UNCERTAINTY OF US CORE GEO-DEMOGRAPHIC DATA

TIGER Tale

Core geo-demographic data in a GIS-compatible format derived from decennial US censuses has greatly benefited marketers, social scientists and governments. It has also inspired development of similar datasets in other countries. However, the US Census Bureauâ€™s Topologically Integrated Geographic Encoding and Referencing (TIGER) data has many limitations, including missing or erroneous data, inaccurate or deliberately misleading responses, misinterpretation by casual users, and reinterpretation by various groups bent on teasing out information. There is also potential for criminal misuse of derived information.

One of the biggest sources of uncertainty concerns undercounting of various groups, many but not all socially marginalised. So-called â€˜snowbirdsâ€™, retired Americans with above-average incomes, in many regions representing a large proportion of the winter population are, for example, also hard to enumerate. Another undercounted group comprises undocumented migrant workers in the agricultural sector and extractive industries such as oil, gas and mining, not all low-skilled and/or illegal aliens.

Biased Characteristics

Undocumented workers are a large, rapidly growing and increasingly diverse population; todayâ€™s estimates vary from 2 to 13.5 million, a group economically crucial to the global competitiveness of the US. Most will have been missed by the census due to their never having responded to census instruments and not being imputed to dwellings that appeared inhabited during enumeration. When imputed residences are counted such people are assigned as additional inhabitants with the attributed characteristics of other residents of the block. While this may be appropriate for the general population, undocumented workers tend to comprise a higher proportion of men (sharing facilities and working shifts) and the families may have more children and be poorer and less educated.

Identity Theft

The US Census Bureau (USCB) makes many efforts to account for illegal aliens, for example by making clear that census responses and forms are not going to be shared with law enforcement agencies. However, after the 9/11/2001 attacks the USCB did share information with other government agencies on the number of â€˜Arabsâ€™ per zip code. Having materials and outreach efforts in multiple languages would have modest success. A secure national identity card would be the longer-term solution to gaining a better handle on population size. This approach, if adopted, would also have the effect of uncovering vast numbers of persons with multiple identities and persons who have long since died but whose relatives are fraudulently collecting benefits. Unification of birth certificate, driverâ€™s licence, social security, phone company, US Census and other databases would also have the benefit of making identify theft more difficult and finding some of the estimated one million fugitives.

Shunning Census

Undercounts and over-counts may occur during procedures for tallying non-respondents to the initial distribution of census forms. The State of Utah, for example, lost a congressional seat by being undercounted. Utah argued in court filings that Mormon families, who constitute a majority of Utah residents, have on average more children than typical Ameri-
can families and should have a larger residential population imputed to them. Mormons who were out of the country on missionary work were not counted either. Interestingly, Utah did not try to argue that the State held as many as ten thousand polygamist families. Other unconventional groups, such as the branch Davidians, followers of Rajneesh, and the Amish are also unlikely to be accurately counted. The Amish, although long settled, are likely to shun census takers just as they avoid voter registration, jury service and military conscription. Certainly the attempt to establish Rajneshpurnam in 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 ten residences and held as many as two thousand actual permanent residents, hosting up to fifteen thousand visitors during annual festivals. So an, ultimately failed, attempt was made to win control of county elections.

Wrong Counts

Over-counts, which occur less than undercounts, may appear in areas with many seasonally occupied holiday homes and in situations where a child away at school is counted twice, they themselves responding to the census and others at home responding on their behalf. Further, federal workers working outside the country may not be treated as ordinary uncounted expatriates but rather as living where they are â€˜basedâ€™; so the home-base community gains added population even if those persons spend only a few days there annually or are gone for years at a time. Most analysts simply count popu-
lation. However, in some areas the presence of institutions such as prisons may cause a biased count. Huntsville in Texas, for example, has a high population of â€˜Hispanicsâ€™ (Figure 1), but here the highest concentrations are associated with five state prisons, a county jail and a university dormitory. Six thousand of the 35,000 inhabitants are prisoners, while Walker County, of which Huntsville is the county seat, has approximately thirteen thousand prisoners in a population of slightly over sixty thousand. In some American counties more than half the population is composed of prisoners, a fact that unless taken into account distorts many aspects of the use of geo-demographic data.

Periodicity

The ten-year period between counts in the US Census dates back to its inception in 1790. Today families in the US move on average every four to five years, often out of state. Since most census data becomes available one to three years after enumeration, it is most likely that characteristics and estimates of population will be inaccurate. Both yearly and future estimates make use of models involving birth and death rates for various age cohorts, and migration and immigration data. One solution is the American Community Survey (ACS). While today lacking funding, the ACS is a roving snapshot of selected communities taken in the intervening years to supplement census finding with understanding of changes taking place in â€˜outâ€™ years.

Municipal Boundaries

US Census estimates of population within a municipality depend on local authorities updating the boundaries in the TIGER files as growth and annexations occur. However, in most cases municipal boundaries are unrelated to census enumeration area boundaries. One example of a resulting error is shown in Figure 2: the heavy (red) line represents what the US Census believed to be the municipal boundary for the city of Conroe in Texas as it entered the 2000 census. In fact, at that time the correct boundary was represented by the larger shaded area. Not only was this larger, containing many thousands of additional residents as a result of population increase from approximately 27,000 to approximately 42,000 between 1990 and 2000, but the TIGER data also reflected areas incorporated into the city in 1990 and not in 2000.

Misinterpretation

Accuracy of response is often affected by misinterpretation of census questions. Generally an official will supply information on any institutionalised person, such as a prisoner. Thus age, gender and income (often zero) may be accurately reported but race will be reported on behalf of the prisoner. In the case of a bureaucrat making a judgement of ethnicity some objective measure of the accuracy of this assignment might be needed. It makes a difference to corrections and community supervision efforts whether 10% or 40% of New York prisoners are â€˜Hispanicâ€™. Erroneous answers may be given deliberately or by accident, such as under-reporting of income or over-estimating house value. Some individuals may interpret some questions as being offensive, such as those on marital status, born out of wedlock or mixed-race children, and living in unconventional households, such as those with multiple wives. Users may also misinterpret data, such as assuming that data for 2000 is valid for the â€˜outâ€™ years through to 2012, handling of the institutionalised population such as prisoners, and by confounding terms such as â€˜median family incomeâ€™ with â€˜personal incomeâ€™ and â€˜householdâ€™ with â€˜familyâ€™.

Data Misuse

Geodemographic data is often used in studies for which the instrument of census-data collection and dissemination has not been designed, and this may lead to misuse. An example is estimation of homosexual population from counts of same-sex couples living together in the same household. Further, blocks, block-groups and tracts are in TIGER typically bounded by street segments representing centrelines for roads or motorways or, occasionally, railways or water features that in the single-layer, obsolete TIGER data structure form the boundaries of the smallest enumeration area. Block boundaries can vary tremendously and many blocks may have low population or even none at all (Figure 3). However, basic demographic information such as race, age and gender is available down to block level, which may compromise privacy and confidentiality, areas which the USCB seeks to safeguard. This is especially true when data is published on the internet, which allows anyone with a browser to identify characteristics of individual families; it is thus possible to search for blocks in cities containing a single, mixed-race couple. What is more, internet publication may endanger future census response as people sense impingement upon their privacy. In the 2000 census some outright refusals to respond received widespread publicity, and one publicised instance of resistance might represent thousands of individ-
uals who felt that the USCB had no right to probe into their personal lives. Restriction of data for subdivisions having less than fifty families (~150 persons) would be a good approach. While some of the problems of low population and â€˜exposedâ€™ minorities are unavoidable, newer geo-database structures could be used to â€˜shieldâ€™ low population subdivisions from release of data. A more appropriate approach would automatically re-aggregate certain subdivisions into larger encompassing areas, suppressing release of data concerning those areas with low population per se and lumping them together with adjacent and surrounding subdivisions of populations adequate to serve to protect the privacy of respondents.

Disappearing Cajuns

The census â€˜long formâ€™, randomly administered to one in six respondents for the decennial census, includes questions on the ethnicity and national origin of respond-ents. One group whose numbers have drastically changed due to the way this question is posed are Cajuns. Cajuns are the descendants of French Canadians and residents of Santo Domingo who came to Louisiana in the eight-eenth century as refugees. They are concentrated in southern Louisiana, as Figure 4 shows. In 1990 the census included â€˜Cajunâ€™ as an example of a possible nationality, while in 2000 the term was missing and anyone responding by writing in Cajun was lumped together with other French Canadians, a group concentrated in the north-east of the US. The result was the apparent disappearance of over 400,000 â€˜Cajunsâ€™ between 1990 and 2000. Related problems are caused by confusion over what the term â€˜Hispanicâ€™ means, and also by persons who report an ethnic origin that does not relate to an actual nationality, such as someone from Turkey reporting himself a â€˜Kurdâ€™.

Concluding Remarks

TIGER has various issues that continue to cause problems. Solutions exist, some as simple as using population estimates or correctly accounting for institutionalised persons and others that would take advantage of the growing sophistication of geospatial data structures available in commercial off-the-shelf GIS software packages. One can only hope that through educational efforts and improvements in content and database tools the potential for misuse and misinterpretation, and the proportion of erroneous data, will decline in the future.

Acknowledgements

Thanks are due to Jennifer Lorca, Dr William Bosworth and Peter Wagner.

Further Reading

Â· Leipnik, Mark R. and Sanjay S. Mehta (2005), Geographic Information Systems (GIS) in E-Marketing, in Advances in Electronic Marketing, Idea Group Publishing, editors Irvine Clarke III and Theresa B. Flaherty, Chapter XI, pp 193-209.

Â· Peters, Alan and Heather MacDonald (2004), Unlocking the Census with GIS, ERSI Press Redlands CA, ISBN 1-58948-113-5.

Â· Leipnik, Mark R., Sanjay S. Mehta, and Balasundram Maniam (2003), An Introduction to TIGER 2000, Marketing Management Association Spring Educators Conference, Chicago IL, March 12-14.

Â· Brewer Cynthia and Trudy Suchan (2001), Mapping Census 2000: The Geography of U.S. Diversity, ESRI Press Redlands CA, ISBN 1-58498-014-7.

Â· Mehta, Sanjay S., Leipnik, Mark R. and Balasundram Maniam (1999), Application of GIS in Small and Medium Enterprises, Journal of Business and Entrepreneurship, Vol. 11 No. 2, pp 77-88.