (bias) and precision (variance/margin of error), refining the survey weights reduces bias in the ACS but it also leads to variance in the sample weights.

6 Big Data and Public Statistics

Without traditional survey data from national statistical agencies, like the USCB, it is difficult to contextualize big data, its hard to know who is (and who is not) represented in big data. It is difficult to know if there are demographic, geographic, and or economic biases in the coverage of big data without traditional census data as a baseline. Ironically, as this baseline data declines in quality, many of the populations most in need of urban services are least well served by the traditional census data and quite possibly the big data as well—consider the example of young children in poverty discussed in the introduction.

In the preceding sections we identified several key data gaps and methodological decisions that might be addressed with big data:

1. Sampling is constrained by a lack of detailed high geographic and demographic resolution population data.
2. Small area geographies are not "designed" and this leads to degradation in the quality of estimates and the utility of the published data.
3. Weights are complex and difficult to accurately estimate without additional data.

In this section we outline how big data might be used to address these issues. This section is by no means exhaustive, the aim more to draw attention to the potential for new forms of data to mitigate emerging problems with neighborhood statistics. It is also important to note that, for reasons discussed in the conclusion, this section is largely speculative, that is, very few of the ideas we propose have seen implementation.

So far this paper has emphasized the mechanics of the construction of the ACS—sampling, the provision of small area estimates, the provision of annual estimates, and the estimate of survey weights. The prior discussion had a fair amount of technical detail because such detail is necessary in order to understand how novel forms of "big" data might be integrated into the production process. Directly integrating big data into the production of estimates is not the only way to use new forms of data concurrently with traditional national statistics, but in this paper the emphasis is on such an approach.

It should be apparent that the data gaps and methodological choices we have identified thus far are intertwined. For example, the use of sampling necessitates the estimation of survey weights which are complicated to estimate when very little is known about the target population in the areas under investigation. Spatial and temporal resolution are related because the reliability of the estimate depends on the number of surveys, which accrue over time, and the size (population) and composition of the area under investigation.

The lack of detailed small area population controls is makes it very difficult to estimate the weight for each survey. Since the US Census Bureau does not know how many low income Caucasian males live in each census tract it is difficult to know if the number of surveys returned by low income Caucasian males higher or lower than expected—this affects the weight assigned to a response. For example, imagine a hypothetical census tract with 2000 housing units and a population of 4000 people. 10 % of the population is low-income white males and this tract was sampled at a 5 % rate, one would expect 10 % of the completed surveys to be filled in by low-income white males. However, if this group is less likely than others to respond perhaps the only 2 % of the completed surveys would be completed by low-white males. If the number of low-income white males was known in advance one could "up-weight" these responses to make sure that in the final data low income-white males represented 10 % of the population. However, the census bureau has no idea how many low-income white males are in each census tract. This is where big data might help.

If, for example, the number of low-income white males could be estimated by using credit reporting data, social media profiles, or administrative records from other government agencies, then a lot of the guesswork in deciding how to weight survey responses could be eliminated. It's important to realize that these forms of "big" data might not be of the highest quality. However, they could be used to establish meaningful benchmarks for sub-populations making simple comparisons of "big" and traditional data possible. While it would be difficult to say which data was "correct" it is reasonable to suggest that large discrepancies would warrant closer inspection and would highlight key differences in the coverage of the various data sets. These coverage differences are not especially well understood at the time of writing.

A more sophisticated strategy would be to employ what is called are called "model assisted estimation" strategies (see Särndal 1992). Model assisted estimation is a set of strategies for using ancillary data and regression models to estimate survey weights. Currently, the ACS uses a model assisted strategy called "Generalized Regression Estimator" (GREG). In the ACS GREG takes advantage of person-level administrative data on age, race, and gender of residents from auxiliary sources such as the Social Security Administration, the Internal Revenue Service, and previous decennial census tabulations. The procedure builds two parallel datasets for each census tract: one using the administrative data on all people in the tract, and the second using administrative data for only the surveyed housing units. The second dataset can be viewed, and tested, as an estimate of the demographic attributes of the first—e.g., proportions of males aged 30–44, non-Hispanic blacks, etc. A weighted least squares regression is then run on the second dataset, in which the dependent variable is weighted HU counts and the independent variables are the various weighted attribute counts.

The strength of model assisted estimation procedure depends entirely on the quality of the regression. A well-fit regression should reduce overall uncertainty in the final ACS estimates by reducing the variance of the weights, while a poorly fit regression can actually increase the margin of error. The data used in models assisted estimation in the ACS is terrible for its intended purpose, that is age, sex, and race are only loosely correlated with many of the economic and demographic characteristics of most interest to urban planners and policy makers. In spite of these weaknesses Age, Sex, and Race data are used because they are available to the USCB from other Federal agencies, more sensitive data, like income, is not incorporated into estimates.

However, data on homeownership, home values, spending patterns, employment, education and many other attributes may be obtainable through big data sets and this could be used to improve the quality of estimates through model assisted estimation. For example, housing data from cadastral records and home sales could be (spatially) incorporated into the ACS weighting strategy. The exact home value of each house is unknown, so they are unusable as hard benchmarks. But, it is possible to approximate the value of each house based upon location, characteristics, and nearby sales. Even if it was not possible to directly match survey respondent to records in other datasets, it might be possible to geospatially impute such

characteristics. For example, recent nearby home sales might be used to estimate the value of a respondents' home. This approximation is used to great effect by the mortgage industry and by local governments for property tax assessments. Since these models are approximations, the data may enter the weighting phase as "soft" benchmarks (i.e. implemented a mixed effects models). It is not appropriate for the weights to exactly duplicate the estimated home value, but it is appropriate for the weights to approximate the estimated home value. For example, Porter et al. (2013) use the prevalence of Spanish language Google queries to improve census estimates of the Hispanic population. Carefully chosen controls have the potential to dramatically reduce the bias and margin of error in ACS estimate for certain variables. The estimates most likely to be impacted are socioeconomic variables, which are poorly correlated with the currently available demographic benchmarks, and thus have a disproportionately large margin of error.

A second mechanism for using big data to improve estimates is through zone design. Census geographies are designed to be stable over time, that is, local committees at some point designed them in the past (often 3o years ago) and they have only evolved through splits and merges with other census tracts. Splits and merges can only occur when the tract population crosses some critical threshold. The size and shape of census fundamentally affects the quality of estimates. Larger population census tracts, because they generally have more surveys supporting estimates have higher quality data. However, as geographies grow in size there is potential to loose information on intra urban variation. However, information loss does not necessarily occur as a result of changes in zone size. Consider two adjacent census tracts that are very similar to each other in terms of ethnic composition, housing stock, and economic characteristics. The cost of combining these two census tracts into a single area is very small. That is, on a thematic map these two adjacent areas would likely appear as a single unit (because they would be the same legend color because they would likely have the same value). Combining similar places together boosts the number of completed surveys and thus reduces the margin of error. The challenge is how does one tell if adjacent places are similar (or not) when the margins of error on key variables are very large? Again, big data, if it provides a reasonable approximation of the characteristics of the places at high spatial resolutions it maybe possible to combine lower level census geographies into units large enough to provide high quality estimates. For example, Spielman and Folch (2015) develop an algorithm to combine existing lower-level census geographies, like tracts and block groups, into larger geographies while producing new estimates for census variables such that the new estimates leverage the larger population size and have smaller margins of error. For example, they demonstrate that even for variables like childhood poverty, it is possible to produce usable estimates for the city of Chicago by intelligently combining census geographies into new "regions". This strategy results in some loss of geographic detail, but the loss is minimized by ensuring that only similar and proximal geographies are merged together

7 Conclusion

Little (2012) argues that a fundamental philosophical shift is necessary within both federal statistical agencies and among data users, "we should see the traditional survey as one of an array of data sources, including administrative records, and other information gleaned from cyberspace. Tying this information together to yield cost-effective and reliable estimates…" However, Little also notes that for the Census "combining information from a variety of data sources is attractive in principle, but difficult in practice" (Little 2012, p. 309). By understanding the causes of uncertainty in the ACS the implications of Little's statement become clear, there is enormous potential to mash-up multiple forms of information to provide a more detailed picture of US cities.

However, there are major barriers to incorporating non-traditional forms of data into official neighborhood statistics. The reasons for this range from organizational to technical. Institutionally, there is a resistance to barriers to the adoption of non-standard forms of data in public statistics. This resistance stems from the fact such data sources are outside of the control of the agencies producing the estimates are relying on such data, that may be subject to changes in quality and availability, poses a problem for the tight production schedules faced by national statistical agencies. Technically, it is often unclear how to best leverage such information, while we have outlined some possibilities they are difficult to test given the sensitive and protected nature of census/survey data itself. Very few people have access to this protected data, it is protected by statute, and thus must be handled in very cumbersome secure computing environments. This makes it difficult to "prove" or "test" concepts. In the US and UK there are some efforts underway to publish synthetic data to allow research on/with highly detailed micro data without releasing the data itself. The barriers to innovative data fusion are unlikely to be resolved and until clear and compelling examples are developed that push national statistical agencies away from their current practices.

To summarize, the growing enthusiasm over big data makes it easy to disregard the decline of traditional forms of public statistics. As these data decline in quality it becomes difficult to plan, provide services, or understand changes in cities. The enthusiasm over big data should be tempered by a holistic view of the current data economy. While it is true that many new data systems have come online in the last 10 years, it is also true that many critical public data sources are withering. Is big data a substitute for the carefully constructed, nationally representative, high resolution census data that many practicing planners and policymakers rely upon? I think not, and while federal budgets are unlikely to change enough to yield a change to the quality of federal statistical programs, the use of new forms of data to improve old forms of data is a promising avenue for investigation.

References

Alexander CH (2002) Still rolling: Leslie Kish's "rolling samples" and the American Community Survey. Surv Methodol 28(1):35–42

Anderson MJ, Citro C, Salvo JJ (2011) Encyclopedia of the US Census: from the Constitution to the American Community Survey. CQ Press, Washington, DC

Fay RE, Train GF (1995) Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In Proceedings of the Government Statistics Section, American Statistical Association, pp 154–159

Fuller WA (2011) Sampling statistics. Wiley, Hoboken, NJ

Against the smart city (The city is here for you to use) by Adam Greenfield Kindle Edition, 152 pages, 2013

National Research Council (2013) Nonresponse in social science surveys: a research agenda. In: Tourangeau R, Plewes TJ (eds) Panel on a research agenda for the future of social science data collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC

Kish L (1990) Rolling samples and censuses. Surv Methodol 16(1):63–79

Kish L (2002) Combining multipopulation statistics. J Stat Plan Inference 102(1):109–118

Kitchin (2014) Big Data & Society 1(1)2053951714528481; DOI: 10.1177/2053951714528481

Little RJ (2012) Calibrated Bayes: an alternative inferential paradigm for official statistics. J Off Stat 28(3):309–372

MacDonald H (2006) The American community survey: warmer (more current), but fuzzier (less precise) than the decennial census. J Am Plan Assoc 72(4):491–503

Navarro F (2012) An introduction to ACS statistical methods and lessons learned. Measuring people in place conference, Boulder, CO. http://www.colorado.edu/ibs/cupc/workshops/measuring_people_in_place/themes/theme1/asiala.pdf. Accessed 30 Dec 2012

Porter AT, Holan SH, Wikle CK, Cressie N (2014) Spatial Fay-Herriot models for small area estimation with functional covariates. Spat Stat 10:27–42

Rao JNK (2003) Small area estimation, vol 327. Wiley-Interscience, New York

Särndal C-E (1992) Model assisted survey sampling. Springer Science & Business Media, New York

Spielman SE, Folch DC (2015) Reducing uncertainty in the American Community Survey through data-driven regionalization. PLoS One 10(2):e0115626

Starsinic M (2005) American Community Survey: improving reliability for small area estimates. In Proceedings of the 2005 Joint Statistical Meetings on CD-ROM, pp 3592–3599

Starsinic M, Tersine A (2007) Analysis of variance estimates from American Community Survey multiyear estimates. In: Proceedings of the section on survey research methods. American Statistical Association, Alexandria, VA, pp 3011–3017

U.S. Census Bureau (2009a) Design and methodology. American Community Survey. U.S. Government Printing Office, Washington, DC