UNITED NATIONS SECRETARIAT
Department of Economic and Social Affairs
Statistics Division

ESA/STAT/AC.115/9
May 2007
English only

United Nations Expert Group Meeting on
Contemporary Practices in Census Mapping and
Use of Geographical Information Systems
29 May-1 June 2007
United Nations, New York

Geocoding: concepts and approaches to data collection*

Prepared by

Lisa Jordan
Department of Geography and Center for Demography and Population Health
Florida State University

* This document is being reproduced without formal editing.

Geocoding: Concepts and Approaches to Data Collection

Introduction

The use of Geographic Information Systems (GIS) to collect, study, and release census data serves to improve planning, public health, and public administration and management. Important GIS applications for census bureaus include geocoding address records, spatial analysis and forecasting of census information, and supporting dissemination of census data, particularly with spatial queries of census data using internet mapping interfaces. *Geocoding* describes the process by which addresses are converted to geographic coordinates, such as longitude and latitude, and is the focus of this paper.

One project of census geography entails geocoding to assist in the distribution of questionnaires (or surveys) and collection of responses, as well as using geocoded responses for the dissemination of census information for geographic regions that are useful for decision-making. This involves linking digital address files to geographic files that document the spatial and address characteristics of road segments. Often, the geographic coordinates are estimated, either based on an interpolation of where an address is likely to be situated along a road segment, or with other simplifying assumptions, such as the center or intersection of the street. While geocoding can be done with GPS, the address and line files can reduce the time devoted to geocoding (Karimi, Durcik, and Rasdorf 2004).

Geocoding is particularly important for introducing more fine grained geographic resolution to census information, wherever Census forms are mailed to the population. Geocoding census data also involves the aggregation of population information to geographic areas larger than households. This paper describes the theory behind geocoding, as well as applications, limitations, and future developments of geocoding.

Theory (Place Hierarchies)

In many countries, the size and organization of the state necessitates sub-sets or hierarchies of administrative units. These internal political boundaries within countries have important relevance for census geography. Countries can be divided into regions, states (provinces, cantons, oblasts, etc.), and intra-state regions (such as counties, communes, etc.). Below these larger areal units, countries may be interested in collecting information at a finer resolution. In the U.S., tracts or block groups are examples of small areas for which census data are aggregated. These areas are additive in the sense that block groups are subsets of tracts and tracts are subsets of counties. In some countries, such as Sweden and Finland, census data are aggregated to a grid, so changes

in political boundaries (such as county bounds), do not affect aggregation values from previous years (Martin 2006).

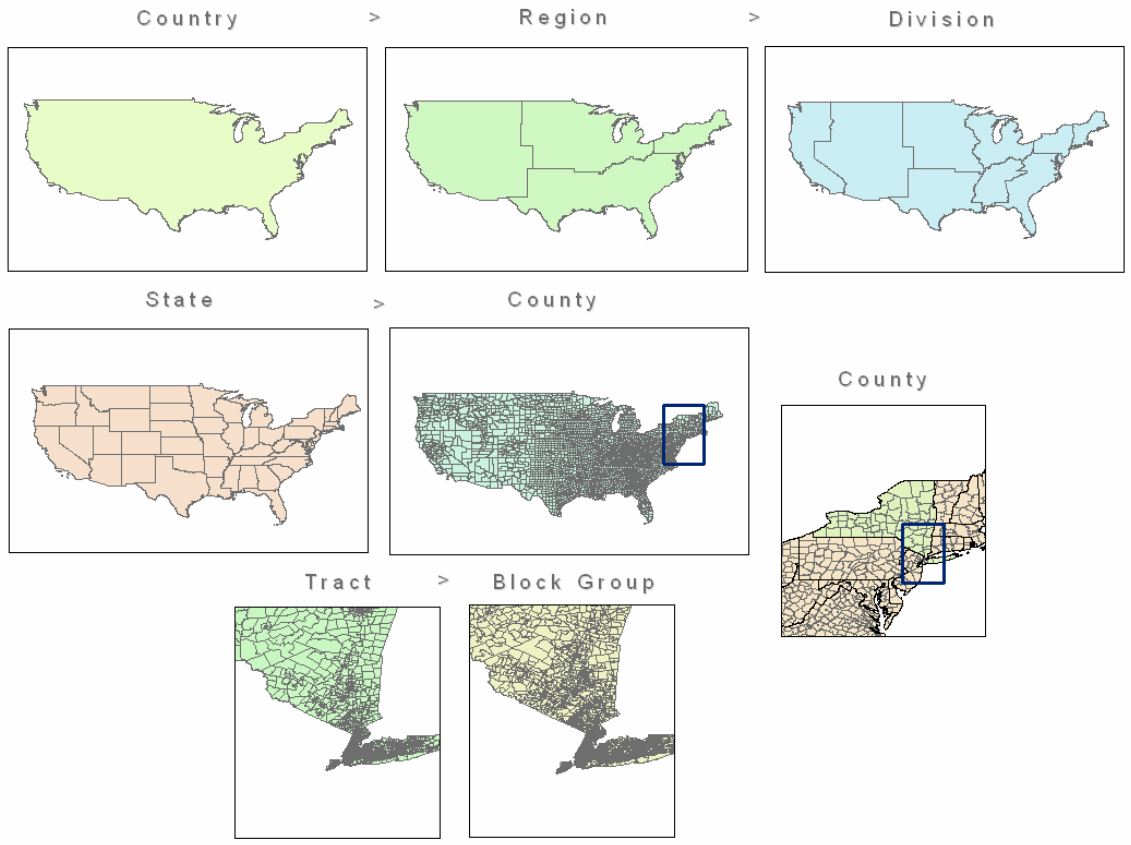

Figure 1a. Example of census hierarchies in the United States

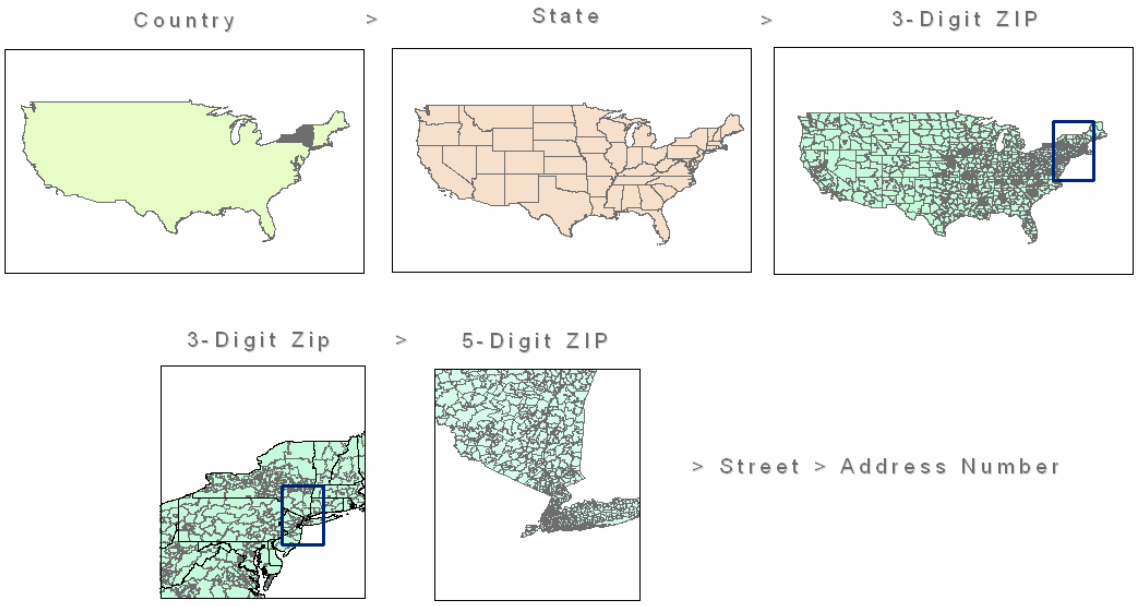

Figure 1b. Example of address hierarchies in the United States

Address information, for geocoding hierarchies, often does not match census hierarchies. An illustration of the differences between census and address hierarchies is shown in Figure 1. After geocoding by latitude and longitude, the census records may then be aggregated to any of the census geographies using a GIS. The first step in geocoding is to link address files with the address lists contained in geographic files. This involves parsing (separating street names from address numbers) and standardizing (changing all variants of address appendages, Street, St, ST, etc. to a standard). Beginning in the 1960s, the U.S. Census Bureau started address matching, and included phonetic translations of street names in the reference files to improve address matching (SOUNDEX) (O'Reagan and Saalfeld 1987); other criteria, including the extent of overlapping letters, are also used in geocoding (Davis and Fonseca 2007).

Reference places (such as monuments, statues, local names) and indirect references (such as telephone numbers or highway exit numbers) may also be used to supplement information from a postal code for address matching (Davis and Fonseca 2007). This is of particular relevance in emerging countries. Once address matching is complete, the process of assigning geographic coordinates to each address follows. One typical scheme is address interpolation, where address values are distributed evenly along a road segment. To do this, geographic reference files or roads files need to include street segments that are identified by name *and* the range of street addresses for each side of the street. Whenever this information is not included, addresses may be assigned a latitude and longitude for the street intersection or the center of the street segment. Clearly, the quality of the geographic reference file and the interpolation method influences the quality of the geocoding results (Karimi, Durcik, and Rasdorf 2004). As Karimi, Durcik, and Rasdorf (2004) report, the quality of a reference dataset is determined by its completeness, correctness, consistency, currency, and spatial accuracy (171).

The development of the geographic reference file types DIME and TIGER has made a notable contribution to census mapping. DIME (Dual Independent Map Encoding), used initially in 1968, included information on street intersections and directions (O'Reagan and Saalfeld 1987). Also, TIGER (topologically integrated geographic encoding and referencing) files are a prime example of a geographic reference database (Bolstad 2005). However, due to incompleteness and inaccuracies, the 2000 Census relied on traditional collection methods, using TIGER files to supplement census-taking, rather than having a GIS-driven census (Martin 2006). It is worth noting that for aggregation and display of census records, when it is not possible to assign a geographic coordinate to a particular record, it is usually still possible to assign a geographic code, which is associated with an area (possibly tract, county, and state). The smaller the area that is defined by the geocode, the more flexible the results will be for subsequent uses.

Applications

Geographic hierarchies used by the census are important for (1) the dissemination of information and (2) aggregation of data to protect privacy (Martin 2006). Local census information is highly valuable for: planning, public health, public finance, evaluation and assessment of policy, transportation management, emergency management, and for research. Furthermore, local census information helps to build national and international spatial data infrastructure (Martin 2006). For example, geographic information on streets serves to verify and validate remote sensing imagery. Streets databases are useful to government (local, national, and international) and business.

In addition to use in decennial censuses, geocoding allows for spatial sampling, which is more representative of the population in a given area (Kumar 2007; Longley et al. 2005). Further, internet mapping of geocoded census information serves government, businesses, and citizens. It allows anyone, with the use of a computer, to survey the demographic landscape of an area.[1] Public and private organizations can target aid to particular populations with relative disadvantage or require investments in educational infrastructure. It also creates a quick reference tool for responding to emergencies, where many people are affected.

Geocoding plays an important role in public health and epidemiology literature, particularly in the US (Yang et al. 2004; Shi 2007; Ziao et al. 2007). Geocoding is valuable for linking addresses, or point information, to neighborhood characteristics defined by the aggregation areas for which the census has released information (Ziao et al. 2007). Geocoding has helped in the development of health surveillance systems used to update medical cases in real time, and to retrieve time and location specific queries for a large number of diseases (Grigg et al. 2006). Such systems provide the information necessarily for an immediate response to contain outbreaks. Medical and public health professionals can also quickly survey and visualize a large amount of spatial and temporal information to arrive at conclusions about at-risk areas that need attention.

Limitations

Despite the usefulness and advantages of geocoding, there are a number of limitations to its use. First, geocoding often involves a heavy investment in the creation of address and geographic reference files. These files may need to be generated for the first time, an expense which may be prohibitive. Furthermore, new development and changes in the street-system are on-going, so address files and street files must be

[1] For an excellent example, see gCensus, developed by Imran Haque (http://gecensus.stanford.edu/). This web-software allows users to choose any place in the U.S., download data on any Census question, and to create a map that can be viewed in Google Earth. By collecting the data and releasing it to the public, creative applications have become public domain, even though they were not generated directly by the Census Bureau.

constantly updated. The quality of geographic reference files may vary significantly, being poorer in more remote or rural areas.

Second, geocoding techniques that were developed in a North American context are built on a number of potentially problematic assumptions that limit their applicability to other countries and regions. In particular, geocoding techniques assume the presence of named streets that follow a linear address scheme, allowing for address interpolation. Some locations defy these interpolation schemes because streets or roads do not have names, are informal (dirt roads), or do not follow a linear address scheme. In Japan, for example, street addresses are not ordered, but are based on other criteria such as year of construction (Longley et al. 2005). Additionally, undercounted populations, particularly people in shanty towns and transient populations, often evade the census altogether. Geocoding operates best when individual households are uniquely defined and relatively static, so communal living arrangements, non-traditional living arrangements, and informal housing fit less neatly into address matching or geographic referencing.

The Geocoding Certainty Indicator (GCI) may be used to filter out information below a certain quality threshold, or the GCI may be used to document uncertainty in the spatial analysis process (Davis and Fonseca 2007). The GCI provides a score based on discrepancies throughout the geocoding process, where a lower score is received for an imperfect match due to misspellings or where the address information is incomplete. Documentation of uncertainty is valuable for understanding ways to learn from and improve the geocoding process.

Future Developments

Some aspects of geocoding are unlikely to change significantly in the future. The underlying logic and algorithms of the process have been long established leading a 1987 U.S. Census Report to state, "we do not expect decisive improvement in either standardization or linking programs" (O'Reagan and Saalfeld 1987). However, significant potential exists in the integration of geocoding with other techniques in GIS. For example, combining remotely sensed land-use/land-cover maps with street files and census data can provide a better picture of where people reside (Mennis 2003; Balk et al. 2004). There is also increasing interest in documenting the differences between where people live and where people work, building on census records (Bhaduri et al. 2002).

Finally, there is always the possibility for creative Census-taking: the Pune slum Census is one example (Sen, Hobson, and Joshi 2003). In a painstaking process, administrators trained slum-dwellers in survey collection and GIS. The product was a GIS representation of a slum area that was not previously on the radar screen of Indian government officials. Household locations were mapped, and attributes for each household and the neighborhood more broadly were incorporated. The Census bought a sense of empowerment and certainly increased the visibility of the people and living conditions in Pune slum, which, in fact, dispelled many previous assumptions about the area.

Geocoding is one technique among many in a suite of GIS tools that are useful for the collection, analysis, and distribution of Census information. In an international context, it is highly valuable for each country to contribute to an international spatial data infrastructure, which supports information on people all over the globe. Census data provides the critical link to help us understand how societies around the world develop, change, interact in a global economy, and affect/are affected by environmental change. The Census is the most geographic of any data collected, illuminating place-based differences and enabling local and international governance to respond to the needs of many.

Bibliography

Balk, Deborah, Francesca Pozzi, Gregory Yetman, Uwe Deichmann, and Andy Nelson. 2007. *The Distribution of People and the Dimension of Place: Methodologies to Improve the Global Estimation of Urban Extents.* CIESIN, Columbia University 2004 [cited 27 Feb 2007]. Available from http://sedac.ciesin.org/gpw/docs/UR_paper_webdraft1.pdf.

Bhaduri, Budhendra, Edward Bright, Phillip Coleman, and Jerome Dobson. 2002. LandScan: Locating people is what matters. *Geoinformatices*:34-36.

Bolstad, Paul. 2005. *GIS Fundamentals*. White Bear Lake, MN: Eider Press.

Davis, Clodoveu A. Jr., and Frederico T. Fonseca. 2007. Assessing the Certainty of Locations Produced by an Address Geocoding System. *Geoinformatica* 11:103-129.

Grigg, M, B Alfred, C Keller, and JA Steele. 2006. Implementation of an Internet-based Geographic Information System: The Florida Experience. *Journal of Public Health Management and Practice* 12 (2):139-145.

Karimi, Hassan A., Matej Durcik, and William Rasdorf. 2004. Evaluation of Uncertainties Associated with Geocoding Techiques. *Computer-Aided Civil and Intrastructural Engineering* 19:170-195.

Kumar, Naresh. 2007. Spatial Sampling Design for a Demography and Health Survey. In *Annual Meeting of the Population Association of America*. New York, NY.

Longley, Paul A., Michael F. Goodchild, David J. Maguire, and David W. Rhind. 2005. Georeferencing. In *Geographic Information Systems and Science.* West Sussex, England: John Wiley & Sons Ltd.

Martin, David. 2006. Last of the Censuses? The Future of Small Area Population Data. *Transactions of the Institute of British Geographers* 31 (1):6-18.

Mennis, Jeremy. 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer* 55 (1):31-42.

O'Reagan, R. Thomas, and Alan J. Saalfeld. 1987. Geocoding Theory and Practice at the Bureau of the Census. Washington, DC: Bureau of the Census Statistical Research Division Report Series.

Sen, Srinanda, Jane Hobson, and Pratima Joshi. 2003. The Pune Slum Census: creating a socio-economic and spatial information base on a GIS for integrated and inclusive city development. *Habitat International* 27:595-611.

Shi, X. 2007. Evaluating the uncertainty caused by Post Office Box addresses in environmental health studies: A restricted Monte Carlo approach. *International Journal of Geographical Information Science* 21 (3):325-340.

Yang, Duck-Hye, Lucy Mackey Bilaver, Oscar Hayes, and Robert Goerge. 2004. Improving Geocoding Practices: Evaluation of Geocoding Tools. *Journal of Medical Systems* 28 (4):361-371.

Ziao, Hong, Celement K. Gwede, Gebre Kiros, and Katherine Milla. 2007. Analysis of prostate cancer incidence using geographic information system and multilevel modeling. *Journal of the National Medical Association* 99 (3):218-225.