Uncertainty, Misinterpretation and Misuse of U.S. Geo-demographic Attribute Data.

Dr. Mark R. Leipnik. Dept of Geography, Sam Houston State University.
Sanjay S. Mehta. Department of Marketing, Sam Houston State University.

Abstract

Geospatial data stands on two figurative legs: *geographic data* and related *attribute data*. In this paper, the accuracy, completeness and appropriateness of the *attribute data* containing responses to questions on forms for the U.S. decennial census and American Community Survey will be discussed. Specifically, issues of under-counts, mobile populations, falsification and misinterpretation of instruments, issues of privacy and handling of low population areas and analysis of special groups such as gays and expatriates will be featured. The potential and actual misuse and misinterpretation of TIGER data in marketing studies along with potential solutions available to cope with errors and reduce uncertainty in the use and analysis geo-demographic attribute data is a major emphasis of this paper.

Introduction.

The availability of detailed geodemographic data in a format compatible with advanced GIS software such as ArcGIS, on residents of the United States of America who responded to the decennial censuses of 1980, 1990 and 2000 has been of incalculable benefit to market researchers, social scientists, and local and national government officials among others. However, the U.S. Census Bureau's TIGER data has many limitations and issues of accuracy and completeness associated with it. Some are derived from the rather primitive data structure chosen to manage the geographic vector data contained in TIGER line and boundary files. Also the locational accuracy and completeness of geographic TIGER data is open to serious question. The attribute data linked to TIGER data is also frequently in error, most commonly due to missing address ranges, misapplied street names or misspellings of street names. Many efforts are underway to correct these deficiencies. The attribute data derived from the census questionnaires themselves is the focus of this paper. Basically there are several issues related to the attribute data. There is uncertainty concerning the accuracy and completeness of the data. There is the potential for misinterpretation and reinterpretation of the data and finally there is the potential for misuse of information derived from the census, most notably by persons engaged in unwanted sales and marketing efforts, but perhaps more ominously by criminals or even terrorists.

Sources of Uncertainty

Of all the multiple sources of uncertainty related to TIGER demographic data, one of the most significant is the likelihood of undercounts of various groups in the society, notably undocumented workers/illegal aliens. Also migrant workers, expatriates, college age singles, ethnic groups like the Roma (Gypsies) and the homeless. Many of these groups are on the margins of society. However, one group of migrants with average or above average incomes which is also hard to enumerate accurately are the "snowbirds". These peripatetic recreational vehicle driving retired Americans are a significant factor in the winter populations of many regions. Another undercounted group are migrant workers and workers in extractive industries such as oil and gas and mining, while migrant workers are often undocumented/illegal aliens particularly those working in the agricultural sector, not all such workers are low skill and/or illegal aliens. Examples of groups that are highly mobile but mostly documented resident high skill workers include oil field workers, workers in off-shore fishing, workers in mobile wheat harvesting crews and miners. In the past, miners may have been more settled, but current capital-intensive mining requires large numbers of workers in the brief early stages of construction and then a smaller on-going labor requirements, creating a highly mobile population of miners.

Undocumented workers are a large rapidly growing and increasingly diverse population that is economically crucial to many local economies as well as to the global competitiveness of the United States. However, estimates of the size of this population vary enormously. Low estimates place the number at approximately 2 million and thus probably constitute the largest block in the officially estimated 3.5 million undercount in the 2000 census. Few demographers believe the low end estimate is accurate however. More reasonable estimates such as that of demographers at Northeastern University place the number of illegal immigrants in 2000 at approximately 13.5 million, some but not all of whom would have been enumerated by the census. Most of these undocumented workers/illegal immigrants would have been missed by the census because many would because they never responded to census instruments and were not imputed to dwellings which appeared to be inhabited but where the residents failed to respond to initial and follow up attempts at contact by census enumerators. Imputed residents are counted, but assigned only as additional residents with the attributed characteristics of the other residents of the same block, block group, tract etc. But that begs the question that if the undocumented non-responsive households are similar to those from that area that do respond to the census. The resounding answer has to be no they are not. They are likely to have a higher proportion of men of working age, those that have families are likely to have more children, and they are likely to be poorer, less well educated and thus generally unrepresentative of the population as a whole. Even for those imputed populations of unresponsive undocumented workers that have been located based or identification of residential housing, the density of persons per dwelling unit and thus total numbers are likely to be higher than estimated based on national averages. This is because many are residing in unconventional housing and sharing facilities sometimes in shifts. This is true whether the housing takes the form of lofts in the Bronx full of immigrants from Fujian paying off the snakeheads who smuggled them in the country, bunk beds in industrial facilities holding Thai and Laotian women held virtual prisoner

in sweat shops in Southern California, or semi-migrant farm workers from Mexico, Central America and Haiti living in trailer "towns" on private property like those of La Perra Falba, Arizona or Belle Glade, Florida or the tens of thousands of slaughter house and confined animal feeding operation workers living 18 to a trailer and sleeping in successive shifts in communities like Guymon Oklahoma or Garden City, Kansas.

The U.S. Census Bureau (USCB) makes many commendable efforts to account for illegal/undocumented aliens. For example, an effort is made to make it clear that census responses and forms are not going to be shared with law enforcement or “La Migra”. However, it is often doubtful that these statements, even if capable of reaching the multi-lingual often semi-literate audience toward which they are direct are in fact believed. When the USCB shares information with other governmental agencies on the number of “Arabs” per zip code it calls into question the confidentiality of responses, but that is largely irrelevant for most illegal aliens who are simply going to be suspicious of any dealings with the U.S. federal government. So having materials and outreach efforts in the Spanish and other languages, having more canvassers in areas with large and growing foreign born populations and other outreach efforts are likely to have been only modestly successful at best. A secure national identity card is the likely longer term solution to getting a better handle on the magnitude of the U.S. population. This approach if adopted would also have the effect of discovering vast numbers of persons with multiple identities and not an insignificant numbers of persons who have died long ago but whose relatives are collecting disability or social security benefits fraudulently on their behalf. Unification of birth certificate, drivers license, social security, IRS, phone company, 911, U.S. Census and other databases would also have the added benefit of making identify theft more difficult and finding some of the 1 million estimated fugitives from justice (mostly child support evaders) that are estimated to be present in the U.S.

Another source of uncertainty relates to the manner in which population is attributed to residences which do not respond to initial distribution of census forms and are not available when house to house canvassing of non-responding residences is done subsequently (typically in late spring or early summer). In the procedures for tallying non-respondents there is potential for both undercounts as well as over-counts. These counts and estimates can have consequences such as the count for the State of Utah not being sufficient to get Utah rather than North Carolina an additional seat in congress. Utah argued in court filings that Mormon families (who constitute a majority of Utah residents) have on average more children than typical American families and should have a larger residential population attributed to them. Also Mormons who where out of the country on mission work were not counted as residents. Interestingly, Utah did not try to argue that there are as many as 25,000 polygamist families in the State, some of which have as many as 60 childen and thus an even higher number should be attributed to those households that do not respond to the census in Utah.

Study of the polygamist community of Colorado City, Arizona although not specifically affecting the Utah count does indicate many interesting things about the way in which unconventional families respond to the census. Interestingly, although technically the men folk are mostly law-breakers, Colorado City residents do respond

with enough information to indicate that the city has the highest family size (6.9) of any tract in Arizona in the 2000 census. It also has one of the fastest growth rates in Arizona, a fast growing state and census data indicate that few of the children over 16 are in school rather than in the work force (few go to college). What is not revealed is that the family size should be even higher than reported, that many men have five or more wives, while many other young men are absent from the community working in itinerant construction jobs often never to return. This structure allows older established men that return to the community to obtain multiple wives without having to recruit women from outside the community. How many more polygamists outside the established communities of Colorado City and the smaller community of Hildale Utah there are is difficult to establish, but polygamist groups in Western Utah, Box Elder County Utah, Western Nevada, South Western Wyoming and most recently in Schleicher County, Texas exist.

Mormon Polygamists are only one unconventional group that may be undercounted. The Branch Davidians, and Rajneesh's followers, UFO cultists, "Moonies", etc are not likely to be counted accurately and while not large groups really they can be locally important. The Amish although long settled are likely to shun census takers just as they avoid voter registration, jury service and military conscription. Certainty, the attempt to establish Rajneshpurnam in South Western Wasco County Oregon had a major demographic effect there. Interestingly enumeration or more precisely voter registration played a very key role in the story of Rajneesh and his downfall. The 18,000 acre property on which he wished to build a city of several hundred thousand residents was zoned for only 10 residences and held as many as 2,000 actual permanent residents and hosted up to 15,000 visitors during annual festivals. So an attempt was made to import homeless persons to gain population in order to gain control of county elections. When this failed, use of food poisoning and other illegal tactics were resorted to ultimately causing the downfall of the group and return to more normal population numbers in the area. The Census seems not to register any of these changes partly because they happened in out years, but also because of the secretive nature of the cult, which seemed to want registered voters but not government intrusion.

Over-counts are of smaller magnitude than undercounts but in addition to the situation where an area with many vacation homes receives attributed population that may only reside there seasonally, another potential source of over-counts is situations where a parent or guardian of a child living away from home while at school or due to divorce, is counted twice when they themselves respond to the census or where the other ex-spouse or step parent or both sets of grandparents respond. The vast and growing number of families with joint custody, step children and other non-traditional living arrangements complicates this issue greatly. Yet another potential source of effective over-counts relates to the treatment of federal workers that are working outside the country, particular those in the armed services. Unlike other expatriates who are not counted Federal employees (mostly military service members) are counted as living where they are "based". The convention is that these persons are "based" at a military facilities such as Fort Hood, Texas (the largest population base the U.S. with two army divisions stationed there) or the Marine Corps base at Camp Pendleton, California, or the Navy Base at Bellingham, Washington, or Minot Air Force Base North Dakota etc

wherever they actually happen to be deployed to. So when a detachment like the 82 Army Airborne Division, the 1st Marine Expeditionary Force, or an Air Combat Wing is deployed to Iraq, Afghanistan or Diego Garcia in the Indian Ocean or somewhere in the U.S. for a temporary Duty Assignment, the home base gets recognized as the home of the members of that detachment even if those person spend only a fraction of any year there or are gone from the base for several years at a time. With increasingly long and back to back deployments the magnitude of the over-count of effective population from this practice is growing.

Other sources of uncertainty relate to the periodicity of census taking. While many advanced countries rely on population registries to keep a current count on population and residency numbers and locations, the 10 year period between counts under the U.S. Census dates to its inception in 1790. That a society relying on leather bound multi-column books and quill pens to ascribed numbers attested to by a family patriarch and then recording name, age, number of children, servants and slaves is now obsolete is clear, but so is the 10 year separation between data collection efforts that went with the 1790 census. Although many countries have followed the U.S. lead by adopting use of a ten year interval, others like Sweden use a shorter interval and many use population registries to keep continuously count on numbers supplemented with a census for gathering more detailed information. Today families in the U.S. move on average every 4 to 5 years, with many moves being out of place, tract and even state, and that coupled with the lag time of 1 to 3 years for release of various components of the census data, makes the likelihood that population estimates and imputed demographic characteristics being in error far higher. That the census count is used to redistrict political jurisdictions means that increasing the frequency of the count is highly problematic. One solution is to use the Amercan Community Survey (ACS). The ACS is a roving snapshot of selected communities done continuously in the intervening years to supplement the understanding of changes taking place in the out years. However, the ACS just gives the barest flavor of changes, because it can only cover a tiny faction of relevant tracts and the program lacks the resources to do entire large urban areas or extensive rural areas. Hence, we are likely to only get a look at a few mid-sized nicely contained communities that are far from typical of either major urban or disparate rural areas.

Population estimates for boom towns that happen to either boom or bust during the time of the census enumeration will lead to erroneous results in the count and character of population being enumerated. Boom towns related to extractive industries such as gold mining are particularly likely to suffer actual population fluctuations. However, these may not be captured. One might have a boom in 1992 and a bust in 1996 in the price of gold and related booms in populations in Ely and Elko, Nevada and areas in South Dakota, and miss both the boom and the bust with data for 1990 and 2000. In fact something like this did happen. However, even if the boom was in 2000, one might miss it with census taking methods. Many miners are highly mobile, living in temporary trailer parks and sharing beds in mobile homes. They are unlikely to consider these boom towns as a permanent home and are unlikely to respond to census instruments. Even if a

correct count of mobile homes was made, the number of inhabitants and the characteristics attributed to them would be likely in error.

Misinterpretation

Even when there is a fairly complete response to either the long or short form from an enumerated person, a range of issues arises with respect to the accuracy of the responses to the survey instrument. Some of these errors stem from misinterpretation of the instrument. For example, interpretation of the meaning of the term Hispanic or Cajun or the differences among and between race, ethnicity and nationality can all cause confusion and miscoding of responses. Other census questions may elicit deliberately or accidentally inaccurate information; such as under reporting estimated income, or over estimating housing values. Some questions may be deemed offensive to some individuals, such as those regarding marital status, parenting, out of wedlock or mixed race children, living in unconventional households (such as those with multiple wives), etc.

In addition to errors and omissions from census responses, there are a wide range of misinterpretations that can befuddle users of U.S. geodemographic data. Common examples of misinterpretations by users of TIGER data are the assumption that data for 2000 is valid in the out years through 2012 or so whenever newer data should become available. Also the handling of the institutionalized population in census data causes numerous misinterpretations to occur. Finally there is a common tendency to confound terms in the census such as family with household and median family income with personal income. Many of the misinterpretations are easier to rectify than the uncertainty related to incompleteness. For example properly taking into account institutionalized populations can help users adjust estimates like trade area market potential to not include as potential customers those persons that are or were incarcerated.

One of the most accurate of all counts is that of institutionalized persons. It does not rely on the choice of respondents or attribution of population to non-responding residences. Rather, a bureaucrat with the U.S. Bureau of Prisons, Texas Department of Criminal Justice or Los Angeles County Sheriff's office will supply information on the number of incarcerated persons, age gender, race, etc. For those with long term sentences there is no question where they are residing at the time of the report and perhaps until the next census. What is a little murky are estimating the number also claimed by relatives and elimination of double counts. Another issue is related to the characteristics of those prisoners. Thus age, gender and income (often zero) will be reported accurately, but race will be reported on behalf of the prisoner. For a category like African American that is likely to be reasonably accurate, but "Hispanics" may well not be reported accurately. It has been noted that Puerto Ricans in New York are not reported as such or as "other Hispanics" but rather as whites by the prison system there. While one cannot quibble with someone who chooses to identify himself as "Hispanic" or as any other race or nationality on a census form, when a bureaucrat makes the judgment of ethnicity, one can argue that some objective measure of the accuracy of the assignment might be needed. It

makes a difference to corrections and community supervision efforts if 10% or 40% of New York Prisoners are “Hispanic” or if it is not possible to differentiate among Cuban. Puerto Rican, Salvadorian and Mexican prisoners because they have all been for ease of classification been termed “white”. It is worth noting that many officials in law enforcement prefer to class Hispanics as white if there is no down side to doing so since it helps minimize charges of racial profiling or disproportionate imprisonment as well as saving officials from understanding the complexities of language, skin color and racial categories not understood outside the home country which account for identities that are recognized among many Latin American populations.

Misuses

Related to misinterpretations are cases where the census data has been used to infer information which was not directly solicited in the census. An example of the later would be generation of estimates of gay and lesbian populations based on counts of same sex couples living together in the same household, but attempts to use geodemographic data in studies of healthiness, political affiliation and religiosity as well as sexual orientation have been made. While these are not exactly misuses of the data, in any situation where some characteristics such as sexual orientation is inferred from an instrument not designed to illicit such information, the conclusions reached are likely to be questionable and may result in misinterpretation.

The last category of issues related to use of census data is the actual misuse of the data. Since those engaged in deliberate misuse of the data are not generally forth coming as to their activities, this section must be the most brief and speculative. However, it is clear that there are several characteristics of census data collection and dissemination methods that lead it to being susceptible to misuse. Most notable is the existence of a vast number of low population blocks, block groups, and tracts. The advent of geodemographic data on the internet coupled with the existence of numerous low population census subdivisions renders the privacy protections that the USCB seeks to insure less than totally effective. While some of the problems of low populations and “exposed” minorities are unavoidable given the diverse distribution of U.S. population, it is certainly feasible using newer geo-database structures to “shield” low population subdivisions from release of data. Thus for blocks perhaps restriction of data for subdivisions having less than 50 families (~150 persons) would be a good approach. Low population blocks, block groups, tracts, is a problem likely to grow worse over time as population growth occurs in areas that currently have zero population. The internet allows anyone with a browser to idenbtify characteristics of individual families in these low population areas. Thus using only the internet persons with bad intent could find the often rather small blocks in cities and towns with a single mixed race couple as residents. That racists have preyed with particular virulence on such persons in the past is indisputable. The USCB’s own data only allows identification of the block and possibly the street name and address range of these families. But the independent advent of high

resolution plan view aerial photography on sites such as Teraserver allows the entry of a street name and county or city name and state and the rapid provision of 1 meter plan view aerial photography usually less than 10 years out of date of the family's home. Thus not only can a person of a given race or mixed race be individually identified but in many cases the home in which they reside can be viewed from above. This invasion of perceived privacy is will be made worse by the coming advent of oblique aerial photography at even higher resolution in the near future.

That geodemographic data is already being misused by direct marketers is undeniable, although it is usually only a part of a wide cross section of data that is available from credit agencies and data compilers. Recent scandals involving loss of the personal information for hundreds of thousands of Americans involving such established companies as Choice Point, Bank of America and other firms highlights the issue of identity theft and privacy of geospatial data. That data provided freely without compensation to the Census Bureau can also compromise privacy is an issue that will only grow more troublesome. In the 2000 census enumeration effort, there where some inklings of organized residence to response to the census and some well publicized cases of outright refusal to respond. This refusal is actually punishable by imposition of a modest fine. It is likely that for every case of publiciced resistance, there were thousands of individuals who felt that the Census Bureau had no right to probe into whether they had indoor plumbing, were a grandparent caring for a minor child, were an unmarried mother or a mixed race couple. If persons begin to misuse census data and that misuse becomes widely publicized, the resistance and aversion to providing personal details to a faceless governmental entity that then proceeds to broadcast that information globally to anyone that can navigate to the American Fact Finder Web site will no doubt grow.

To prevent such an outcome, the USCB needs to be more concerned about the effect of low population subdivisions on confidentiality. In fact, the solution to this problem is now available in the form of a more sophisticated and intelligent data structure for TIGER, one that would take advantage of the object-oriented geodatabase format built into ArcGIS. A more appropriate approach would automatically re-aggregate low population census subdivisions into larger encompassing areas, suppressing release of the data for those areas with low populations per se and lumping them together with adjacent and surrounding subdivisions that did have adequate population to protect the privacy of responses. This approach may have its problems, due to existence of low population, places and counties and even whole states with low populations of certain groups. Thus to reveal anything about the demographic characteristics of a near or below 100 population county like Loving County Texas compromises privacy of respondents, To reveal the location or even number of African Americans present in Sublette County Wyoming (Population 3 African Americans in 1990 and 7 African Americans in 2000) also runs the risk of compromising privacy. Even revealing the number of native Hawaiians in some state such as North Dakota could be a source of concern. Thus for a block with a single family, an outside user of census data would only learn that no data was available on that block, but that that block along with an adjacent block housed 57 families, that 50 African American families resided in any of 10 adjacent block groups, that ten counties in the Texas panhandle had a total of 97 mixed race families and that the

states of North Dakota, South Dakota and Minnesota and Wisconsin together held 57 native Hawaiians. Another alternative would be to sever the street centerline structure from the block boundaries and redraw all blocks to minimize the number of blocks that have low or zero populations (zero populations pose an issue of becoming low population blocks etc in the next census). This solution is also one that takes advantage of the ability of GIS to manage multiple layers and re-aggregate data among and between subdivision using spatial joins and other techniques poorly developed when the now obsolete TIGER data structure was initially developed.

Conclusion.

The advent of the U.S. Census Bureau's TIGER set of geodemographic data was a ground breaking event in the evolution of national geospatial infrastructure in the U.S. and also a source of inspiration for development of similar data sets in other countries, notably Australia and Canada which have followed very similar models. However, TIGER has various issues that continue to cause problems. These issues include those related to missing or erroneous data, inaccurate or deliberately misleading responses, misinterpretation by a growing number of ever more casual users, reinterpretation by various groups bent on teasing out information such as lifestyle segmentation profiles, or sexual orientation information and last but not least future responses to TIGER are endangered by the growing potential of the data to impinge of the perceived privacy of respondents due to the problem of small population blocks, tracts, etc.

Solutions to all of these issues exist, some as simple as using more recent adjusted population figures based on projections or the ACS or properly accounting for institutionalized population and others that would take advantage of the growing sophistication of geospatial data structures available in commercial off the shelf software packages like ArcGIS. That the more sophisticated solutions are unlikely to be adopted due to budgetary limitations is unfortunate, but it will be unlikely to stem the growing use by market researchers, social scientists, municipal officials and increasingly the general public of TIGER data nor its concomitant misuse. One can only hope that through educational efforts and improvements in content and databases tools, that the proportion of misuse, misinterpretation and the proportion of erroneous data will fall over coming years…

Acknowledgements.

Dr. Leipnik would like to acknowledge the assistance of Jennifer Lorca a the University of Texas at Austin, Dr. William Bosworth at Lehman College in the Bronx, New York, Peter Wagner at The Prison Policy Initiative.

Author Information

Mark R. Leipnik, Associate Professor, Geography Department. Sam Houston State University, 1900 Ave I. Huntsville, Texas 77341. Phone (936) 294-3698, Fax (936) 294 3940, email: leipnik@shsu.edu

Sanjay S. Mehta. Associate Professor, Marketing Department. Sam Houston State University, Ave J I. Huntsville, Texas 77341. Phone (936) 294-1312. email: mehta@shsu.edu