UNITED NATIONS
Department of Economic and Social Affairs
Statistics Division

ESA/STAT/AC…/1
March 2008
English only

United Nations Expert Group Meeting on
Measuring the Economically Active Population in Censuses
7-10 April 2008, New York

HANDBOOK ON GEOGRAPHIC DATABASES AND CENSUS MAPPING
(Draft version)

Draft 7 March 2008

UN Handbook on Geographic Databases and Census Mapping[1]

Introduction

This revised edition of the *UN Handbook on Digital Mapping and Geographic Information Systems (GIS)*, now provisionally retitled the *UN Handbook on Geographic Databases and Census Mapping*, seeks to recognize and build on the successes of the first edition.

Major technological advances including the widespread availability of personal computers, global positioning systems (GPS) and low-cost aerial and satellite imagery, have put new tools in the hands of national statistical organizations (NSOs) to collect better – more accurate, timely, and unbiased – information about their populations. The emergence of new technologies is indeed the driving force behind this substantive revision of the handbook, acknowledging the new geospatial applications but also recognizing that adopting such new methods will challenge the leadership of NSOs and evince changes in their organizations.

This revised handbook argues that some of the biggest challenges for NSOs are not merely technical but also organizational, institutional and managerial. Most member countries have

[1] This revised edition is carried out by Dr. David Rain, consultant to the UN, in collaboration with UNSD.

begun to make use of geographic information science and technology appropriate to the scale and scope of data collection needs Many countries are discovering that they can leverage the strengths of other government agencies through what it is referred to as National Spatial Data Infrastructures (NSDI)[2]. Institutional issues such as funding, staffing, and project management basics, while not themselves technical, will have a bearing on the success of geospatial census projects.

Provided here are constructive options for the reorganization of census mapping and analytical tasks centered around digital geo-referenced databases or 'geodatabases.' Structuring the statistical organization around geographically referenced data calls for careful planning, not only because of upfront investment required for GIS, but also for the work required *to develop the capacity* to analyze census data and deliver products to the public in a timely way using effective communication skills. The success of GIS as an industry is built on the power of geospatial information to problem-solve, and the chance to utilize this power is in the hands of the NSOs.

Developing GIS capacity may mean expanding the existing "cartographic unit" to a much larger and more versatile geographic core with the capability to serve all census mapping needs. Achieving this re-organization may require a continuously funded operation, staffed with expertise to help the NSO perform its tasks decade-round. A dedicated GIS core will require a team with not only GIS skills but also a more distant and elevated operational goal of modernizing census-taking. Trained personnel will follow a strict timetable that ensures that

[2] NSDI is a combination of technology, policies, standards and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data. Conceptual parts of NSDI: 1. Institutional framework defines the policies, legal and administrative support to create, maintain and apply the standards to fundamental data sets; 2. Standards define technical characteristics to fundamental data sets; 3. Fundamental data sets require geodetic framework, topographic and cadastre data bases; 4. Technological framework allows the users to identify and receive the access to fundamental data sets (from GSDI Cookbook, 2000).

corrected detailed maps be delivered to enumerators in the field in time for the census operation. A digital Enumeration Area (EA) database will include population estimates of localities as well as EA boundaries that are defined geospatially to represent small areas. Tasks will require continuous measurement including a combination of fieldwork (which can be costly when provisioning crews in all-terrain vehicles) and remotely sensed data collection – resulting in a more accurate and cost-effective census.

Planning for the operational implementation of a geographic program cannot happen soon enough. Due to long lead times for implementing complicated organizational restructuring, planning needs to be initiated years in advance. For the enumeration, highly detailed large-scale maps will always be needed for enumerators and crew leaders. Producing more *granular* geographic data, with dissemination units composed of enumeration areas or population clusters, at a fine enough scale to capture variation and allow use with other data layers in a geographic information system (GIS), will involve considerable planning on the part of the NSO. Establishing a dedicated geographic core or data hub using an enterprise-wide relational database system may also require organizational restructuring. Spatial analysts use specialized software to conduct research and inform decision-making within the government and civil society. Foundational data layers such as demography and administrative boundaries that are produced by the NSO can be shared among many users, eliminating the cost of duplicative efforts. These two layers are basic components of any National Spatial Data Infrastructure. Indeed, for spatial data infrastructure (SDI) efforts, NSOs' contribution is likely to be:

First, a spatial geographic database with polygonal and attribute information for the enumeration areas (EAs) of the country, the units for which the census territory is originally allocated to census-takers. A common digital base can assist with censuses of agriculture as well as

population, as recent experience in Brazil and Russia suggests. Census data can be released at the EA level or aggregated into new small-area dissemination units such as population clusters;

Second, a library of digital administrative boundaries ranging from the provincial to the municipal levels (perhaps even at the level of the land parcel). If the boundaries are fully surveyed and field corrected by the NSO, then substantial cost savings will result when this contribution can save the re-surveying of territory for other purposes such as electoral redistricting;

Third, a national gazetteer including geographic names and coordinates of population settlements (referred to as "P-Codes" within UN Humanitarian community) which when paired with population estimates can be used by humanitarian organizations for development and emergency response. Additional data on housing units in vector format can be similarly used.

For geographic and demographic data in GIS format to be shared among other organizations, NSOs will need to follow standards for geographic referencing established at the national level. Consideration needs to be given especially to the administrative classification system to be used to organize territory for the census. In this document hereafter this will be referred to as '*geo-coding*,' for it serves as a linking point for demographic information and the location on the earth surface.

Above all, we highlight the need to develop realistic plans to harness the power of GIS and other geospatial technologies to modernize census operations and get better results and analysis, and to extend the scope of census cartography to the national spatial framework of the country.

Scope, purpose and outline of the handbook

The rapid recent developments in digital mapping and geographic analysis technology and the increasing demand for geo-referenced small area population data continue to be the main motivation for this handbook. Any country embarking on a census project will need to plan extensively to minimize the costs and maximize the benefits from the required mapping activities. Besides informing key supporters of geographic initiatives within NSOs, this handbook is aimed at providing technical and methodological background information to support the choice of a suitable set of tools and procedures for a given country.

Given the multitude of available options and the differences in conditions and available resources among countries, clearly the choices will be different in each case. This handbook is therefore not a step-by-step cook-book. Each country needs to evaluate how available mapping options fit into the context of their own census program and national plans. Issues such as existing geographic resources in the country, existing technology resources and staff, available funds and the time frame allocated to complete the geographic tasks for the census will determine the best mix of technology and approaches for each individual case.

This handbook respects the grand tradition of census cartography and does not argue that traditional analog mapping techniques that have been used successfully in many countries are completely obsolete. The main reference on the topic, the U.S. Census Bureau's book Mapping for Censuses and Surveys (1978) continues to be an invaluable resource for beginners as well as experienced cartographers. In particular, the chapters on organization and control of a mapping program, enumeration area delineation, and statistical areas continue to be relevant, and this handbook makes extensive use of material from the older book. As technology has progressed,

however, there are now better ways of doing many of the census mapping tasks. This handbook therefore aims to complement the earlier guidelines by providing information on recent technologies.

The main chapters of this handbook assume a basic knowledge of GIS and cartographic concepts. For readers less familiar with these subjects, Annexes 1 and 2 provide a brief overview of both topics. In particular cartographic projections and coordinate systems are a more important topic in a project utilizing GIS than in a traditional approach based on sketch maps.

Chapter by chapter summaries

The main chapters of this handbook are divided into those topics relevant to activities before, during, and after the enumeration. The new revised handbook organization reflects the operational decision in many member countries to effect a transition to digital census operations.

Chapter 1 consists of an executive summary for statistical agency heads that covers the role of mapping and geospatial information in census data collection; then the chapter covers the following topics: managing the geospatial census; costs and benefits of a geospatial program; critical success factors; planning the census mapping process; identifying user needs, staffing and training; and institutional collaboration through the vehicle of spatial data infrastructures-what they are, and making use of them for base data; sharing data.

Chapter 2 provides technical content on the step-by-step process of assembling a digital enumeration area (EA)-level database, including geodatabase basics, the mechanics of data input, geographic coding ('geocoding'), and EA delineation. This section is for the data processing

manager or cartography/GIS chief to use for the practical establishment of a digital geographic database.

Chapter 3 spells out the process of incorporating technical advances such as GPS and remote sensing, adding a section on integrating these geospatial tools with ground-based work. Employed here is an approach borrowed from remote sensing science called "change detection," which seeks to direct emphasis to areas that have changed since the prior census. A final new section in Chapter 3 covers the warehousing and retrieval of digital geographic information through spatial database engines or spatially-enabled relational databases.

Chapter 4 covers the process of creating the maps needed for enumeration, with an operational focus that picks up where the geographic database discussion left off. Included are sections on process flow for enumerator maps, compilation, determining relevant layers for enumerators and supervisors, printing and distribution basics, a project management approach to enumeration, and a plan for contingencies in the event of delays.

Chapter 5 employs a similar approach for the postcensal phase, but again builds on the notion of the geographic database (geocoded to specific locations) as the central data interface. New sections here include a list of products from the census, the aggregation of census collection units (also known as enumeration areas or EAs), dissemination units to display results, disclosure and privacy issues (relevant for the release of small-area data) and the option for a national statistical agency to raise revenues selling value-added products such as data CD-ROMs or DVDs. It also includes geographic data products like map viewers, attributed spatial files for use in commercial GIS packages or on-line products. Finally, the chapter addresses data maintenance issues for a continuously updated census geo-database and provides arguments for continuous or updated

measurement in the field. A new section on addressing systems will also be included. The annex section, providing background information on GIS, R/S, GPS and some mapping and geospatial principles, is updated from the 2000 edition, with a new section added on where to go for assistance.

Executive summary for national statistical heads and other decision-makers

The following section is aimed at the managers of national statistical organizations (NSOs), census chiefs, and geographic section heads. It covers managerial or institutional issues (i.e., the non-GIS technical content), making a case using country experiences that NSOs can produce geo-referenced data with a higher degree of accuracy in less time using new technologies. Included are examples that describe a set of country experiences illustrating the utility of geospatial technologies for census work.

Planning for a geospatial census will require some reorganization of the NSO since there is no one technology solution that will allow the NSO to modernize the census. The handbook presents a range of choices that allow the NSO to adopt new technologies at the appropriate comfort level. It conceives the options as a "palette" that the NSO can choose from, with specific tasks that can be scaled and adopted in whole or part depending on the size of the country and the NSO. Throughout the process it is important to stress the need to begin planning early and ask for help when it is needed.

Text box1: Four countries that have implemented digital programs with results and efficiency gains:

Namibia began its digital mapping and GIS program in preparation for its 2001 census, with its main aim being the efficient production of base maps needed for fieldwork. GIS infrastructure was established through the use of a consulting firm, with spatial data needed for the census captured and digitized using GPS. Namibia has a land area of about 824,000 square kilometers with a population of 1.8 million, so covering the entire territory was a logistical challenge. The Central Bureau of Statistics (CBS) demarcated enumeration areas for 13 regions and 107 constituencies and geocoded them using unique identifiers composed of nine digits. Additional boundary layers including national parks, farms, communal lands, townlands, and localities were also created. Some challenges included working with an organizational structure that did not include GIS, lack of trained personnel, lack of training, initial lack of spatial data, boundary problems, and inaccessibility of some areas. To address human resource gaps, Namibia has built a partnership with the Polytechnic Institute to create specific curricula in GIS and ICT. Some of the software used in the digital mapping program included GEO-MEDIA 6.0 and ArcGIS 9.2, ER Mapper, IDRISI Andes, GeoPDF and MapGuide. For its 2011 census, Namibia will establish a web-based GIS using open source software called Postgress that will enable users to create their own maps. Data will be stored and accessed centrally using Oracle. CBS is planning to capture dwelling units and use them demarcate EAs. (contact: Ottilie Mwazi)

Bhutan conducted its 2005 census of population and housing using GIS. The National Statistics Bureau (NSB) ensured complete coverage in its delineation of 6,800 enumeration areas, for a country of 47,000 square kilometers and a total population of approximately 2.3 million. Every structure including shelters, temporary housing, and caves were located using GPS during the house listing operation. Spatial information was extracted from 1: 50,000-scale topographic maps, and EAs were created to nest within administrative units at the Dzongkhag and chiwog levels. For data dissemination, NBS set up ArcSDE and ArcIMS Server and migrated a geodatabase of administrative boundaries at the Dzongkhag level, using a Windows server reachable through the NSB web site. Some of the disseminated indicator maps include employment, health, housing, water access, energy, and sanitation. Plans call for using the digital base layers for planning the next Bhutan census. (contact: Thinley Jyamtshow Wangdi)

St Lucia began to develop its GIS capabilities in 1995 for its census of agriculture. The Central Statistical Office (CSO) realized that the hand-drawn and contour survey maps used in the past were inadequate for the task of locating housing units for a census of population. Through cooperative assistance with the Land and Surveys Department, the physical planning unit within the Ministry of Planning, the forestry department, the Ministry of Agriculture, and on the private sector side with Cable and Wireless and St. Lucia Electricity Services Ltd., the CSO digitized enumeration district maps and used high-end GPS to capture spatial data from the field. During preparation for the 2001 census, the CSO obtained cloud-free satellite imagery from the Land and Surveys Department. In 2004, the CSO was able to make use of aerial photos of the island through the assistance of the Survey and Mapping Department of the Ministry of Physical Planning. The CSO designed settlement boundaries using latitude-longitudes for every housing unit, thus allowing virtually any area within St. Lucia to be defined. Settlements were geocoded using a nine-digit number. During the buildup to the conversion to digital mapping operations, St. Lucia experienced human resource challenges including staff attrition. Even so, it was able to provide GIS training to its existing staff and complete its designated tasks. (contact: Sherma Lawrence)

Insert additional country (perhaps a more developed, large country) as a fourth example.

Censuses continue to be one of society's most important tools for understanding human populations scientifically. Geospatial technology has transformed the way that information can be presented for societal benefit and for the promotion of social, economic, or sustainable development. GIS promises measurable productivity gains in almost any industry, using the organizing central principle of geography, that space and place matter. Organizing information geographically, through ordered display of data that is centered around location, can effectively tie social observations to their location.

The role of GIS in national statistical organizations is changing, and increasingly countries see the merit in organizing government information around a spatial model that implicitly incorporates geography. Lateral communication through the mechanism of the geographic

database promises, and in many cases has delivered, gains in such important areas as efficiency and customer service, when government agencies communicate results and inquiries among themselves through the use of GIS. The role of GIS in improving government efficiency is highlighted in many studies of "e-government" (O'Looney 2002, Garson 2003, Khosrow-Pour 2005, White 2007).

Because of the large upfront investment, in terms of both hardware and software and re-organization, GIS use in census operations needs to be carefully planned. GIS industry leader David Rhind, from United Kingdom, acknowledges GIS not just as a technology but increasingly as a part of the way in which commerce, government, and academia – all operating in some sense as businesses – operate. Like any strategic investment, GIS costs and benefits need to be weighed carefully. Roger Tomlinson, from Canada, maintains that good planning leads to success, whether for GIS or any other technology implementation. A census is an elaborate national undertaking that will influence policy for years to come. Plan for this by thinking about the products the NSO will produce over the course of the next few years. Working backward from publication to the planning and data needs, think about conventional census products – including tabulations, totals, age and sex breakdowns – that will need to be released. Also think about new offerings such as atlases, DVDs with disaggregated data or small-area data, or microdata, and electronic maps that could meet the needs of a host of new data users. Using the power of geography, internet delivery of products and services are within your reach.

As Tomlinson also points out, GIS is a particularly horizontal technology, with wide-ranging applications across the industrial and intellectual landscape. The planning process will be covered in more detail later in this chapter. For now, consider your agency's strategic purpose. Most likely your main goal is doing an accurate census, on time and under budget. What specific

ministerial objectives or mandates placed on your organization affect census plans? Tomlinson asks organizations to consider how can their strategic purposes can be improved by using new technology such as GPS and geodatabases. He recommends doing extensive cost-benefit analyses over multiple time horizons and asking users about their needs. Plan your agency's products after talking with users and developing an information product description that can then be used to meet requirements. Later in the process, create a data design using the requirements you identified, then choose a logical data model for the data you create and determine system requirements; finally plan the implementation.

In the realm of the census, considering the potential sea of change brought on by new geospatial technologies, NSOs should plan a reorganization of processing and dissemination systems to deliver the products that customers demand. Potentially widespread impacts could be felt on the organizational structure, starting with the option of centering operations around a data hub or information core that stores and serves data within the organization and to the greater society through analysts and disseminators, especially using the Internet. A 'geospatial one-stop' approach can help coordinate delivery of census products to thousands of new data users.

A geo-centric census means organizing data geographically. For many NSOs on the cusp of fully embracing digital capabilities, often the investment is in analog-to-digital conversion of paper enumeration area maps, which involves careful scanning and correction procedures so they can be used as a basis for new digital geographic databases. These new geographic databases can then be compared against remotely-sensed imagery such as aerial photos or satellite images, and field-corrected using GPS. Incorporating technology such as scanning, satellite imagery, and GPS allows the mapping unit to focus efforts on the areas most in need of updating since the previous census.

For any new technical undertaking, a key issue is building capacity. Many NSO managers do not think they have the budget, nor the institutional capacity, to re-organize their agencies. Consider this less a question of budget and more to the scale of the planning horizon. Think ahead five or even ten years; the NSO will need a far finer level of detail in its information products than what can be provided at present.

Traditionally, the role of maps in the census has been to support enumeration and to present aggregate census results in cartographic form. Cartographic automation, GIS and other geospatial tools have greatly expanded this role. In addition to enabling more efficient production of enumerator maps and thematic maps of census results, GIS now plays a key role in the analysis of population and household data and in census data dissemination. Advances in technology and new tasks for GIS using new data sources like remote sensing and GPS-enabled locational recording have expanded the power of geographic representation within the NSO. Also, web-based mapping tools have extended the outreach of census information.

Mapping has been an integral part of census taking for a long time. Very few enumerations during the last several census rounds were executed without the help of detailed maps. In general terms, digital mapping serves several purposes in the census process:

- *Maps ensure consistency and facilitate census operations (pre-enumeration).*

  The census office needs to ensure that every household and person in the country is counted, and that no households or individuals are counted twice. For this purpose, census geographers partition the national territory into small reporting units. Maps thus provide an essential control device that guarantees consistency and accuracy of the census.

- *Maps support data collection and can help monitor census activities (during enumeration).*

  During the census, maps ensure that enumerators can easily identify their assigned set of households. Maps are also issued to census supervisors to support planning and control tasks. Maps can thus also play a role in monitoring the progress of census operations. This allows supervisors to identify problem areas and implement remedial action quickly.

- *Maps make it easier to present, analyze and disseminate census results (post-enumeration).*

  Cartographic presentation of census results provides a powerful means for visualizing the results of a census. This supports the identification of local patterns of important demographic and social indicators. Maps are thus an integral part of policy analysis in the public and private sectors.

The next section of this introduction provides a brief overview of the underpinning reasons that encourage many NSOs to move from traditional cartographic methods to digital mapping and GIS. Rapid developments in digital mapping and geographic databases have been the motivation for preparing this handbook. We will then discuss why census offices are being asked to provide timely census data in geographically referenced form, and discuss some costs and benefits of developing geospatial capacity. Lastly, some recommendations are made for helping the NSO participate fully in national spatial data infrastructure efforts.

From maps to geographic databases: the mapping "revolution" continues

People have used maps for centuries to represent their environment. Maps are used to show locations, distances, directions and the size of areas. Maps also display geographic relationships, differences, clusters and patterns. Maps are used for navigation, exploration, illustration and

communication in the public and private sectors. Nearly every area of scientific enquiry uses maps in some form or another. Maps, in short, are an indispensable tools for many aspects of professional and academic work. Today maps are but one form of information display subsumed under a broader concept of geographic information, and the form this geographic information most often takes is the geographically referenced database (or geodatabase).

Cartography was influenced by the information revolution somewhat later than other fields. Early computers were good at storing numbers and text. Maps, in contrast, are very complex and digital mapping requires large data storage capacity and fast computing resources. Furthermore, mapping is fundamentally a graphical application and early computers had limited graphical output capabilities. The earliest computer mapping applications implemented on computers in the 1960s therefore did not find wide application beyond a few government and academic projects. It was not until the 1980s when commercial geographic information systems reached a level of capability that led to their rapid adoption for example in local and regional government, urban planning, environmental agencies, mineral exploration, the utility sectors and commercial marketing and real estate firms.

GIS has benefited greatly from developments in various fields of computing. Better database software allows the management of vast amounts of information that is referenced to digital maps. Computer graphics techniques provide the data models for storage, retrieval and display of geographic objects. Advanced visualization techniques allow us to create increasingly sophisticated representations of our environment. GIS data display functions go far beyond static two-dimensional displays and provide animation and three-dimensional modeling capabilities. Just as the input of textual information is facilitated by optical character recognition, fast, high-

resolution scanning and sophisticated software speed up map data conversion that previously relied exclusively on manual digitizing.

New information sources also shorten the time from project planning to operational database. The most important recent developments have been in navigation, remote sensing, image analysis, data manipulation and Internet mapping. The global positioning system (GPS) revolutionized field data collection in areas ranging from surveying to environmental monitoring and transportation management. A new generation of commercial, high-resolution satellites can deliver pictures of nearly any part of the earth's surface with enough detail to support numerous mapping applications, including those for censuses. The cost of precision digital mapping has fallen significantly due to the close integration of GPS techniques and remote data capture.

Similar advances are occurring in the areas of geographic data dissemination. All major GIS vendors now provide the tools to make geographical databases accessible via the Internet. Government agencies at all levels are embracing this technology to provide access to vast amounts of spatial information to the public cheaply and quickly. The Internet is replacing printed maps and digital media as the most important means of data distribution.

Internet mapping programs are one indication that the tools to utilize digital spatial information are constantly becoming cheaper and easier to use. While high-end GIS packages still require considerable training to be used effectively, desktop mapping packages are no more complicated to use than standard business software. Digital mapping is also becoming ever more closely integrated in standard computer applications such as spreadsheets, graphics and business management software.

Statistical offices were some of the early adopters of GIS. Population, social and economic statistics are the foundation of public planning and management. The spatial distribution of socioeconomic indicators guides policy decisions on regional development, service provision and many other areas. Digital techniques allow better management, faster retrieval and improved presentation of such data. There has therefore always been a close linkage between geography and statistics—as reflected, for instance, by the fact that the national statistical and mapping agencies in many Latin American countries are housed under the same roof. This close integration of GIS in statistical applications yields large benefits to national statistical offices as it reduces the cost and time required to collect, compile and distribute information. GIS allows the statistical office to produce a greater number of services, thereby considerably increasing the return on investment in data collection.

Increasing demand for local area statistical data

The benefits of geographic data automation in statistics are shared by the users of census and survey data. The data integration functions provided by geographic information systems, which allow the linking of information from many different subject areas, have led to a much wider use of statistical information. This in turn has increased the pressure on statistical offices to produce high-quality spatially referenced information for small geographic units. The types of applications for such data are almost limitless. Some examples are:

- Emergency planning and humanitarian response: Agencies can prepare for the eventuality of natural disasters by identifying highly populated areas that can be difficult to evacuate in case of fires, earthquakes, volcano eruptions or tsunamis. Following a major natural disaster, some of the early questions asked include: which villages are affected? What is the size of their

population? How many people were killed, injured and made homeless? What is the status of infrastructure, particularly roads and bridges, health centers, schools, water supply systems and government building, etc.? If digital maps of population distribution and housing characteristics could be easily overlaid with elevation and slope data, transportation networks, and other geographic information of the affected area by the disaster, it is possible to generate reliable estimates of the number of people affected, their needs in terms of medical aid, food and shelter, and particularly their location. Standard "P-Codes" for populated places eases the logjam of locating affected areas in need of assistance.

- Flood plain modeling: Major flooding is an increasing risk in many of the world's watersheds. Digital elevation and hydrological data in combination with small-area census statistics allows planners to make detailed assessments to reduce the risk for populations in flood-prone areas and for emergency management planning. Insurance companies in some countries use the same tools to assess risk levels of home owners which leads to a fairer assessment of premiums.

- Planning of social and educational services: A main task of local and regional government is to ensure that all parts of the country have equal access to government services such as health care and education. Small area census data on age and social characteristics allows planners to forecast demand for various services. In combination with GIS data on transport infrastructure, this information allows better distribution of resources among existing service centers and more rational decisions concerning the location of new facilities.

- Poverty analysis: In countries where income or consumption data are not collected during a census, household characteristics are an important indicator for the welfare of various

population groups. Small area census data in combination with spatially referenced information on infrastructure and agro-ecological conditions can be used to estimate poverty incidence and the location of poor communities. This information improves targeting of poverty alleviation schemes by channeling resources to areas of greatest need while avoiding leakage of subsidies to non-poor communities.

- Utility service planning: Private and public water, gas, electricity and telecommunications utilities not only use GIS to manage their physical infrastructure, but also use spatial analysis of demographic data to assess current and future demand for services. Digital census data—together with digital terrain models—have been a key component in the design of location-based services around the world.

- Labor force analysis: Whether it is a private company looking for a suitable site to locate a factory or a government agency attempting to match labor supply and demand, small area census data are an important element in employment related analysis. Journey-to-work analysis in which the location of jobs and the residence of employees are compared is critical to transportation planning.

- Marketing analysis: Companies use small area census data to plan the location of new stores and warehouses, to manage customer service information and to target advertising. An entire branch of GIS—termed variously business geographics or geo-demographics—has flourished. In fact, the strong demand for these types of analysis has been a major driving force for the development of inexpensive, easy-to-use desktop mapping packages.

- Voting district delineation: In representative democracies parliamentary representation is based on the principle of equal weight for each vote. To guarantee this principle, small area

population figures are used to delineate voting districts of approximately equal size. For example, in the United States this is the main basis for the decennial census required by the constitution where GIS and census data are employed in the design of electoral districts.

- Epidemiological analysis: Small area census data in combination with health incidence and bio-physical data allow health officials to estimate the population at risk of certain infectious and vector-borne diseases. Knowing how many people in the country are potentially affected by malaria or bilharzia, for instance, allows planners to estimate the resources required for eradication measures. Identifying where these risk groups are located supports prioritization and implementation of intervention activities.

- Flood plain modeling: Major flooding appears to be an increasing risk in many of the world's watersheds. Digital elevation and hydrological data in combination with small-area census statistics allows planners to make detailed assessments to reduce the risk for populations in flood-prone areas and for emergency management planning. Insurance companies use the same tools to assess risk levels of home owners which leads to a fairer assessment of premiums.

- Agriculture: Geographic information on agro-ecological conditions and production data together with small area data on the demand for food products facilitate the analysis of food security issues. Famine early warning systems have been set up in many countries characterized by fragile ecosystems to prevent major food crises.

The efficiency argument—long-term cost savings

This section discusses the costs involved and the potential benefits realized in using a digital cartographic or GIS approach in census mapping. There are a variety of options ranging from a fully digital in-house mapping capability to using, for example, desktop mapping for presentation of results and dissemination only. In other words, there is no "one size fits all" solution to the introduction of digital mapping in the census process. However, advanced innovation and lowered costs for entry into GIS make conversion at this point certainly within reach as complex, high-end GIS systems become usable on inexpensive desktop computers. The appropriateness to the task is the overriding principle of any cost-benefit analysis.

For various reasons, it is also very difficult to assess the costs and benefits of using GIS quantitatively. For example, many of the benefits may not be realized by the agency paying for the GIS investment, but rather by outsiders who are gaining access to products of higher accuracy or lower cost, or who may obtain products that were previously not available at all. This also highlights the difference between "cheap" and "cost-effective." The cheapest option for producing census maps, especially in countries where labor costs are low, may still be the traditional manual approach. From a societal standpoint, however, it may be more cost effective to invest initially more money in a digital approach because the digital output products will realize much larger long-term benefits inside and outside the census or statistical agency, creating truly a national initiative.

Textbox2: For the discussion of technological and cost hurdles (Lesotho or equivalent—*still need*)

Cambodia undertook EA and village boundary mapping as a part of its 2003 health survey. Census GIS activities have been part of the Department of Demographic Statistics of the National Institute of Statistics. Census mapping was coordinated with the Ministry of Land Management and the Ministry of Interior. A staff of 30 was assigned to delineate EAs as units which have no more than 120 households. A

village questionnaire was also used to produce an EA database. Among the functions of the EA database were operations to label, check in, code, edit, perform data entry, verify, and clean boundary data. Barcodes were used to facilitate tracking of EA maps. For its 2008 census, the NIS plans to disseminate results on-line using the "CamInfo" system that is based on DevInfo. Cambodia estimates that the cost of cartography and EA delineation, combined with that of data processing, represents approximately 15 percent of the total cost of the census. Particularly costly from a resource standpoint are field vehicles and computer (and computer peripheral equipment), although these are often provided by donors.

Budgetary investments in GIS are heavily front-loaded, which means that the major costs are incurred early on in a project while tangible benefits may only materialize long into the project cycle. When contrasting the costs and benefits of a traditional mapping approach with digital cartography, one notices that in the traditional mapping case, maps are re-created manually for each census (this is also known as *"reinventing the wheel"*). Too often in the past, census mapping was purely project based. A few years before the census, a team was assembled to quickly produce census sketch maps by hand which are only used for the enumeration. Several years later the process started again for the next census.

Having a long-term strategy for census mapping means viewing census cartography and geographic database development as a continuous process with regular maintenance of databases by a permanent core staff that receives frequent training. In the analog case, the costs tend to be higher than the benefits, since the hardcopy maps are useful for census purposes only. In the digital case, an initially large investment results in lower maintenance and updating costs and sustainable benefits in the long run. The long-term benefits are significantly higher because the process results in a digital database with multiple uses by the NSO and other government agencies.

Costs and benefits

As with the adoption of any new technology, GIS presents challenges to the existing organization of statistical offices and should be adopted only after costs and benefits have been carefully considered. Cost components include systems design, hardware and software acquisition, prototype development, human resource planning, training, geodatabase design, transitional costs, data acquisition, data capture, quality control and system maintenance. The transition to geospatial data incurs costs associated with digitization, namely the conversion of analog enumeration area maps to geospatial formats.

Benefits of migrating to geospatial systems for census work can be broken into two categories, efficiency benefits and effectiveness benefits. *Efficiency* refers to the amount of output that can be obtained per unit of input. In terms of a census, it means doing more for less money. Efficiency benefits include cost savings or productivity gains, time savings, greater credibility and authority of geospatial data products, better service, increased accuracy, improved consistency, and income generation. These can be realized by the census organization itself with proper planning.

Effectiveness benefits pertain to the impacts of policies or programs that benefit from improved information, including societal benefits realized by the users of census-derived statistical data. Effectiveness benefits include improved analysis at a scale more appropriate for local and regional studies, more informed policymaking, more extensive data sharing with other government agencies or nongovernmental organizations, and improved outreach to the public at large. Note that Annex 8 covers each of these benefits in more detail.

Critical success factors for GIS implementation in the national statistical agency

Besides the obvious costs that can be quantified for a given GIS project, there are a number of stumbling blocks that may cause a project to fail or to fall short of realizing its full potential. Mostly, such problems result from a lack of planning, the choice of inappropriate hardware and software, and various organizational mistakes. Surveys of real-world GIS projects reveal a set of critical success factors shared by successful GIS implementations. The following list of critical success factors is adapted and expanded from Johnson (1997).

1. A ***champion*** or key person to promote GIS development within the organization.

2. High-level management support.

3. Decision to invest in GIS is need-based and problem-driven rather than technology-driven.

4. Detailed strategic, operational and managerial planning based on a realistic assessment of costs and effort involved.

5. Clear goals and objectives defined for the GIS department.

6. GIS education and training for affected employees *and* management.

7. Staff continuity—the ability to retain skilled staff members.

8. GIS treated not as an independent add-on, but as an integral part of the overall information management strategy.

9. Completion of a user needs assessment and a priori definition of output products.

10. Development of cooperative agreements with other interested parties, including spatial data infrastructure arrangements.

11. Clear implementation schedule.

12. Defined long-term funding plan including cost recovery and data pricing strategies.

13. Accurate estimates for maintenance and associated costs.

14. Explicit operational procedures guiding the use of GIS facilities.

15. Well-established quality control / quality assurance procedures.

16. Clear specifications, requirements and benchmarks to deal effectively with vendors and contractors.

17. Well-defined written contracts with vendors, consultants, partners and clients within and outside the government.

18. Bonuses, incentives, short contracts for retaining employees.

19. Completion of a prototype pilot project to test appropriateness of equipment, software and procedures.

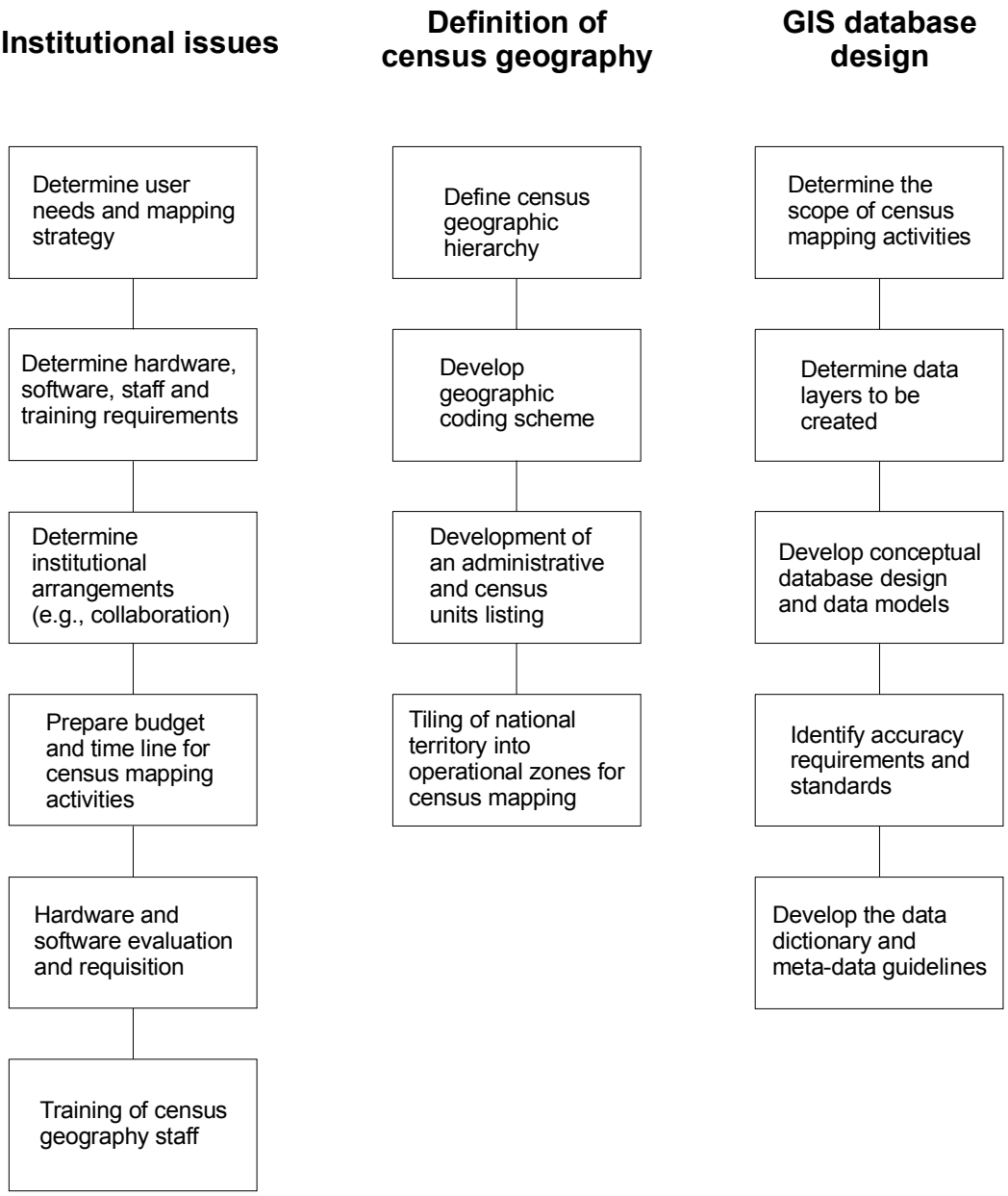
20. Frequent milestones and delivery of output products to encourage adherence to pre-set time frames.

21. Outreach and marketing including published successes.

Planning the census cartographic process

This section deals with preliminary organizational tasks in a census mapping project and with critical design issues that determine the nature of the resulting databases and thus the range of applications that it will support. The success of the actual data conversion process depends on a well-designed institutional environment and a well-planned operational strategy. The planning steps are divided here into institutional issues such as staffing and cooperation with other agencies, the definition of the census geography, and the design of the GIS database. As illustrated in Figure 1.1, these stages can be carried out more or less simultaneously, and many of the choices depend also on the chosen data conversion strategy.

Figure 1.1: Stages in planning census cartographic work



Needs assessment and determination of mapping options

User needs assessment

One of the first steps in a census mapping project is a detailed needs assessment followed by an investigation of feasible census mapping options. The census mapping agency must then reconcile user expectations with what is feasible given available resources.

A successful census planning process requires extensive consultations with the main users of the information that will be produced in the census. This process should be embedded in the general consultation program for the census. As the demand for spatially referenced census data increases, consultations concerning mapping products will receive a more prominent role in this process. Institutions that use statistical maps should be included in the advisory panels that provide input in the census planning process. The census office must consult with three main groups in the planning stages:

a) Persons and institutions participating in the census operations. In order to obtain full information about resources and potential bottlenecks, the census mapping agency must carry out an intensive survey of available human resources in the country, available equipment that can be used, existing digital and analog map products, and ongoing or planned mapping activities by other public and private entities. Avoiding duplication of efforts is key to reducing the cost of census mapping and to timely delivery of census map products.

b) Census geographic data product users. These will mainly come from other government departments, the academic research community, and the private sector.

c) The general public. With access to computers and Internet mapping options, private users will are also an important user group. Citizens may, for example, want to obtain statistical information about their own neighborhood or a neighborhood they intend to move to. With the current rapid changes in technology, the census office must plan carefully to anticipate demand for their data.

Determination of products

User needs will determine the range of output products that need to be completed at the end of the census mapping cycle. Products created by the NSO, which are discussed in more detail in Chapter 5, include:

- A set of digital enumeration area maps, or derived dissemination units, that are designed to enable the production of all output products that will be disseminated to government departments and the general public;
- Geographic boundary files in a digital format for all statistical reporting units for which census indicators will be tabulated;
- Listings of all statistical and administrative reporting units, including towns and villages, their variant names, and geographic coordinates;
- Geographic equivalency files that indicate how current reporting units relate to those used in previous censuses, or how one set of reporting units relates to another set;
- Vector layers containing feature data such as landmarks, roads, schools, hospitals and clinics, which can be used when analyzing population data spatially;

- Street index listings for all major urban areas;
- Centroid files that provide a representative geographic point reference for each reporting unit;
- Gazetteers that provide geographic coordinates for all population settlements and other important geographic features in the country.

User requirements are the most important determinants of a census mapping design. These of course must be weighed against available resources in the budgetary cycle. Various other factors determine the choice of the mapping strategy. Among these are:

- Available financial and human resources;
- Existing digital and analog map products;
- The degree of integration between the mapping and statistical agencies in the country, or other relevant agencies;
- Technical capabilities in the statistical office and in collaborating agencies;
- The trade-off between use of technology which may require foreign exchange and lead to dependence on outside technology, and increased use of low-technology labor which may provide a beneficial boost to local economies;
- The size of the country, both population and areal extent; and
- The time frame available to plan and carry out the census mapping process.

Geographic data options

Each country starts its geographic efforts from its own base of existing information, budgets, technical capabilities and available time frames. There exists therefore a multiplicity of paths towards a fully digital map database for census collection and dissemination purposes. A partial list of available options—in increasing order of complexity—follows:

- Production of digital maps created on the basis of existing sketch maps;
- Geo-referenced enumeration area maps that can be properly integrated with other digital geographic databases;
- Inclusion of geographic reference layers, showing, for instance, roads, rivers, landmarks, point features and other features—these can be included as simple images from scanned maps, or designed as a unit's own geographic database;
- A digital postal address registry where addresses are matched automatically or semi-automatically to digital road databases; and
- A digital database of precisely located dwelling units, created with the aid of geographic positioning systems.

This list is incomplete and is provided here for illustrative purposes. All of these issues are discussed in detail elsewhere in this handbook. The most appropriate census mapping strategy for a country will consist of a tailor-made approach that considers the country's needs and resources. While a cook-book type approach is thus not feasible, this handbook

will discuss the range of available technical and logistical options. From these, the census office must then select the subset of techniques and procedures that best fit the needs of the country.

Staffing, responsibilities and training requirements

Motivated and well-trained staff are a key factor that will determine the success or failure of a census geography project. The goals of a mapping project are equivalent whether the maps are produced by hand or by computer. But the use of computers requires a number of new skills from census cartographic staff since similar products are created using different techniques. Furthermore, a digital GIS database is useful for many more purposes. A census office is thus likely to fulfill additional demands for products and services which were not available before. Every member of the census cartographic staff should therefore have some degree of computer literacy.

Much of the expertise required in the traditional, manual census mapping approach is relevant also for a digital mapping project. Rather than completely replacing existing skills, the digital mapping approach requires additional expertise in computer methods. Thus, only relatively little of the expertise of cartographers and geographers on the staff is obsolete, but the demands on their job skills have increased. For instance, traditionally trained cartographers will no longer need some techniques of manual mapmaking such as lettering, negative scribing or drafting with pen or pencil. Instead, after receiving training in computer methods, they will be able to use their background in map design and cartographic communication to produce well-designed enumeration area or thematic maps using a GIS or desktop mapping package. It is often easier to train a subject

specialist in computer techniques than to train a computer expert in a substantive applications area.

The following sections detail the profile of tasks for which staff are required in a digital census mapping project. The same staff members in a census office may be able to perform several of these tasks as required in different stages of the census project.

Planning. In the early stages of the project, a group of people should be formed that will develop the overall strategy for digital census mapping. This requires people trained in geography, GIS, and computer applications, who have experience in census mapping. In addition to census office staff, the planning group can include representatives from the national mapping agency and other interested government organizations, data user groups, or outside consultants. Technical advisers from NSOs in countries that have already switched to digital census cartography or from international organizations should be involved in the planning process as they can provide useful input in the planning process and help cost out alternatives.

Project leadership. Leading the planning process is the census mapping project leader who also supervises the implementation of the digital census mapping strategy. The project leader should have a background in geography, computer science or a similar field with training in GIS and digital mapping. Experience in census cartography, ideally from a previous enumeration in the country is highly desirable. Management experience or management training is necessary to supervise budgeting, personnel management and scheduling. Good communication skills will facilitate cooperation with other parts of the census project and collaborating agencies. The project leader also has to keep up-to-date

on GIS developments and trends, and must be prepared to adapt the census mapping strategy if conditions change or better solutions become available.

GIS data conversion. Data conversion specialists are responsible for the actual implementation of the conversion of map information to digital database format. They have training in relevant GIS techniques such as scanning, digitizing, and editing of GIS databases, and attribute database development using relational database management systems. The data conversion specialists must determine the most efficient way to develop the digital database and supervise technical staff. Knowledge of new data sources and the technology implications of getting new material to help with the creation of geographic databases are essential as well.

Map scanning and digitizing. While scanning has gained new converts and has assumed a prominent new place in the data input strategies of many NSOs, digitizing remains an trusted option. The technical know-how can be acquired relatively quickly by persons who do not have professional training in geography or a similar field. However, digitizing is an extremely repetitive tasks and requires good concentration, attention to details and a good understanding of the structure of digital geographic databases. The best performing staff should also receive training in quality control/quality assurance approaches. This training requirement pertains as well to heads-up digitizing.

Cartographic design. Cartographers will be in charge of designing all map products including enumeration area maps, supervisory maps and thematic maps of census results. They must have a background in map design, cartographic communication, and training

in GIS and digital mapping. Trained cartographers will have most of the required skills, but should receive sufficient training in computer methods.

Field work. The requirements for census mapping field work have changed with the techniques used for digital map production. As global positioning systems have become an essential tool for field data collection, field staff must now be trained in the operation of these systems and possibly also in the use of laptop or handheld computers used for downloading and displaying these data in the field. While a professional background in geography or surveying is not necessary, field staff must receive training to use the new tools properly.

Systems administration. It was already stated that NSOs need to invest in dedicated information technology (IT) hubs that can serve GIS as well as other IT/data processing needs. Timely completion of a census geographic database project depends on the smooth operation of computer equipment. A systems administrator is in charge of maintaining computer hardware and software systems with the goal of minimizing downtime, supporting census cartographic staff and ensuring data security (e.g., data backups). Even if they are not directly involved in census mapping activities, systems administrators are vital members of the cartographic team, since almost every aspect of the work depends on a well-functioning computer system. Administration of computer systems for the geography branch of the census office can in some instances be covered by general computer support staff in the agency.

Special requirements. Depending on the census mapping strategy that is adopted, some additional expertise may need to be present in the census mapping organization. For

instance, if updating of census maps will make significant use of remote sensing products, an analyst trained in digital image analysis should be on the staff. Other experts that may be required are operators of a high-volume map scanning system, or staff members with experience in database management software systems and computer programming. Still others may have graphics capabilities, web-deployment, or customer relations skills. All such skills are helpful in the development of databases and optimization of software systems.

Levels of training. In many countries there may be a shortage of trained GIS experts that can be recruited on a permanent or temporary basis for the census mapping project. The census office must therefore evaluate training options to ensure that existing and new staff have the proper knowledge required for successful completion of the project. Usually, staff trained in traditional geographical techniques who have some computer literacy, will have little difficulty adapting to digital techniques after going through training. Different types of training will be required for various purposes:

- Short seminars to raise the awareness of the digital census mapping program should be conducted for all staff of the census office, including staff from other sections, and managers. This will foster the integration of the digital mapping project into the overall census process. Better utilization of census mapping products by other census office branches will be another benefit of broad dissemination of information. Such seminars can be conducted by the project leader or specialist census cartographic staff.

- Training for repetitive tasks such as scanning, digitizing or editing can involve short in-house seminars with subsequent on-the-job training. Products developed by new staff should receive close scrutiny to identify whether staff need additional instructions or training or possible re-allocation to other duties.

- The core geographic staff involved in census mapping should receive additional training in GIS and digital mapping techniques. Since training can be expensive, only permanent staff members should be sent to courses conducted by universities, vendors or other organizations in the country or abroad. Individuals that have been trained in this way should take a leading role in informing and training additional staff. A large number of people can be trained by using a hierarchical "training-of-trainers" approach, which is particularly appropriate for a decentralized approach to census mapping.

- Applications of specialized techniques such as digital image processing or advanced computer database applications usually require a professional degree or equivalent experience. If no suitable staff can be hired, the census office should consider, well in advance of the actual mapping project, to send a staff member to a university for training. Many universities and training centers around the world now specialize in professional one- or two-year degree courses in GIS, remote sensing and related techniques. See Annex 7 for listings of institutions offering training in geospatial information and techniques.

Institutional cooperation: National Spatial Data Infrastructure: Ensuring compatibility with other government departments

In many countries, multiple government agencies produce digital geographic databases. National mapping agencies increasingly use fully digital techniques in the entire map making process. But other government departments including the transportation, health, environment and water resources units also use GIS to manage the information they collect and use for analysis and planning. Additionally, private sector companies, for example in the utilities, telecommunications and mining sectors, have realized the advantages of managing their information needs in digital geographic form.

Building infrastructure for geographic information is becoming recognized as just as important to member countries as the building of roads, telecommunication networks, and the provision of other basic services. It is now widely accepted that developing national data infrastructures will better facilitate the availability and access to spatial data for government organizations, the private sector, universities and civil society in general. Indeed, spatial data infrastructures allow the NSO to gain access to expertise within the country as well as existing digital sources such as base data in geospatial formats for use in GIS programs. In this framework, NSOs are seen as information nodes in networks of data and content providers that span the country.

National spatial data infrastructures now exist in over 100 countries in the world and likely, an NSDI already exists in some form or another in your country. Some possible sources of information and expertise that the NSO could benefit from building their spatial data infrastructure include the national mapping authorities, other agencies

including environment and planning ministries, and the military. A basic organizational structure for an NSDI includes a ministry in charge, a lead agency, a forum or network of data producers and users, a steering committee, and technical working groups.

Establishing contact with the existing NSDI requires finding the contact person in the lead agency and contacting them. Most likely the main contact in a typical country is the national mapping organization. Examples of data that could be shared include a database of scanned topographic maps or digital vector coverages at a scale of 1:100,000 or finer that could be used as input for the creation of enumeration area boundaries; elevation data, hydrography, and transportation networks—all of which can serve to help delineate areas for census canvassing.

Textbox3: Examples (3) of data sharing agreements and collaborations

1. UNFPA and Fiji hardware sharing: The development of new technologies for census geography provide powerful tools for increasing the efficiency of the census-taking. The experience in Fiji, where the actual enumeration was followed by a team from the Fiji Bureau of Statistics that geo-referenced every single housing unit occupied by a household using Geographical Positioning System (GPS) devices is one of the examples of how to add value to the census, as this procedure will enable a much more accurate future mapping application and facilitate future censuses and statistical exercises. Funding for the GPS exercise was facilitated by UNFPA, which has proposed to extend the use of the GPS devices to other Pacific Island countries that would like to conduct the same or similar exercises. This sharing agreement is valuable tool for capacity building and harnessing valuable information during the census.

2. Secretariat of the Pacific Island Applied Geoscience Commission (SOPAC) and the Pacific island countries Data Sharing & Processing Agreement: The costs of acquiring satellite imagery that would enable using Geographical Information System (GIS) applications for

census geography, such as delineation of enumeration areas, are considerably lower compared to just several years ago. However, these costs are still not easily born by the majority of national statistical offices. International and National agencies can greatly contribute their existing datasets to aid in census taking activities. In that context, where the Secretariat of the Pacific Island Applied Geoscience Commission (SOPAC) has access to satellite imagery, and it provides it free of charge for all of the Pacific Islands. SOPAC also can provide the scanning and geo-registering of existing maps; this represents an excellent starting point for developing census geography on a contemporary platform. National statistical offices are considering using these options provided by SOPAC in evaluating the use of GIS for the population and housing census.

3. Example from large country's data sharing experience, perhaps the US

It is crucial to ensure cooperation at the national level when it comes to acquiring, using and developing GIS applications. The establishment of the GIS focal points and user groups is an optimal way to ensure synergy and avoid duplication and waste of otherwise scant human resources. The need for continuous and well-formatted technical assistance provided by national and international agencies cannot be understated. For many countries, this represents a vital component in building their capacity to conduct complex exercises such as a census. The success of the census very often relies on the success of the public communications and awareness campaign that precedes it; resources for running such a campaign often come from outside donors and international agencies and these need to be fully exploited.

Numerous users inside and outside government agencies require access to these basic geographic databases. Many of these users need access to several databases or use a standard geographic data layer as a template for their own spatial data collection. Such standard data layers that provide the basis for many mapping and data collection activities are termed *framework data*. Some core data layers that form the elements of national spatial data frameworks can include:

- *Geodetic control* – a system of precisely determined geographic control points that serve as the reference for all mapping activities in a country; sometimes called benchmarks;

- *Topographic database (including digital elevation models)* – scanned or vectorized geographic layers showing terrain, other geological features, water bodies, roads, and populated places;

- *Orthoimagery* – air photos or high resolution satellite images which have been processed to have the same geometric accuracy of a topographic map;

- *Administrative boundaries* – subnational governmental units such as provinces and districts that have been delineated on the ground with a high degree of precision;

- *Transportation* – roads, inland waterways, railroads – any infrastructure used to move people or goods;

- *Hydrography* – surface water features; these can be natural such as rivers, lakes or artificial such as canals;

- *Demographic database* – thematic data such as population distribution (or density), female/male counts, age distribution by 5-year age groups, mapped at the finest possible detail – either at the level of the enumeration area or a derived small-area unit such as the population cluster; and

- *Cadastral information* – an official register of rights and interest in land property.

Other basic data layers such as soil types, vegetation zones, geographical names of population settlements, etc., could be added to this list. Most relevant for the census office are the governmental administrative units, since enumeration areas need to be consistent with the boundaries that form the administrative hierarchy in the country and population distribution. But data layers such as transportation and hydrography are also important for census mapping, since roads and rivers form a natural delineation for enumeration areas.

After a census is conducted, the NSO will be called upon to share its results and products with other participants in the national spatial data infrastructure. Enumeration area boundaries with census information such as basic demographic information, and perhaps housing and community characteristics, are an important data source for other government and private organizations. The NSO could also provide a dataset containing the center points (centroids) of enumeration areas. Health sector analysis, for example, requires detailed information about at-risk populations. Transportation sector planning can make good use of data on demand for public transport services. And public and private utilities need to know where to provide increased capacity of electricity, water or telecommunications services.

The concept of a national spatial data infrastructure consisting of basic geographically referenced GIS databases has three implications for census mapping activities:

- The census office has a responsibility to contribute to the national spatial data infrastructure a consistent set of data reporting units which are consistent with the administrative hierarchy and to which socioeconomic and related information can be

linked. In order to ensure that these census maps can be integrated with other data sources, the census mapping organization should adhere to any existing national geographic data standards.

- To ensure compatibility with other data sets and to facilitate census map development, the census mapping authorities should collaborate closely with other government agencies involved in mapping. Apart from ensuring consistent standards and definitions, collaboration will lead to cost reduction, since it helps avoid duplication of efforts.

- NSOs should pay particular attention to issues with *geographic* referencing. In order to overlay different datasets, which cover areas in the same geographic extent, one needs to know how the position of features has been defined, that is, one needs to know the projection and datum, and details of the co-ordinate system, to ensure the correct spatial relationships between features in different datasets.

A statistical agency's contribution to SDI efforts nation-wide may also entail other forms of participation, including the commitment to attend planning meetings and stay informed about development in the country. The NSO will also be required to provide metadata.

Textbox4: International agency participation and coordination

The Global Map Initiative was proposed in 1992 at the Earth Summit held in Rio de Janeiro, Brazil. In 1996, the International Steering Committee for Global Mapping was established to coordinate the network of official mapping organizations from participating countries. By March 2007, 172 countries and areas around the world were contributing data to (or are now developing data) Global Map. The goal of Global Map is to provide coverage of the whole land area of the

Earth at a scale of 1: 1,000,000, with a spatial resolution of 1 kilometer. Data will be updated at approximately five-year intervals to monitor change over time. Global Map is a platform supporting base maps (land use, land cover, vegetation, elevation) in raster and data maps (population centres, drainage, transportation, and boundaries) in vector format. The data on the Global Map is available to all via the internet free of charge for non-profit purposes. For more information, see http://www.globalmap.org

Dev-Info (UNICEF) is a software tool to assist countries in monitoring the Millennium Development Goals and advocate their achievement through policy measures, multi-sectoral strategies and the development of appropriate interventions. DevInfo is a general purpose package for the compilation and presentation of data. Beyond providing a repository of data, DevInfo provides easy to use facilities for querying the database and for producing graphics (tables, graphs, maps) for inclusion in reports and presentations. DevInfo is expected to be a powerful advocacy tool that will contribute to greater MDG awareness and knowledge at country level and to evidence-based policy-making. For more information, see http://www.devinfo.org/

UNSDI or United Nations Spatial Data Infrastructure was proposed in 2005 in order to promote and achieve sustainable development and improve humanitarian and peacekeeping operations. The vision for UNSDI is for a comprehensive, decentralized geospatial information framework that facilitates decision-making at various levels by enabling access, retrieval and dissemination of geospatial data and information in a rapid and secure way. UNSDI enables interoperability between spatial data infrastructures developed for specific purposes within UN agencies, among groups of UN agencies sharing common interests, and among the UN, Member States, and their regional and thematic groupings, and partners. Access, retrieval, and dissemination of geospatial data and services will be enabled by the UNSDI, avoiding duplication within the United Nations. For more information, see http://www.ungiwg.org/unsdi.htm

Standards

To facilitate data exchange between data users it is clearly necessary to coordinate development of geographic databases. In several countries, national geographic data committees have been formed for this purpose, which bring together the key persons in charge of spatial data development. In addition, supra-national organizations such as the European Umbrella Organization for Geographic Information (EUROGI), the Global Spatial Data Infrastructure (GSDI) Association, the International Organization for Standardization (ISO-TC/211) and the OpenGIS Consortium (OGC) are active in defining geographic data standards.

Collaboration

In the process of digital census mapping, the census organization may have the option of collaborating with other government agencies or with the private sector. Both options have been used successfully in different countries. As was mentioned previously, the national mapping agency is the natural first point of contact among government agencies. But other agencies (such as Cadastre, Environment, and Local Government) may also be able to contribute resources or have an interest in sharing the cost of creating a high-quality census database. Among private sector agencies, software and hardware vendors can support the technical side of the census mapping process, either under contract from the census office or in a cost sharing arrangement in which the private company will recoup its investment through the sale of spatially referenced census databases. It should be noted, however, that collaboration with other agencies is desirable but not mandatory. Since the census mapping agency must produce a map base for the census at a given time, it needs to avoid complete dependency on an outside supplier of data.

Any partnership or collaboration must be based on a shared intent and a well-defined agreement. The following elements of the cooperation agreement or letter of understanding need to be specified:

- Formalization: Is a loose collaboration sufficient or do the arrangements need to be highly formalized. A more formalized agreement will take considerable time to put in place, but can avoid later disagreements about rights and responsibilities concerning the development and use of data products. In most instances therefore, a formal, legally binding letter of understanding between the census office and the cooperating agency should be put in place that covers all relevant aspects of the partnership. Such formal contractual arrangements are mandatory when dealing with private sector suppliers of data or services.

- Scope of partnership: Collaborative agreements may cover simply the use of another agency's data, or they may involve the development of a large comprehensive spatial database from scratch.

- Responsibilities: Who will perform which tasks and functions. Issues that need to be addressed include data development, maintenance, data access, project supervision and resource use.

- Benefits: Clearly, the arrangement must be of benefit to all participants unless one agency simply purchases the services of another. It is useful to clarify how the different partners will gain from the arrangements in order to fairly divide tasks and responsibilities.

- Resource requirements: Resources include staff, computing environment, materials, and communication. Resources required for management and project supervision must also be considered.

- Cost sharing: Any direct and indirect costs connected to the activities of the partnership must be divided fairly. Accounting may not be straightforward since contributions can be in cash, data, labor, equipment use, or some other way.

- Cost recovery: If any revenues are generated from the distribution of the final products, they need to be shared with consideration of the costs that are incurred by managing and operating data distribution. This also involves a clear determination of agreed-upon uses and the copyright of the output products.

- Conflict resolution: In the event of disagreements during the course of the project, it is useful to specify a course of conflict resolution in advance.

Summary and conclusions

As we have established in this first chapter, national statistical organizations face a great opportunity to use new geospatial technology including the widespread availability of personal computers, hand-held devices, global positioning systems (GPS) and low-cost aerial and satellite imagery to obtain information about their populations. Planning for a geospatial census will require consideration of funding, staffing, and project management issues. For many NSOs, the task is to enlist new talent and re-organize the enterprise to accommodate it. This means recognizing the challenge of attracting and retaining trained personnel, and many countries are expressing their concerns with these issues.

NSOs require information about national spatial data infrastructure (NSDI) efforts, both for gaining access to valuable base data for use in census planning, and in sharing data and information after the census. The philosophy here is to move NSOs past the commitment phase into operational forward drive and begin with practical matters. NSOs will be able to the answers they need to develop a scalable geocentric digitization plan, scaled appropriately for the country, that will help you realize efficiency and effectiveness benefits. Remember that planning here is paramount. Chapter 2 will begin the process of creating a digital EA level census database for all phases of the census operation, covering the topics of geodatabase basics as well as factors in EA delineation including mechanics, data structures, and metadata.

Chapter 2: Constructing an enumeration area (EA) level database for the census

As Chapter 1 illustrated, one effect of embracing geospatial practices in a national statistical organization is to transform the enterprise around a geospatial information core, in other words for geography to become the pivot around which the NSO rotates. So we begin by tying the technical content of this upcoming chapter to the management material in Chapter 1, showing how the geographic census database becomes the focus of activities, where forms of census information are stored and accessed.

One main goal of the handbook is constructing an operational plan to build an enumeration area (EA) level database for the census, reflecting changes in how a census is conducted. A census is more than a data collection exercise, it is also an opportunity for member countries to develop information technology (IT) capabilities while also advancing development goals. Whereas before it was a one-off exercise done usually every ten years, maintaining an accurate geographic base for the census is now a continuous process.

Converting to digital geo-referenced operations will pay off handsomely in the future as a new generation of products becomes more available to the public. Working toward the goal of an accurate census through the embrace of geospatial technologies is a commendable public service. Because of the commitment required and the investment involved in the re-design of census mapping operations, the NSO can build off the work of one's predecessors and use maps they developed in prior years.

Chapter Two will introduce a concept of geographic coding that differs slightly from that conventionally understood by the GIS industry to mean address-matching. The UN definition of "geocoding" is more broad, representing the connection between statistical observations and real-world locations expressed in terms of latitude and longitude. While this chapter is advancing this new definition of geocoding, it is also emphasizing the importance of traditional (i.e. attribute) coding for a digital census. Simply put, geographic coding is the way that the data know where they are.

Geocoding is designed to cover a continuum of spatial scales from individual housing units to enumeration area level up to higher administrative or national levels, but successful use depends on the country having established a set of administrative areas with known territories and digital representation in the form of computer coding. The coding scheme must reflect one important quality of a geocoding scheme, that of scalability and flexibility, allowing both comprehensive coverage and room for growth.

Administrative hierarchies are based on the idea that inside the territory of the country there are boundaries that serve to demarcate territory at state or provincial and district levels, or for the purposes of voting or health monitoring or postal delivery. Together these various geographies can be stored in a database with the administrative level code and the number of units. For example, units at administrative level two (ADM 2s) are provinces, and units at administrative level three (ADM 3s) are districts. Ideally a GIS facility would have access to these units in GIS format for use in its various projects.

The chapter will then focus on standards and coordination as a management issue. National standards require coordination. If the statistical agency is the authority (i.e., the custodian of the

codes), then there is a need for the NSO to develop a transparent defensible plan. One must highlight, then, the need for regional coordination, especially in big countries.

It is very possible that a given country's NSO is responsible for administrative boundary definitions. However, if the NSO is not the custodian, then it will need to work with the existing geocoding plan within the government, possibly adapting its use for particular census needs, which are to allocate territory to individual census canvassers and crew leaders on the ground. Some regional coordination may be necessary as well.

Among the topics covered in this chapter are geodatabase basics, how a geography-based digital census differs from an analog census, administrative hierarchies; factors in EA delineation including mechanics; data structures; metadata; geographic coding and geocoding; and the process of replacing discrete steps with continuous maintenance, which means doing the census planning with an emphasis on identifying areas on the ground such as rapidly expanding cities that require special attention. Finally, the process flow from analog base maps as input to a seamless EA level geodatabase as the base data product of the census will be covered.

As always, plan carefully! In addition, it is necessary to emphasize again that with growing demand for small-area data, there is a need to assess country needs realistically.

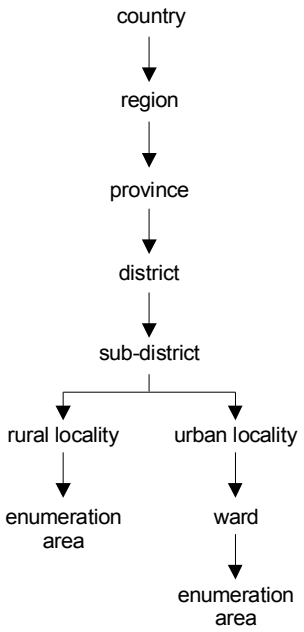
Definition of the national census geography

Administrative hierarchy

One of the earliest decisions in census planning pertains to the administrative areas for which census data will be reported. Administrative areas can be any special geographic unit, but mainly

they are units of administration; i.e., some governmental authority has jurisdiction over the territory. Census preparation involves creating a list of all administrative and statistical reporting units in the country, with the relationships among all types of administrative and reporting unit boundaries defined. Every country has its own specific administrative hierarchy, that is, a system by which the country and each lower level set of administrative units (except the lowest) are subdivided to form the next lower level. For example, for the purposes of the census a country may have been divided into seven hierarchical levels in urban areas and six in rural areas:

Figure 2.1 A census geographic hierarchy

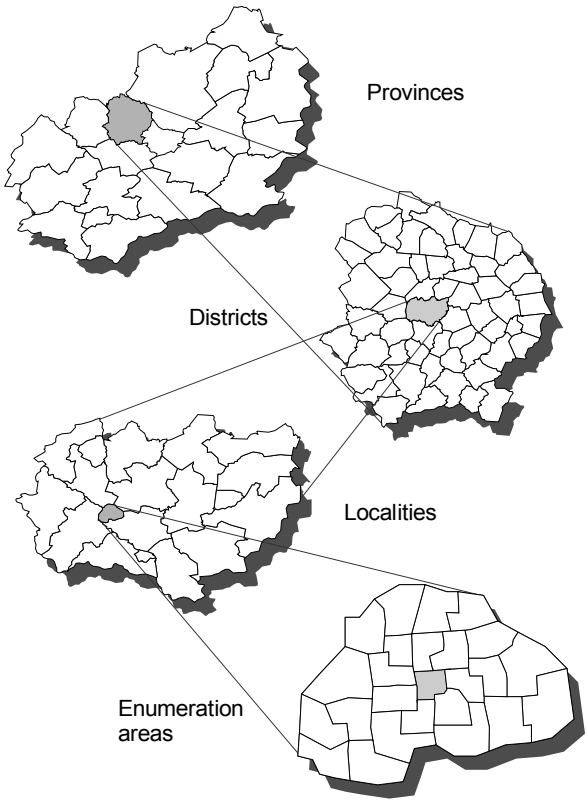


Only some of these hierarchical levels may have actual administrative roles; for example, the province, district and locality levels may have capitals with local government offices that are responsible for those regions. Figure 2.2 illustrates the nesting of administrative and census units using a simple example with only four hierarchical levels. In some instances, however, administrative units may not be completely nested. Especially when considering both

administrative and other statistical reporting units, the census office may need to deal with a very complex system of geographic regions.

Not all levels must be equally important. For example, many countries divide the territory into major regions, which are often geographically defined, such as *North-South-Southwest-East*, or *Mountain-Plains-Coastal*. These regions often do not serve any administrative function, but might still be used to report statistical information.

Figure 2.2: Illustration of a nested administrative hierarchy



Relationship between administrative and other statistical reporting or management units

In addition to administrative units, most countries will have a number of other sets of areas that are used for different purposes and for which census data will need to be compiled. These areas have special uses and may have been defined at a rough scale. Some examples are:

- health regions;
- labor market areas;
- electoral districts;
- postal zones;
- cultural or tribal areas;
- urban agglomeration or metropolitan areas;
- agricultural or economic census units;
- land titling or cadastral units; and
- utility zones (water or electricity supply districts).

More can be learned about these special geographies through interaction with other GIS users in your country or through spatial data infrastructure activities. In their spatial rendering, some areas may not nest perfectly into the administrative hierarchy of the country. In designing enumeration areas, the census mapping agency should consider these reporting units as much as possible in order to facilitate tabulation of census data for these regions. The user requirements

analysis carried out in the census planning stages should provide guidelines as to which non-administrative areas will receive the most consideration. Generally, to guide enumeration area design the census mapping agency should divide all sets of areas into those where compatibility is mandatory, desirable or unlikely, and consider them accordingly.

For some reporting or management zones in the country, digital boundary data may already have been produced by the responsible agencies. For instance, a number of countries that have initiated land reform programs are using GIS to manage land titling databases (cadastral information) and many national postal organizations are using GIS databases of postal codes to facilitate mail delivery. Where digital databases of such units are available, they can support the development of census geographic databases. Where a high degree of compatibility can be achieved, this has the added advantage that statistics for other zones, for example, water demand or voting results, can be combined more easily with demographic and social statistics.

Within the statistical office, other census operations also require the definition of data collection units. Most importantly agricultural and economic censuses are carried out regularly in many countries. Many analytical applications benefit from the joint analysis or population census information with agricultural or economic data. A high level of agreement between the geographic units used to compile these types of data will greatly increase their utility in public and non-governmental applications.

Coding

A digital geographic database in vector format consists of a structured set of points, lines and polygons. Each geographic feature – i.e., each point, line, or area – has a unique identifier which is used by the system internally. This internal identifier is not usually accessible by the user and

should not be modified externally. Needed is a more meaningful identifier that can be used to link the geographic features to the attributes recorded for them. For the enumeration areas and administrative units this link is the unique EA or administrative identifier which is listed in the master file of all geographic areas relevant in the census.

How this identifier is entered is again software-specific. It can be added during the digitizing process by entering the identifier before the feature is digitized. Or it can be added at a later stage by selecting the feature interactively and adding the identifier through a menu interface. For polygon features, some systems require the user to add a label point that is contained in each area unit. While conceptually simple, coding may require considerable time and resources.

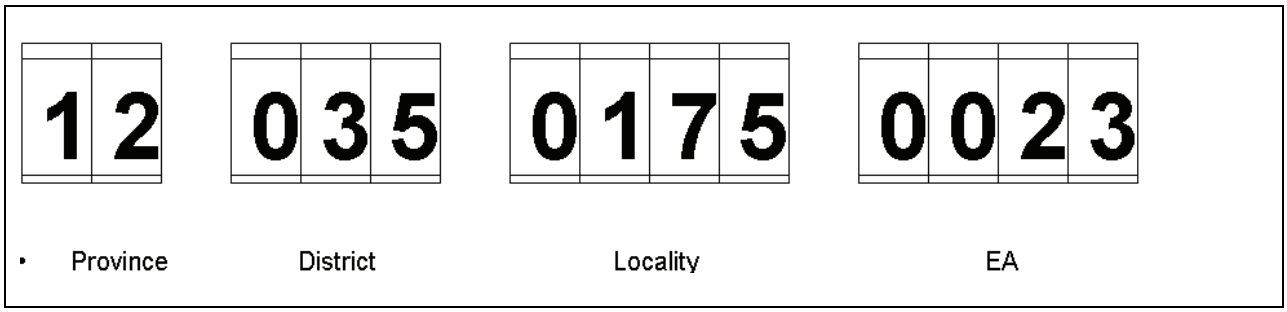
Indeed, a unique code needs to be assigned to each enumeration area. This code is used in data processing to compile enumerated information for households in each EA and to aggregate this information for administrative or statistical zones for publication. This is the numeric code that provides the link between the aggregated census data and the digital EA boundary database stored in a GIS. Ideally the coding scheme needs to be determined on a country by country basis. However, the rules used to assign codes need to be unambiguous and should be designed in collaboration within the NSO, especially between the mapping/geospatial data and those managing the data core. The most important principles when designing a coding scheme are flexibility, intuitiveness and compatibility with other coding schemes in use in the country. The statistical office is often the custodian of coding schemes in the country and should also be the focal point for the design of the census mapping codes.

A hierarchical coding scheme will usually facilitate consistency and clarity of the numeric identifiers. In this approach, geographic units are numbered at each level of the administrative

hierarchy – usually leaving gaps between the numbers to allow for future insertion of newly created zones at that level. For example, at the province level, units may be numbered 5, 10, 15 and so on. A similar scheme would be used for lower level administrative units and for enumeration areas. Since there are often, for example, more districts in a province than provinces in a country, more digits may be required at lower levels. The unique identifier for each smallest level unit – i.e., the enumeration area – then consists simply of the concatenated identifiers of the administrative units into which it falls.

For example, a small country could use the following coding scheme:

Figure 2.3 An enumeration area coding scheme



An EA code of 12 035 0175 0023 means that enumeration area number 23 is located in province 12, district 35 and locality 175. The unique code is stored in the database as a long integer or as a thirteen character string variable. Obviously, the variable type needs to be the same in the census database and in the geographic database. Storage as an integer variable has the advantage that subsets of records can be selected easily using standard database query commands in any database management system or GIS package. For example, the following query will find all enumeration areas within locality number 175 in the database or on the digital map:

SELECT ID > 1203501750000 AND

ID < 1203501760000

Storage of the code as a character variable, on the other hand, can improve consistency, for example through the use of leading zeros. In this case, the code is considered a name made up of ascii characters rather than a sequential number.

In cases where administrative and reporting units are not hierarchical, special coding conventions need to be developed. In any case, it is very important to be completely consistent in defining and using the administrative unit identifiers, since they are the link between the GIS boundaries and the tabular census data. The census office should therefore maintain a master list of EA and administrative units and their respective codes and commit any changes made to the master list to the GIS and census databases. The NSO might also consider publishing a list of EAs including relevant coding, along with the latitudes and longitudes of centroids (central points) and perhaps populations enumerated as well.

Delineation of enumeration areas (EAs)

Enumeration areas are the operational geographic units for the collection of census data and are defined early in the census process. The delineation of enumeration areas is similar whether manual or digital cartographic techniques are used. The design of enumeration areas should take various criteria into account. They should:

- Be mutually exclusive (non-overlapping) and exhaustive (cover the entire country);
- Have boundaries that are easily identifiable on the ground;
- Address the needs of government departments and other data users;

- Be consistent with the administrative hierarchy;
- Be useful also for other types of censuses and data collection activities;
- Be compact without pockets or disjoint sections;
- Be of approximately equal population size;
- Be small enough and accessible to be covered by an enumerator within the census period;
- Be small and flexible enough to allow the widest range of tabulations for different statistical reporting units; and
- Be large enough to guarantee data privacy.

Among these criteria are some that facilitate census data collection, while others pertain to the usefulness of EAs in producing output products — i.e., the relationship between data collection and tabulation units. Bear in mind that the purpose of a census is to produce useful data for administrators, policy makers and other census data users. Maximum flexibility and suitability for producing the best possible output products thus has precedence over convenience of census enumeration.

The size of enumeration areas can be defined in two ways: by area or by population. For census mapping, population size is the more important criterion, but surface area and accessibility also have to be taken into account to ensure that an enumerator can service an EA within the time allotted. The chosen population size varies from country to country and is determined based generally on pretest results. Average population size may also vary between rural and urban areas, since enumeration can proceed more quickly in towns and cities than in the countryside.

Under special circumstances, enumeration areas that are larger or smaller than average may have to be defined. For most practical purposes, the population size of an enumeration area will be in the low to mid hundreds.

Before delineation of EA boundaries, the number of persons living in an area and their geographic distribution need to be estimated. Unless there is information from a recent survey, a registration system or some other information source, these numbers need to be determined by counting the housing units, determining the associated number of households and multiplying by an average household size. The number of housing units can be determined through cartographic field work, cooperation with government officials, extrapolation from previous census results, or, by means of aerial photographs or satellite imagery as discussed in the next chapter.

Enumeration area boundaries need to be clearly observable on the ground. Even if they do not have considerable geographical training, all enumerators need to be able to find the boundaries of the area for which they are responsible. Thus, population sizes between enumeration areas may be varied in order to produce an easily identifiable delineation. Natural features that can be used for this purpose are roads, railroads, creeks and rivers, lakes, fences, or any other feature that defines a sharp boundary. Features with more gradual edges, such as brush, forests or elevation contours such as ridges, are less ideal. In some instances, it is unavoidable to use EA boundaries that are not clearly visible on the ground. In this case, an exact textual description and appropriate annotation on the EA maps is required. Examples are offset lines and extended lines. For example, an EA boundary may run parallel to a specific road at a clearly defined offset. Or a portion of an EA boundary may be defined as the extension of a clearly visible road to another clearly defined feature such as a river or railroad.

Specific issues related to EA delineation will be encountered in many countries. For instance, while villages may be assigned to specific administrative units, the actual boundary delineating the village area may not be defined. Also, special populations, such as transient, nomad or military personnel, need to be assigned a geographic reference. For instance, naval personnel is often assigned to their home ports. Be aware that when planning for locating hard-to-enumerate populations that operational costs are sometimes 10 to 20 times higher than for residential populations in urbanized areas.

Among the criteria for EA delineation is determining the ideal EA size, which is based upon the number of people one enumerator can count in the time period scheduled for data collection. The plan for EA delineation should reflect the overall census plan, stemming from the number of days allotted for enumeration. A census pretest can determine the number of housing units (HUs) that an enumerator can cover per day. For instance, if 16 HUs can be enumerated per day in urban areas but only ten per day in rural areas, and if the period of enumeration is 10 days, then the ideal urban EA would contain 160 HUs and the rural EA would contain 100 HUs. If the average number of persons in an HU is five, then the ideal population size would be 800 for an urban area and 500 for a rural area. Other factors influencing the size of an EA include administrative area boundaries, the visibility of EA boundary features, and the mode and availability of transportation.

Population estimates are essential for proper EA delineation. Local officials can be called upon to provide small-area estimates, or the areas in question can be visited by NSO field personnel. In areas that have not experienced dramatic change, estimates can be adjusted from the previous census based on the time elapsed.

Delineation of supervisory (crew leader) areas

After delineation of EAs, the design of supervisory maps is usually straightforward. Supervisory areas consist of groups of usually eight to twelve contiguous EAs which share some of the same characteristics as enumeration areas. The EAs assigned to the same supervisory area must be compact to minimize travel times and of approximately equal size. They should be included in the same field office area, which usually is defined according to administrative units.

Depending on the size of the country, additional levels of census management areas can be designed. In larger countries these will often coincide with the provincial or regional statistical offices.

Components of a census database

A comprehensive census GIS database consists of a digital map of census enumeration areas and, in most instances, a series of base map layers that provide the context and orientation in the final enumerator maps. Base data layers might be roads, rivers, buildings, or settlements. Each of these will be contained in a separate GIS database. So, roads and rivers, for instance, are both represented as lines, however they will not be stored in the same digital file. Above all features need to be identifiable on EA maps in order to orient enumerators.

Before starting data entry and data conversion, the census cartographic staff should design the structure of all GIS data sets that will be produced. The structural definition will be a detailed description of all conventions and guidelines that the cartographic staff need to follow to ensure consistency of the final output products. Good planning will avoid confusion and incompatibilities later in the process.

The first step is to think about what the final products will look like. The complete digital enumeration area database, for example, will likely consist of the following components:

- The spatial *boundary database* consisting of area features (polygons) that represent the census units;

- The *geographic attributes table*: a database file which is linked internally to the spatial database and contains one record for each polygon. This table contains the unique identifier for each census unit and possibly some additional static, i.e., unchanging, variables such as the unit's area in square kilometers; and

- The *census data tables containing non-spatial attributes*—i.e., the census indicators for the spatial census units. Each of these files must contain the unique identifier of the census unit which provides the link to the corresponding polygon attribute table records. There will be one record for each census unit.

- Additionally, other vector (point or area) features such as landmarks, roads, waterways, schools, hospitals, or other buildings may be useful for orienting field workers during the enumeration. Such features that are recorded during the time of preliminary field canvassing or house listing may prove useful later to other government agencies or nongovernmental organizations, saving time and money.

Consistency with past censuses

A census provides a cross-sectional view of the size and characteristics of the population of a country. One of the most important uses of a census is to analyze changes in the composition of the population over time. This change analysis is often done at fairly aggregate levels only, for

example, at the national or provincial level. However, changes in local areas are equally important, since dynamics in small areas affect local planning decisions. Change analysis at the local level is greatly facilitated if the units of enumeration remain compatible between censuses. Although statistical or other techniques exist that reconcile information for incompatible area units (as described in a later chapter), such short-cuts may introduce errors in subsequent analysis. Most census data users lack the expertise and tools to do such interpolations. The problem of changing the geographic base between censuses is no less serious than changes of definitions of items on the census questionnaire.

In designing the census geography, the census office should therefore attempt, in as much as possible, to preserve boundaries from the previous census. Due to increases in population size, new EAs may have to be defined. In these cases it is always preferable to subdivide an existing EA rather than to change the existing boundaries. An analyst can simply aggregate a subdivided enumeration zone to make the new census data compatible with the information from a previous enumeration. If boundaries are changed, more complicated methods of adjustment are necessary.

One component of EA delineation that can facilitate change analysis is the compilation of compatibility or equivalency files. These list the codes of each enumeration area in the current census and the corresponding code in a previous enumeration. If units were split or aggregated, this is indicated in the files.

The boundary database and geographic attribute table are tightly linked – essentially they represent one data set. During census planning some basic census related information such as housing unit or population estimates and documentation information will be compiled for each enumeration area. This external information about the census units will be stored in separate data

tables in a generic database management system. From there it can be linked as needed to the boundary data through the common identifier – the EA code – in the geographic attributes table. Similarly, after completion of the census, the census information is stored separately in a database management system. To create thematic maps of census results, the boundary and census data are then linked via the unique identifiers in the polygon attribute table. Clearly, to ensure that the census databases that are the product of the data entry and tabulation program will match the geographic boundary files, close cooperation between the census cartographic and data processing sections is required.

Typically, separate databases will be developed for each administrative level or set of statistical areas for which census data are published. When boundaries at any level are updated, the changes will, of course, have to be made also in all other databases that contain these boundaries. The best approach is to make all changes in the master boundary database at the lowest aggregation level (i.e., the EA level database) and to produce each higher level administrative or statistical unit database using standard GIS and database aggregation functions.

Some of the base data layers may be much simpler than the digital census enumeration area map. For example, for a roads database, only a few attributes – name or identifier of the road if available, surface type, and number of lanes – might be collected. In this case it might not be necessary to store the descriptive attribute information in a separate table. For simplicity all attributes can be contained in the geographic attribute table itself.

At certain stages in-between and during census cycles, benchmark data sets should be created. For instance, there should be a unique version of the country's census map database that matches each data collection effort or related statistical application. Separate aggregated boundary data

sets can be produced for each statistical reporting unit for which data are required. These benchmark data sets should be time-stamped and permanently archived. Thus, benchmark data sets created from the same master database may exist for a census in 2010, for a large survey in 2012 and for an election in 2015.

The development of the digital census database will be based on two data sources: the conversion and integration of existing map products that may be in hardcopy or digital form, and the collection of additional data using field work, air photos or satellite images. Collectively, we will use the term *data conversion* to refer to these steps.

The best strategy for data conversion depends on many factors including data availability and time and resource constraints. There will always be a trade-off between the cost of a project, the amount of time required to complete data conversion, and the quality of the final product (Figure 2.4). It is usually only possible two optimize two of the three objectives at the expense of the third. For example, it is possible to create a high quality database quickly, but this will be expensive. One can produce good data cheaply, but this will take a long time. Or, one could develop a database quickly and cheaply, but the quality of the resulting product will be low.

Figure 2.4: Trade-offs in the data conversion process

(after Hohl 1998)

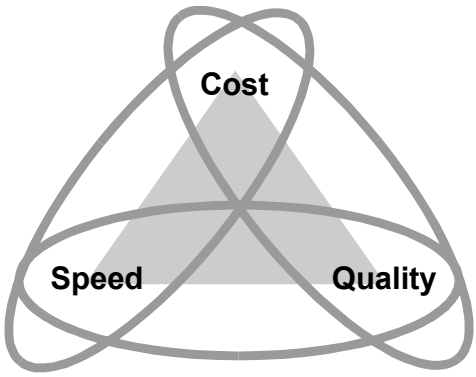
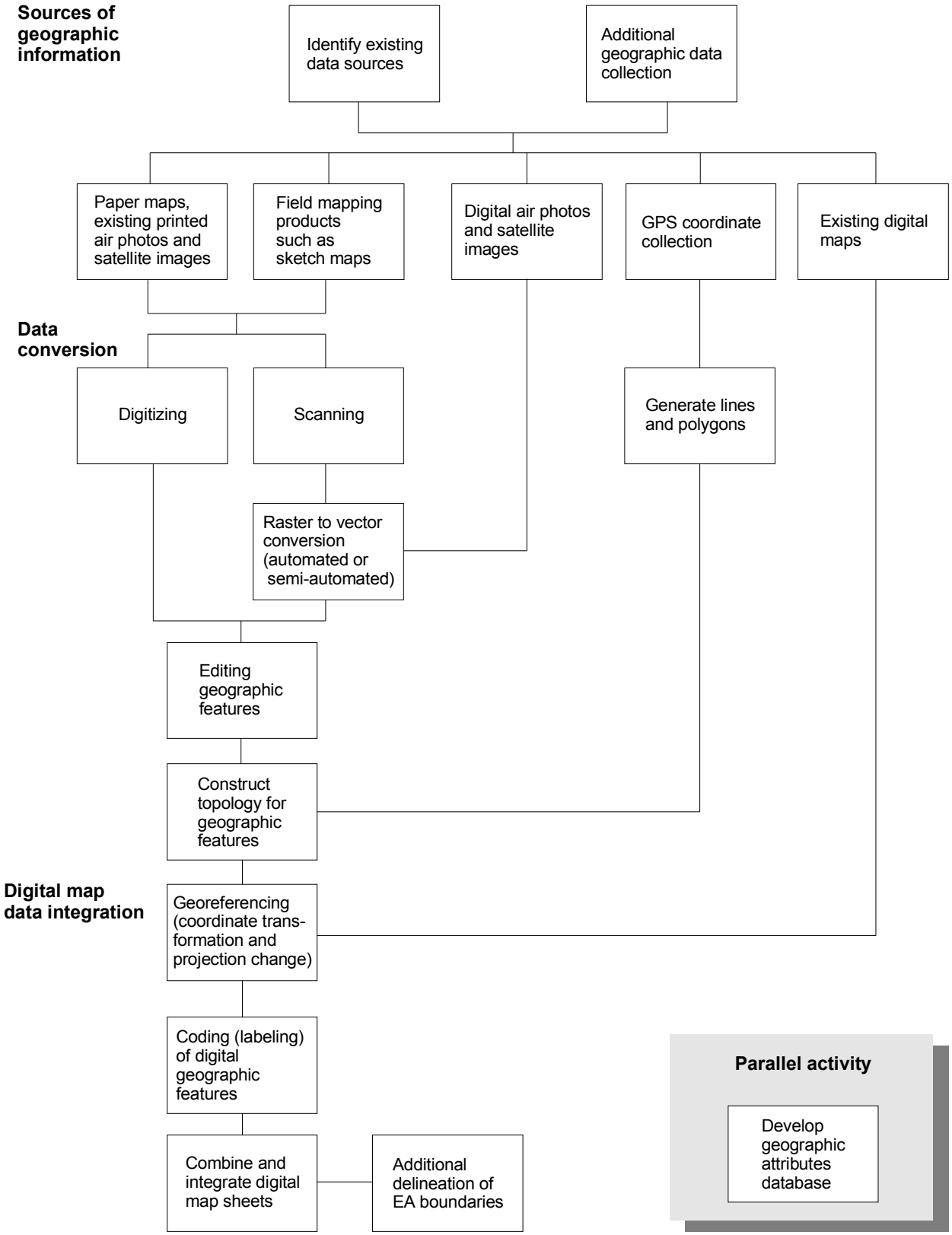


Figure 2.5 outlines the basic steps in the data conversion process that leads to a complete digital census database. A survey of existing digital and hardcopy sources will lead to the identification of data gaps. Existing maps may be outdated, or the scale of available topographic maps may be insufficient for census purposes. For any areas for which existing materials are of insufficient quality, a strategy for field mapping or some other data collection approach must be developed.

Boundaries and point locations of geographic features required for the census — building and village locations, road infrastructure, rivers and any other information used to delineate enumeration areas — must be delineated digitally from published paper maps, sketch maps, printed air photos or satellite images. This is accomplished by scanning with subsequent image to vector conversion, or by digitizing — tracing the features with a mouse-like cursor. Although scanning and digitizing technology is continuously improving, this is still the most tedious part of a data conversion process. Data capture is followed by an editing step, the construction of GIS database topology, and referencing of all coordinates in a proper cartographic map projection (this step can sometimes be integrated with digitizing activities).

Figure 2.5: Stages in the census GIS database development



At the same time, existing digital databases, for example products created by another government agency, and coordinates collected in the field using global positioning systems (GPS) can be imported into a GIS. GPS coordinates may have to be converted from point locations to lines and boundaries that show linear and polygon features such as roads or city blocks. After attaching attribute codes to all database features, digital map sheets that were developed separately can be joined to create a seamless database for an entire region. The completed database will — depending on the scope of mapping activities — show major physical features, landmarks, infrastructure, settlements and individual buildings. Based on this information, census staff can delineate enumeration areas interactively using the geographic reference information as a backdrop.

GPS as a data source for mapping and evaluating EAs will be covered in more detail in Chapter 3, and in one of the Annexes as well. Additional material will be included on using GPS and remotely sensed data to correct the geographic database.

As a parallel activity during the entire data development process, census staff should maintain a list of all administrative and enumeration areas that are delineated in the database, including names, geographic level, and a locational reference. This computerized list is the geographic attributes table and will be linked to the completed GIS database.

The flowchart in Figure 2.5 shows only one of many possible sequences in the data conversion. EA boundaries, in particular, can be delineated at several points during the process. For instance, scanned and properly georeferenced air photos, or high-resolution satellite images, show enough detail that an operator can delineate digital EA boundaries on the screen using the air photos as a backdrop. EA boundaries can also be hand-drawn on suitable paper maps and digitized together

with other information from those hardcopy sources. Other steps may also be performed in a different sequence. For instance, most GIS packages support georeferencing at the beginning of the digitizing process, therefore making an extra step at a later stage unnecessary.

No matter which process is chosen, the census office should evaluate the feasibility of the approach by carrying out a pilot study. This typically involves a test of the methodology on a small sample area. The pilot study will allow to identify problems early on, so that the technology and procedures can be fine-tuned, modified or, in the worst case, abandoned. Information from the pilot tests will also aid scheduling and budgeting activities as they allow a better evaluation of staffing and equipment requirements, and the time required to perform all activities.

The pilot area should be representative for as many regions of the country as possible. In other words, it should include a high degree of variation, covering rural as well as urban areas, regions with characteristic settlement patterns, agricultural lands and zones of dense vegetation or other features that inhibit field data collection.

GIS software and equipment vendors will often be willing to assist in a pilot study, since they hope to benefit from sale of their products if they prove suitable for the census mapping project. Vendors will also provide benchmarking data, which is important for high-capacity applications such as high-volume map production and database access. Some techniques can be easily tested on a part of a country's territory. For instance, global positioning receivers are inexpensive and census staff can carry out evaluations of field data techniques. It may, however, be too expensive to obtain digital air photos for a small pilot test site. In this case, older products or sample air photos for a country in which conditions are similar could be obtained.

Cartographic data sources for EA mapping (secondary data acquisition)

Types of maps required

In nearly all cases, a census cartographic program will have to consult existing hardcopy maps for the production of a digital cartographic database or for updating an existing GIS database. The census geography staff need to obtain all up-to-date maps for the nations' territory, including the following types of maps:

- National overview maps usually at scales between 1:250,000 and 1:5,000,000 depending on the size of the country. These maps should show major civil divisions, the location of urban areas, and major physical features such as important roads, rivers, lakes, elevation, and special points of reference. These maps are used for planning purposes;

- Topographic maps at large and medium cartographic scales. The availability of maps at these scales will vary by country. While some countries have complete coverage at 1:25,000 or 1:50,000, the largest complete map series in others is only 1:100,000 or 1:250,000 scale;

- Town and city maps at large cartographic scales, showing roads, city blocks, parks, etc. These can be at a variety of scales ranging from 1:5,000 to 1:20,000 and be of various vintage including perhaps dated colonial maps and urban plans;

- Maps of administrative units at all levels of civil division; and

- Thematic maps showing population distribution for previous census dates, or any features that may be useful for census mapping.

For incorporation in a GIS database, ideally these maps should all have comprehensive documentation. This includes the geographic referencing information, including the map scale, projection and geographic datum, the map compilation date, compiling agency and complete legend. However, even maps that are not properly georeferenced are useful if they show information relevant to census mapping — especially when they can be easily scanned and brought into a project as an on-screen traceable layer through what has become the new definition of "heads-up digitizing." In such cases, the benefits of additional information will often outweigh the resources required to integrate such data into the census GIS database and the accuracy problems associated with any such product.

Inventory of existing sources

All maps that have been obtained should be well documented and organized according to the organization of the census mapping program — i.e., by census region or district. In addition to hardcopy map sources, digital map sources will increasingly become available from many sources. Digital maps, of course, have the advantage that they can be more easily manipulated and adapted to the purposes of census mapping. However, this is not always completely straightforward. If documentation is absent, it is often not possible to determine the correct projection information and data quality is difficult to evaluate.

Through activities sponsored by a national spatial data infrastructure (NSDI) or through person-to-person contacts, the following agencies and institutions can be contacted to see whether they can contribute useful hardcopy or digital maps:

- National geographic institute / mapping agency. This is the lead agency in the country concerned with mapping. However, in some countries the mapping agency is lacking the

resources necessary to produce topographic maps at large cartographic scales or to convert maps into digital databases;

- Military mapping services. In some countries, the main mapping organization is part of the military. Military mapping organizations are often strong in aerial photography and in the interpretation of remotely sensed data;

- Province, district and municipal governments. Local government organizations increasingly use GIS to manage information about transportation, social services, utility services and planning relevant information;

- Various government or private organizations dealing with spatial data:

  - Geological or hydrological survey;

  - Environmental protection authority;

  - Transport authority;

  - Utility and communication sector companies; and

  - Land titling agencies.

- Donor activities. Project level activities by multinational or bilateral aid organizations sometimes include mapping components. Such projects often have the means to purchase and analyze remotely sensed data or aerial photographs, which can be of great use to the mapping agency.

Importing existing digital data

Direct import of digital data is in most cases the easiest form of digital spatial data conversion. The GIS industry has transitioned from standalone tables, most often in DBASE (.dbf) format, to relational databases such as Oracle or Microsoft Access, or through personal or file geodatabases. Data transfer relies on the exchange of data in mostly proprietary file formats using the import/export functions of commercial GIS packages.

Textbox : choices on commercial off the shelf (COTS) software.

Selection options commercial GIS software

Given the plethora of commercial suppliers, NSOs should evaluate their operational goals and adopt technologies accordingly. Software interoperability may be important to suit both current and future needs. Consult with different agencies and staff on the preferred software platform. An informed choice is the best choice.

COTS (Commercial Off the Shelf) programs can be divided into those offering raster/vector integration and primarily image analysis software.

Raster/Vector Integration

Programs include Environmental Systems Research Institute (ESRI)'s ArcGIS, Intergraph's Geomedia, IDRISI, Maptitude, GRASS, Pitney Bowes' MapInfo, AutoCAD, and Microstation.

ESRI is the market leader and offers extensive format support, flexible functionality for different users, an extensive analysis toolset, database management of various data types, extensive support, training, and an impressive knowledge base.

GeoMedia supports many data types, offers an extensive analysis toolset, and direct access to major geospatial/CAD data formats. GeoMedia incorporates industry-standard relational databases, and is regularly updated with full support and training.

IDRISI is a single packaged product and provides open code for customization. It offers sophisticated raster based analysis, a raster analysis toolset, and extensive vector-based data input and output through CartaLinx.

Maptitude handles both CAD and GIS, has import/export options, offers an address matching feature, and suited for readily available datasets.

Geographic Resources Analysis Support System (GRASS) provides raster/vector capabilities, and has a new topological 2D/3D vector engine and support for vector network analysis. GRASS was the first UNIX GIS, and it offers UNIX functionality, open code, and a global user base.

MapInfo (Pitney Bowes software) provides mapping functionality, but with limited GIS functionality. It traditionally uses Visual Basic, and flexible applications. It has a global base and a strong dissemination tool.

AutoCAD Map offers sophisticated GIS/CAD integration, as well as grid, projection, and topology support. It provides extensive database options, and vector/raster display and analysis. It is menu-based, but costly.

Microstation has largely US and European users. It features CAD-based mapping, but with limited analysis features and limited data format integration and georeferencing. It does offer a simple Google Earth/Google SketchUp interface.

Image Analysis

Raster-based analysis is becoming commonplace in census mapping operations. Many image analysis software can now integrate vector data for refined image analysis. As imagery becomes less expensive

and faster to process and integrate with existing vector datasets, more NSOs will buy image analysis licenses.

Major platforms include Leica Geosystems Imagine, Geomatica (PCI Geomatics), ENVI (ITT visuals), Definiens Professional, and Google Earth.

Leica Geosystems Imagine is raster-based software designed to extract information from images. It handles extensive collections of geospatial data, and provides client-side interaction with spatially aware databases. Imagine extends functionality to topological editing of spatial databases, and offers vector layer support.

Geomatica (PCI Geomatics) handles extensive collections of geospatial data formats, and provides client-side interaction with spatially aware databases, full map production capabilities, attribute management tools for viewing, editing, querying and analysis of attributes, and extensive image processing capabilities.

ENVI 4.3 (ITT Solutions) integrates raster imagery with geographic information systems. Vector layers can be overlain onto image data to easily compare raster and vector information. ENVI offers relatively extensive support and a broad knowledge base. In addition, it handles numerous vector formats (including ArcView shapefiles, ARC/INFO interchange, DXF, Microstation/Intergraph DGN files, USGS DLG files and more). Its linear feature extractor automatically digitizes everything between the seed points, faithfully following curves, jumping gaps, and snapping vectors if desired.

Definiens offers highly sophisticated automated user-defined feature extraction, allowing geo-information to be extracted from any kind of remote sensing imagery. It offers data management connectivity to ArcGIS server, loading and saving vector data to and from databases, and allowing for simultaneous updating of different locations of one large dataset. In addition, one can update portions of large vector datasets. An extension for ArcGIS allows ArcCatalog to define a collection of maps as a Definiens' workspace, enabling users to review and edit entries.

GoogleEarth's strongest selling point is its free imagery. However there are costs associated with feature upgrades such as GoogleEarth Pro. Google Earth offers global coverage, democratizing access for

many users. It is user friendly, has a large knowledge base, and it is useful for quick display of raster and vector data.

Some issues to consider include initial cost of the software as well as maintenance and upgrades, the LAN configuration, training needs, ease of installation, maintenance, documentation and manuals, help-line and vendor support, means of making patches for support, and workforce.

Free and Open Source Software (FOSS) for Desktop Mapping

An alternative to commercial software is free and open source (FOSS) geographic information systems software. FOSS offers a low- or no-cost approach to GIS. FOSS programs are freely downloadable from the internet and provide similar functionality to many commercial programs. FOSS implies that users can access the source code of the application, meaning that NSOs with programming expertise can tailor programs to suit their specific needs. FOSS software is becoming more user friendly, providing the capability for tailored applications for specific uses. The experience of trying FOSS software may prove valuable even if the NSO eventually decides to adopt a commercial product. FOSS desktop mapping programs have traditionally been used by programmers or others with IT experience. This fortunately has changed due to a larger user base and subsequent product development. Programs have become more user friendly, with online training and product support. FOSS programs offer *interoperability*, defined by the Global Spatial Data Infrastructure Association as "The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units."

FOSS programs include:

Quantum GIS (http://qgis.org) QGIS is the most advanced raster/vector FOSS software, with users on six continents and a rapidly evolving suite of features. QGIS offers Grass for digitizing and editing of vector features. Multiple annual software updates keep users ahead of product developments. A strong user community has online support for many technical issues.

Thuban (http://thuban.intevation.org) Thuban is another example of FOSS software with a large user base and online product support. Thuban is implemented using Python and is multi-platform. It can handle both vector and raster data and offers full GIS functionality including object ID and annotation, a legend editor and classification, table queries and joins, projection support, and multi-language support.

Open EV (http://openev.sourceforge.net OpenEV is both a software library and application for viewing and analyzing raster and vector geospatial data. OpenEV can support both 2D and 3D display and can reproject on the fly. It provides powerful image analysis. The OpenEV virtual user community makes use of discussion lists and how-to tips to include novice users.

All software systems provide links to other formats, but the number and functionality of import routines vary among the packages. Problems often occur because software developers are reluctant to publish the exact file formats that their systems use. Competitors then use some form of reverse engineering to figure out the exact file formats to enable their customers to import external files. Consequently, import routines are sometimes unstable and frequently lose some of the information contained in the original data files. In some instances it may be better to go through a third data format, instead of attempting to import another package's exchange file directly. For instance, Autocad's DXF format is supported by most GIS packages and is well documented. DXF export and import functions of other commercial packages are therefore usually quite reliable.

Problems can be reduced if the census GIS or cartographic agency employs a widely used, comprehensive commercial GIS package. High-end systems are more likely to provide import functions for a large number of exchange formats. And it is also more likely that other data producers will be able to provide GIS data in the native format of the GIS package. Import

capabilities are one important criterion for choosing GIS software. Another option is to use a third-party conversion package.

Apart from problems in converting the data files from one format into another, the most often encountered difficulty in using existing digital data is insufficient or absent metadata. Without such information it is difficult to assess the quality of the digital information. Even worse, missing information about the geographic reference framework might make it impossible to convert data from the external data set's coordinate system to the one used by the census organization. Similarly, a missing code book or data dictionary will make it difficult to interpret the geographic and data attributes included in the GIS data sets attribute tables. When data are procured from external sources, the census office should therefore always insist that extensive documentation is provided.

Other possible problems that may need to be addressed include differences in definitions and coding schemes, use of different cartographic reference systems, incompatible spatial scales, and varying accuracy standards which may result in features that should match across two databases to be displaced. Addressing these problems in order to make full use of existing digital maps may require considerable processing and editing.

Geographic data conversion: analog to digital

The process of converting features that are visible on a hardcopy map into digital point, line, polygon and attribute information is called data automation or data conversion. In many GIS projects this step continues to require considerable resources, especially time.

The conversion of hardcopy maps or information from printed aerial photographs or remote sensing images into a digital GIS database involves a series of steps. Although the sequence of steps may vary, the required procedures are similar in each case. After selected point and line features on the map have been converted into digital coordinates in the computer, there is often an additional amount of editing required to deal with any remaining errors or omissions. Following this step, the map coordinates which are initially recorded in units used by the digitizer or scanner need to be converted to the real world coordinates corresponding to the source map's cartographic projection. Some systems allow the determination of the projection prior to digitizing. In this case, the coordinates are converted on the fly during the digitizing process. The end result, of course, is the same.

The next step is to attach consistent codes to the digitized features. For example, each line representing a road would obtain a code that refers to the road status (dirt road, one lane road, two lane highway, etc.) or a unique code that can be linked, for example, to a list of street names. In higher end GIS software packages, this step is followed by the structuring of the database, also called building of topology. In this step, the GIS determines relationships between features in the database. For example, for a roads database, the system will determine intersections between two or more roads and will create nodes at these intersections. For polygon data, the system will determine which lines define the border of each polygon. After the completed digital database has been verified to be error-free, the final step is to add additional attributes. These can be linked to the database permanently, or the additional information about each database feature can be stored in separate files which are linked to the geographic database as needed.

The two main approaches for converting information on hardcopy maps to digital data are scanning and digitizing. Scanning is the automatic process of converting a map into a digital

raster image which can subsequently be converted into digital line work. Digitizing, in contrast, involves the tracing of all required point and line features on a map using a cursor or mouse. On-screen digitizing techniques are used to draw new map layers using scanned maps or images, and content can also be digitized from marked-up map sheets. The two approaches are now discussed in more detail.

Scanning

For many data input tasks, scanning has arguably bypassed digitizing as the main method of spatial data input, mainly because of the potential to automate some tedious data-input steps using large-format feed scanners and interactive vectorization software. There are different types of scanners, but all work basically in the same way. The map is placed upside-down onto the scanning surface where light is directed at the map at an angle. A photosensitive device records the intensity of light reflected for each cell or pixel in a very fine raster grid. In gray scale mode, the light intensity is converted directly into a numeric value, for example into a number between 0 (black) and 255 (white). In binary mode, the light intensity is converted into white or black (0/1) cell values according to a threshold light intensity. In color scanners the light sensitive device is divided into three portions that are sensitive to red, green and blue respectively. The relative intensity of the three color signals, when combined, determine the pixel color. The result of the scanning process is a raster image of the original map which can be stored in a standard image format such as GIF or TIFF. After georeferencing the image — this involves specifying the coordinates of an image corner and the pixel size both in real-world units — it can be displayed in many GIS packages as a backdrop to existing vector data. Usually, however, geographic features from the image are extracted either manually or automatically and converted to vector data.

There are three basic types of scanners in common use:

- **Feed scanners** are now the most commonly used scanner type for large-scale GIS applications. In feed scanners the sensor system is static. Instead, the map is moved across a sensor array. Their accuracy is lower than that of drum scanners, since the map feed can be less precisely controlled than the scanner movement. But their accuracy is usually sufficient for GIS applications, their cost is lower and they typically produce images in less than five minutes. A caveat is that older or fragile documents might be damaged by the feed scanner's rollers. See Figure 2.6.

- **Flat bed or desktop scanners** are found in many offices. They are of relatively small format so that larger maps must be scanned in several parts and joined in the computer. The document is placed upside down on a glass plate and the camera and light source move along the document beneath the glass. The strength of flat bed scanners is their low cost and easy setup and maintenance. They are useful for scanning text documents—for example data tables—which are later interpreted using optical character recognition (OCR) software. They also provide a means to bring small graphics and maps into a computer. They are less suitable for large-scale map conversion tasks, where many large format topographic and thematic maps need to be scanned. Scanning such maps in sections and joining the pieces later in the computer is time consuming and might introduce a large number of errors.

- **Drum scanners** are more expensive and are used for professional applications which require very large precision (e.g., photogrammetry or medical applications). The map is fixed on a rotating drum. A sensor system then moves along the map and registers the light intensity or

color of each pixel. While drum scanners provide very high precision, they are also very expensive and fairly slow. A single scan may take 15-20 minutes.

Figure 2.6: Photo of feed scanner *(source: Ideal.com)*

The scanner settings chosen by the operator have a large impact on the output image characteristics. Choosing the optimal parameters requires a certain amount of experimentation, since it depends on the scanner options, the characteristics of the base maps or photos that are scanned and on the anticipated further processing steps. The most important parameters are:

- Scanning mode. Binary or "line art" is appropriate for monochrome drawings or sketches as well as for color separations where all features are basically of the same type. Grayscale mode preserves variation on a map and subsequent image manipulation can be used to extract only features that have a certain reflectance value in a graphics or image processing system. This is even easier when the maps are scanned in color mode, where, for instance, all features drawn in green on the map can be extracted using a few simple commands;

- Image resolution is measured in dots per inch (dpi). Common scanning resolutions are between 100 and 400 dpi (although air photos are usually scanned at higher resolution on special purpose scanners). A higher scanning resolution preserves more details of the original

map and results in smoother lines in the vectorized GIS data set. But the resulting images will be larger and will require more memory and disk space; a doubling of scanning resolution results in a four times larger image size. The choice depends on the properties of the source document, available hardware and the intended use of the resulting image;

- Brightness, contrast and threshold. These parameters determine the appearance of the resulting image. Brightness determines the overall lightness and darkness of the image. Contrast is used to determine how gray values or subtle color tones are preserved. Higher contrast makes the image appear sharper, but might also lead to a loss of variation and detail. Threshold is a parameter that is used in binary mode to determine how gray values in the original document are converted to black or white pixels. Parameter choice may be quite different depending on whether the goal of scanning is to produce a visually appealing and accurate representation of the source document, or whether the goal is subsequent vectorization. In the latter case, higher contrast or brightness may highlight features in the map and thus facilitate later conversion to vector format; and

- Gamma correction. Brightness and contrast control work well if the pixel values in the image are fairly regularly distributed over the entire gray scale range. This is often not the case. For example, the image might primarily consist of very bright and very dark areas. Gamma correction is a technique that considers the distribution of gray values in the image and adjusts automatically to brighten or darken areas, or to stretch cell values over a wider range of gray values. This technique can often help to preserve subtle variations in the image.

Scanning the source document is only the first and fairly straightforward step. Since the end result of the conversion process is a digital geographic database of points and lines, the scanned

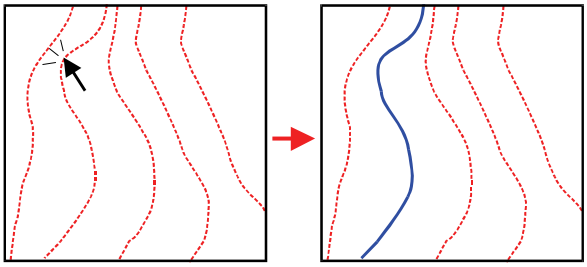
information contained on the raster images needs to be converted into coordinate information. This process is called *raster to vector conversion*. Until recently this step has been the weak link in the scanning process, which is why digitizing has usually been the preferred way of data entry. Recent advances in software development, pattern recognition techniques and processing speeds have led to major advances in this field.

Raster to vector conversion can be performed in automatic, semi-automatic or manual mode. In automatic mode, specialized software converts all lines on the raster image into sequences of coordinates automatically. Since thick lines on the map result in lines on the raster image that are several pixels wide, the automated raster to vector process starts with a line thinning algorithm. The next step is to determine the coordinates for each pixel that define the line, followed possibly by the removal of coordinates which are redundant — i.e., straight lines which can be represented by fewer coordinates. Conversion software also usually allows the user to specify tolerance levels. For example, features that consist of only one or a few pixels may actually represent dirt spots on the source maps and could be deleted automatically. Also, if the image has been scanned using a color scanner, raster to vector software often allows the user to specify line codes to be assigned to colors. This is useful for extracting different types of features into separate GIS data layers. For example, rivers may be represented in blue on the source map, while roads are drawn in black and administrative unit boundaries in red.

In semi-automatic mode, the operator clicks on each line that needs to be converted (see Figure 2.7). The system then traces that line to the nearest intersections and converts it into a vector representation. This has the advantage that the operator can select only a subset of features on the map, for example, all roads but not the rivers. Finally, in manual mode, the scanned raster image

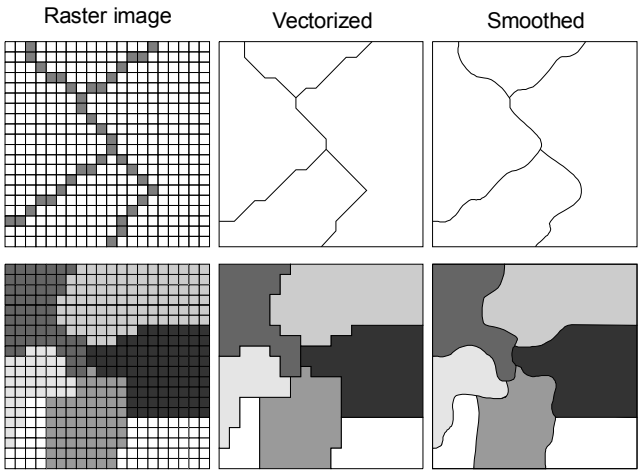
is simply used as a backdrop on the computer screen. Coordinates are created by tracing features on the scanned image using a mouse, similar to heads-up digitizing mentioned before.

Figure 2.7: Semi-automatic vectorization



If linear or area features are converted automatically from relatively low resolution raster images into vector format, the resulting line work may show unnatural sharp edges. It is common practice to smooth the vector data using spline or generalization functions available in GIS packages. Figure 2.8 shows examples for a line and a polygon data set.

Figure 2.8: Vectorization and smoothing of scanned image data



Some additional considerations

There are a number of considerations when planning a data conversion project based on map scanning. Proper preparation of the base map before scanning can significantly improve output quality. Maps should be flat and clean. Any tape residue that might be present on the map should be removed, since it may leave traces on the scanner surface. Faint features on the map can be highlighted using a pen or marker. Similarly, the operator can retrace screened line symbols and fill cross-hatched polygons to produce solid lines and fills which will facilitate automatic vectorization. Alternatively, these changes can also be made on the scanned image before vectorization. Any raster based graphics package can be used for this purpose. However, it is often easier to make these changes by hand. A water based marker or wax lead pencil should be used since petroleum based markers may damage the scanner's glass surface and graphite pencil marks reflect the light in a way that might make them invisible. For photographs, matte finishing will bring better results than glossy paper.

An additional step is often introduced for the conversion of relatively complex maps which show many different features (e.g., topographic maps) or of maps that are of bad quality. For such map data sources accuracy can be improved and post-processing efforts reduced by first tracing all required map features on transparent media such as mylar. Although this increases operator workload, tracing often turns out to be faster in the end since it reduces the time required for editing and error correction. The traced source document that is scanned subsequently is clearer and contains only those features that are actually needed. This procedure is employed in most large-scale professional scanning applications. Drafting can be avoided if the original color separations of published source maps are accessible. These can often be obtained for the national topographic map series. Each separation contains only a subset of the features of the printed map which makes it much easier to separate features into separate data layers.

Despite these preliminary steps, the scanned images might still require further processing before running the vectorization routines. Such processing may include further image enhancement such as sharpening or contrast enhancement as well as removing speckles or doing interactive pixel-level changes. A raster-oriented graphics package or the vectorization software itself will provide the necessary functions.

GIS packages that support raster data provide raster-to-vector conversion routines. Mainly these are designed for converting between raster GIS and vector GIS data and not to convert complex, scanned images into clean vector features. For a large-scale vectorization project a special purpose package is more appropriate. There are now many standalone commercial and non-commercial raster-to-vector packages available (for example, Vextractor, AbleVector, and PTracer), as well as software extensions such as ESRI's ArcScan. The available options differ between these products. Some offer de-skewing of the scanned images, or optical character recognition of map annotation which can be saved as attributes for the resulting vector features. Prices vary greatly. The data conversion staff should thus carefully compare available options and functions with the requirements of the data conversion tasks.

Advantages and disadvantages of scanning

Advantages of scanning:

- Scanned maps can be used as image backdrops for vector information. For instance, scanned topographic maps can be used in combination with digitized EA boundaries for the production of enumerator maps;

- Clear base maps or original color separations can be vectorized relatively easily using raster-to-vector conversion software; and

- Small-format scanners are relatively inexpensive and provide quick data capture.

Disadvantages of scanning:

- Converting large maps with small format scanners requires tedious re-assembly of the individual parts;

- While large format, high-throughput scanners are expensive, their cost can be justified if they are put to use on large-scale map scanning and vectorization resulting in a digital geodatabase; and

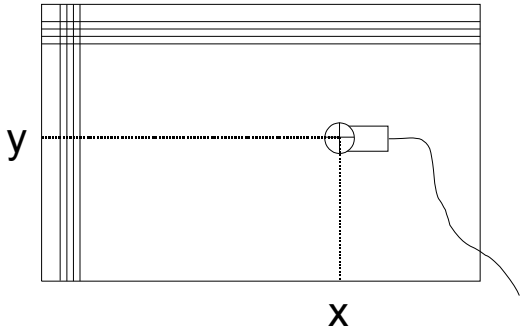
- Despite recent advances in vectorization software, considerable manual editing and attribute labeling may still be required.

Digitizing

Manual digitizing has traditionally been the most common approach for spatial data automation. Manual digitizing requires a digitizing board which may range in size from small tablets of 30x30 cm to large digitizing tables of 120x180 cm. Larger digitizing tables facilitate the digitization of larger map sheets. On a small tablet a large map will have to be digitized in several pieces and combined later. In the process of digitizing, the map is fixed to the digitizing board using masking tape. Ideally the map should be flat and not torn or folded. Paper often shrinks, especially in humid conditions, and this shrinkage introduces distortions that will be carried into the digital map database.

The first step in digitizing is to determine a number of precisely defined *control points* on the map (usually at least four). These control points serve two purposes. Firstly, if a large map is digitized in several stages and the map has to be removed from the digitizing table occasionally, the control points allow the exact re-registration of the map on the digitizing board. Secondly, control points are chosen for which the real-world coordinates in the base map's projection system are known. A good choice for control points are therefore the intersections of the graticule of latitude and longitudes that are shown on many topographic maps. In the georeferencing step which precedes or follows the digitizing of point and line features, this information is used to convert the coordinates measured in centimeters or inches on the digitizing tablet into the real world coordinates — usually in meters or feet — of the map projection. After selection of the control points, the operator traces line features on the map using a cursor that communicates with the digitizing board. The board contains a grid of wires (part of which is shown in Figure 2.9). This grid creates an electromagnetic field. The cursor contains a metal coil so that the digitizing board and cursor act as a transmitter and receiver. This allows the cursor to determine the nearest wires in the x and y direction. The exact position is found to a high degree of precision through interpolation. Features that are digitized are immediately drawn on the computer screen. This allows the operator to monitor which boundaries have been captured and whether any major errors have been introduced.

Figure 2.9: Digitizing table



Coordinates are recorded in point, distance or stream mode. In point mode, the operator pushes a button on the cursor every time a line changes direction. For curved lines the number of coordinates recorded will determine how smoothly the line will appear in the GIS database. In distance mode, a coordinate is automatically recorded when the operator has moved the cursor by a specified distance. Finally, in stream mode, the cursor automatically records coordinates at pre-specified time intervals. In distance and stream mode, there is a danger that complex line segments with many curves may be recorded with too few coordinates. Long straight segments, may in contrast yield many redundant points. The point mode, which leaves the choice of coordinate density, is usually the preferred mode of digitizing for experienced operators.

Digitizing is often tedious and tiring to the operators. Apart from ensuring that operators are well trained, it is therefore important to provide a good operating environment including an ergonomically appropriate digitizer setup. Consistent GIS software macros that guide the operator and quality control procedures will minimize errors during digitizing and reduce the time required for later editing.

During digitizing the operator has the option of assigning feature codes to each line or point that is captured. For instance, different types of administrative boundaries can be assigned codes

from one for province boundaries to three for district boundaries. In some topologically structured GIS systems, the user also has to add a label point to each digitized polygon. This can be done manually during digitizing or automatically before topology is constructed. This label point provides the link between the polygon and the geographic attribute table which contains data about the polygon (see Annex 1).

A form of data input that does not use a digitizing tablet is sometimes called heads-up digitizing. Heads-up digitizing now has two meanings. In the old method, the operator traced map features on a transparency and attached this map to the computer screen. Using a GIS data entry module or simply a graphics package that supports a GIS compatible graphics format, lines or points can now be digitized with the mouse (Figure 2.10). In the new method of heads-up digitizing, a scanned map image is used digitally to trace the outlines into a GIS layer. The operator uses a scanned map, air photo or satellite image as a backdrop. The image has been georeferenced, in that the image has been converted to a format with the same projection and real-world coordinate system as the other layers in the GIS project. The analyst converts the image by using control points and “tethering” the image to known locations both in the rest of the layers and in the real world. Good control points include street intersections and landmarks. The analyst then traces features with a mouse off the scanned image, creating a new layer in the process.

Figure 2.10: Heads-up digitizing

Advantages and disadvantages of digitizing

Advantages of digitizing:

- Digitizing is easy to learn and thus does not require expensive skilled labor;
- Attribute information can be added during the digitizing process; and
- High accuracy can be achieved through manual digitizing; i.e., there is usually no loss of accuracy compared to the source map.

Disadvantages of digitizing:

- Digitizing is tedious possibly leading to operator fatigue and resulting quality problems which may require considerable post-processing;

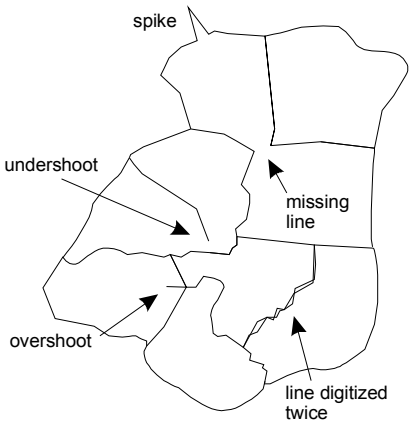
- Manual digitizing is quite slow. Large-scale data conversion projects may thus require a large number of operators and digitizing tables; and

- In contrast to primary data collection using GPS or aerial photography, the accuracy of digitized maps is limited by the quality of the source material.

Editing

The objective in converting geographic information from analog into digital form is to produce an accurate representation of the original map data. This means that all lines that connect on the map must also connect in the digital database. There should be no missing features and no duplicate lines. Manual digitizing is error prone. The most common types of errors are shown in Figure 2.11. Similarly, after raster to vector conversion disconnected line segments need to be manually joined. This happens, for instance, where small roads or rivers drawn with a thin line symbol cross major roads that are drawn as thick lines. If the minor roads or rivers are extracted into a separate map layer, there will be gaps in the road network at intersections with major roads.

Figure 2.11: Some common digitizing errors



Some of the common digitizing errors shown in Figure 2.11 can be avoided by using the digitizing software's snap tolerances that are defined by the user. For example, the user might specify that all endpoints of a line that are closer than 1 mm from another line will automatically be connected (snapped) to that line. Small sliver polygons that are created when a line is digitized twice can also be automatically removed. However, only some of the problems can be resolved in this way. Manual correction of digitizing errors after careful comparison of the original and the digitized map remains a necessary component of the data conversion process.

Constructing topology

Transforming geographic data so that points become nodes of polygons requires defining them so that they know what they are in relation to other objects. This illustrates the concept of ***topology***, defined as study of the properties of geometric figures that are not changed by distortion. Topologically correct data differs from tables or graphical objects without topology (often referred to as 'spaghetti'). Objects with topology know their locations in absolute space and also know their nearest neighbors. The construction of digital map topology supports the editing process. For example, it allows the user to identify problems such as polygons that are not completely closed. Feature topology describes the spatial relationships between connecting or adjacent geographic features such as roads connecting at intersections (see Annex 1 on GIS). Structuring a GIS database topologically involves the identification of these spatial relationships and their description in the database. How this is actually done is software-specific. Storing the topological information facilitates analysis, since many GIS operations do not actually require coordinate information, but are based only on topology. For example, a district's neighbors can be determined from a database table that lists for each line the polygon to the right and the one to the left (see Annex A1).

The user typically does not have to worry about how the GIS stores topological information. Provided that the digital database is clean — that is all lines are connected and polygons are properly identified — a GIS function is used to build topology and create all necessary internal data files. This function will only perform successfully if the map database does not contain any errors. Building topology thus also acts as a test of database integrity.

Digital map integration

A census mapping project should take advantage of all suitable geospatial data sources. These are likely to be stored in different formats, using varying map scales and cartographic projections. Integrating these heterogeneous data sources requires considerable knowledge of GIS data integration methods, if the goal is to produce a complete and seamless digital census map database. The following sections discuss the most important methods that facilitate digital map data integration.

Georeferencing

The coordinates captured with a digitizer or scanner are relative coordinates measured in the x and y direction usually in centimeters or inches from the data input device's origin — usually the lower left corner. If several adjacent map sheets are digitized, they will clearly not fit when their digitized map sheets are later pasted together in the database. In fact, they would be drawn on top of each other since they are all referenced in the same segment of the digitizer's coordinate system. Similarly, existing georeferenced GIS layers for the same area or coordinates collected using a global positioning system will not be compatible with the digitized maps since they are referenced in a real-world coordinate system. For this reason, the digitized point and line coordinates need to be converted from digitizing units into the real-world map coordinates which

are measured in meters or feet (see also Annex 2). As pointed out earlier, this step can be done in most systems either at the start of digitizing or after spatial data automation has been completed.

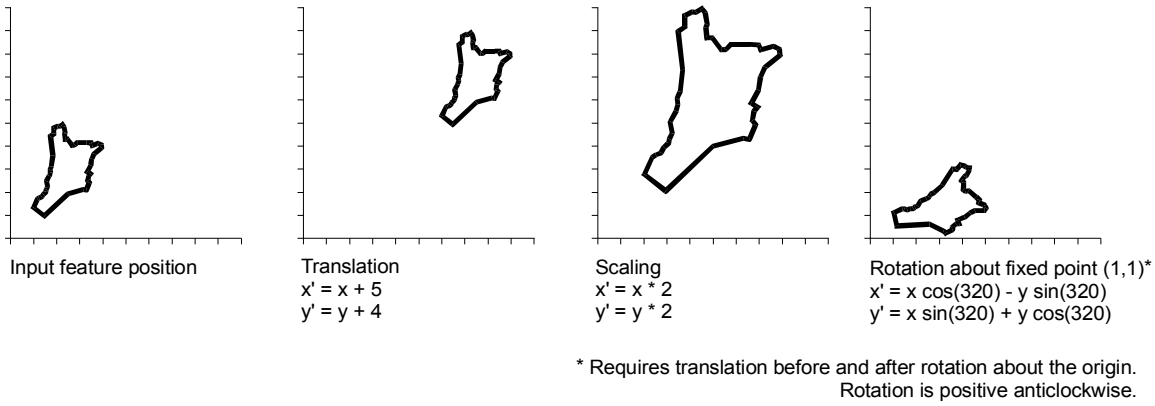
Nearly all GIS packages provide the functions necessary for georeferencing. The user needs to specify a number of control points for which the real-world coordinates are known. Based on the input coordinate data in digitizing units and the real-world output coordinates, the system computes a set of transformation parameters that perform the following transformations (see Figure 2.12):

- Translation: The geographic feature is moved to a new position simply by adding (or subtracting) constant values to the x and y coordinates. The offset will usually be different for x and y;

- Scaling: The feature is enlarged or reduced by multiplying the x and y coordinates by a factor for the x and y coordinates respectively. The scaling is usually done relative to the origin of the coordinate system; and

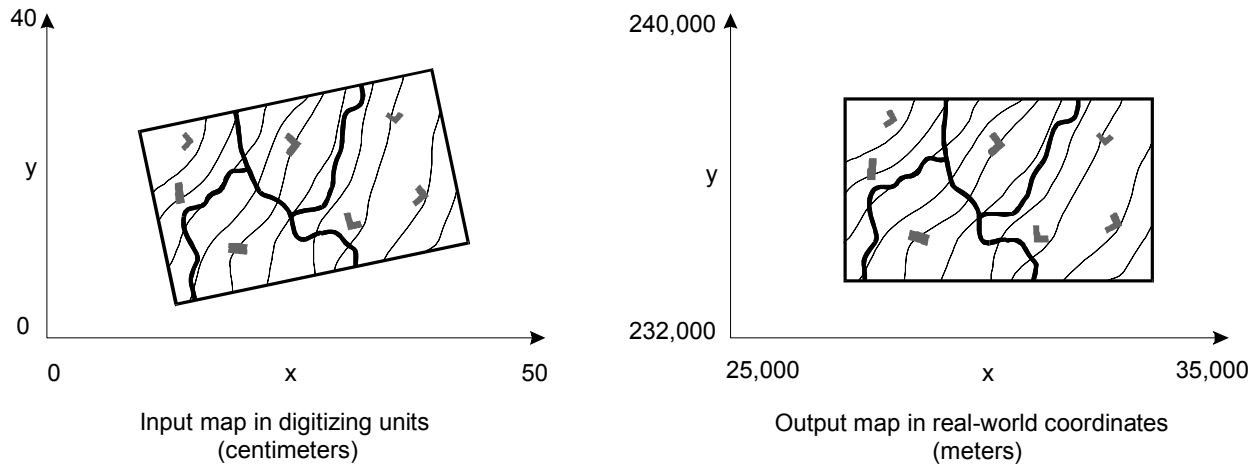
- Rotation: The geographic feature is rotated about the coordinate systems origin by a given angle. Rotation will make sure that the resulting digital map has the proper orientation even if the paper map had not been correctly aligned on the digitizing board.

Figure 2.12: Translation, scaling, rotation



Note that the shape of the digitized features does not change in this transformation as it would in a projection change. Only the relative size and orientation of the objects is modified. After the correct translation, scaling and rotation parameters have been computed, the system applies these parameters to all point and line coordinates in the database. The output is a map that looks very similar, but is now registered in the proper coordinate system that was used in the production of the original base map (see Figure 2.13). It is important to ensure that the error in this operation is minimized. The system usually provides information on the error in the estimation of transformation parameters for each point, which is helpful to detect errors in specifying the control points' real-world coordinates. More technical details are given in an example in Annex 2.

Figure 2.13: Map in digitizing units; map in real-world coordinates



A serious problem occurs when the map projection and coordinate system of the source paper map are unknown. Unfortunately, this problem is encountered quite frequently since many paper maps, especially thematic maps, do not contain this information. Two options available in this case are to try a large number of possible map projections (the standard projection used in the country's mapping programs is a good candidate), or to use what is known as rubbersheeting.

Rubbersheeting requires a large number of control points that are well distributed across the map. Sometimes a digital map of country and administrative boundaries or any other clearly defined points that are also present in the digitized map can be used to find links between corresponding points. The system then uses the coordinates of the input and output coordinates to compute higher-order polynomial transformations. Typically, the error introduced in rubbersheeting is quite large, and this operation should therefore be avoided if at all possible. However, in some instances, where the input maps clearly do not conform to any well-defined projection, rubbersheeting is a viable option to make use of available geographic information. A good example in the context of census mapping is the georeferencing of hand-drawn sketch

maps. Section A2.6 in Annex 2 provides a practical example of georeferencing that illustrates the process of converting, for instance, a digitized map into a properly referenced digital database.

Projection and datum change

Related to the transformation process that converts the coordinates of digital map features without changing their shape is projection change. When converting from one projection to another, the shape and distortion of map features does change, although the changes may be all but imperceptible at large cartographic scales.

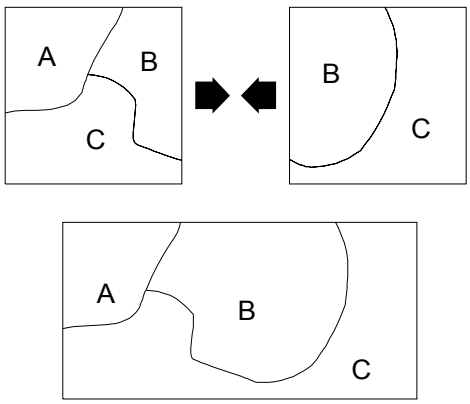
Projection change is necessary when maps that were digitized from different map sheets need to be assembled into a seamless database. Often, maps issued at different map scales use different projections. In other instances, a mapping agency may have changed the standard projection used for mapping in the country, so that older map sheets may use a different projection from those map sheets that were revised more recently. Similarly, the mapping agency may have modified the geographic datum, which establishes the reference framework for cartographic work in the country, so that older topographic maps, for example, use a slightly different coordinate system, than newer maps.

Projections and geographic datums are discussed in more detail in Annex 2. It will be useful for the census mapping agency to have trained staff or access to experts from an outside agency to advise on the most appropriate strategy for reconciling projections and related issues to produce a consistent national census map base. The actual technical steps of projection change will require relatively little effort, since all commercial GIS provide the required projection change functions.

Integration of separate map segments

The purpose of a digital mapping project is to produce a seamless database for a large region or an entire country. At medium or large cartographic scales (e.g., 1:250,000 or larger), base map information will be contained on separate topographic map sheets. These are digitized separately and the resulting digital map sheets are joined in the GIS (see Figure 2.14).

Figure 2.14: Joining adjacent digital map sheets



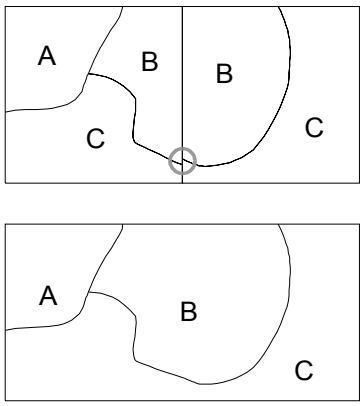
Usually this process is straightforward. But the match between map sheets might not always be perfect. Features that span both sheets — e.g., roads or boundaries — might be displaced at the map boundaries (Figure 2.15). Errors could have been introduced during digitizing, or the errors may actually be present on the source map sheets. For instance, adjacent map sheets may have been produced at different times, so that newer features such as new roads do not continue across map sheet boundaries or they are represented by different symbols.

The problem is particularly serious if there is no complete coverage for the entire country at the desired map scale, so that map sheets of different scales with different feature densities need to be integrated. This problem is often encountered when integrating map sheets at the urban/rural interface, where large scale city maps need to be matched to smaller scale rural maps. Due to the

variations in cartographic generalization, features may or may not be present on the smaller scale maps, or their symbology may be different in the two map series. Integration of such maps requires considerable judgment and experience.

The process of fixing these errors is called edgematching. It is usually performed manually involving a considerable amount of editing. If the displacement is not too large and the features are compatible across map sheets, features can be connected using automatic edgematching functions provided by some GIS packages.

Figure 2.15: Edgematching after joining adjacent map sheets



The census geodatabase

Through constructing a geodatabase, the NSO begins to realize the benefits of organizing itself around a geographic model. For designing census data models, an NSO must consider needs of its data users as well as statutory and constitutional provisions for census work. Generally a conceptual model can be developed to link basic collection geography from the enumeration, layers such as EAs, physical features (including elevation, slope, and aspect) and other layers that affect census-taking including point and area features such as landmarks. Vector topology

has been mentioned in the previous section as playing a key role particularly in boundary and administrative area delineation.

While vector topology is important, other data formats may also need to be brought in a GIS project, including scanned maps in the form of raster files, satellite images, aerial photos, CAD files (which can be brought in with topology and attributes or as featureless "spaghetti"), and global positioning system data in streams of points.

Among the advances in computing in the late 1990s and early 2000s was the development of object-based data in the computing realm. In object-oriented programming, objects can be defined as having literally thousands of different characteristics that when imported in a GIS, allow them to "behave" predictably in simulated conditions and permit sophisticated modeling such as travel-route analysis.

Geographic databases (hereafter referred to as geodatabases) are more than spreadsheets. Entity types can be defined as having specific properties that govern behavior in the real world. The EA as a geographic unit is a kind of object whose function is to delineate territory for the census canvassing operation. Morphologically, the EA is contiguous, it nests within administrative units, and it is composed of population-based units.

Relational databases

Before discussing specific structures of the census GIS database, the concepts of relational databases will be reviewed. All large operational GISs are built on geodatabases; they are arguably the most important part of the GIS. Geodatabases form the basis for all queries,

analysis, and decision-making. A DBMS, or database management system, is where databases are stored.

A geographic object can be defined as an integrated package of geometry, properties and methods. Objects of the same general type are grouped together as object classes, with individual objects in the class referred to as "instances." In many GIS systems, each object class is stored physically as a database table, with each row an object and each property a column. For example, the object class "Street light" can include instances such as "gas lamp," "sodium vapor lamp," and "mercury vapor lamp."

The relational database model is used to store, retrieve and manipulate tables of data that refer to the geographic features in the coordinate database. It is based on the entity-relationship model.

In a geographic context, an *entity* can be administrative or census units, or any other spatial feature for which characteristics will be compiled. For example, an entity might represent the feature "enumeration area" (Figure 2.16). Individual enumeration areas in a district or country are instances of this entity and will be represented as rows in the entity's table. The entity type, in contrast, refers to the structure of the database table: the attributes of the entity which are stored in the columns of the table. For an enumeration area this may be the unique identifier, surface area, population, the code of the crew leader (CL) area that the EA is assigned to, and so on. Note that the entity type only refers to the generic definition of the database table, not to the actual values recorded for each instance. One or more attributes (columns) in the entity type are used as keys or identifiers. One of those is the primary key which serves as the unique identifier for an entity type. For an enumeration area database this would be the EA code.

Figure 2.16: Example of an entity table – enumeration area

Entity: Enumeration areas

Type (attributes)

Instances

EA-Code	Area	Pop	CL-Code
723101	32.1	763	88
723102	28.4	593	88
723103	19.1	838	88
723201	34.6	832	88
723202	25.7	632	89
723203	28.3	839	89
723204	12.4	388	89
…	…	…	…

Primary key

Varieties of relational database and geodatabase structure

Database management systems (DBMSs) can be divided into various types, including relational, object, and object-relational. Relational datebases consist of a set of tables, each being a two-dimensional array of records containing attributes about objects under study. While they are flexible and useful, they are not designed to handle rich data types such as geography, where object topology and relationships can be complicated. Commercial and open-source RDBMS programs include Microsoft Access and Oracle.

Object-DBMS are designed to address a central weakness in RDMS, namely its inability to store complete objects directly in the database. ODBMS can store objects persistently and provide object-oriented query tools. Object-relational DBMS are a hybrid object-relational database consisting of an RDMSs engine with an extensibility framework for handling objects. Ideally an ORDBMS has the following components: query parser for SQL (structured query language) queries, a query optimizer, query language, indexing, storage management, transaction services, and replication.

Software companies have answered the need for spatial capability in their relational databases through the use of geographic DBMS extensions. These are major DBMSs with spatial database extensions. Examples include IBM's DB 2 Spatial Extender, Informix Spatial Datablade, and Oracle Spatial. These software products handle points, lines, and polygons as feature types that can be aggregated to richer types using topology and linear referencing capabilities, indexing with R-tree and quadtree methods.

There are special challenges involved with extending DBMSs to store geographic data. Object data models are geometry-centric, in that they model the world as collections of objects such as points, lines, polygons, or rasters. Operations are performed on the geometry as separate procedures using programs or scripts. But this is overly simplistic for geographic systems, especially when they contain many entities with large numbers of properties, complex relationships, and sophisticated behavior.

Relations define the association between entities. For instance, a table describing enumeration areas can be linked to a table for the entity crew leader area. This table has attributes such as the name of the crew leader, the regional office responsible, and contact information. The primary key in this table is the crew leader code (CL-Code), which is also present in the EA table. A relational database management system can thus join the two tables so that each instance in the EA table is matched with the corresponding instance in the CL area table.

The process of designing a relational database structure through a series of steps is called *normalization*. The outcome is a database with minimum redundancy. In other words, the data are organized in a number of tables so that values that are repeated many times are avoided. This

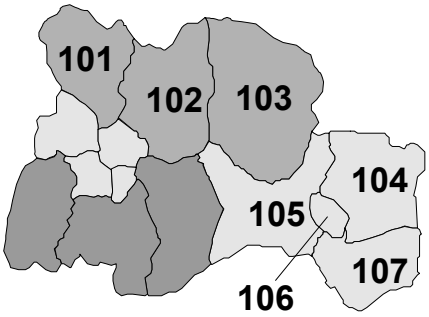
reduces storage space and avoids errors that might be introduced in standard database operations such as insertion, deletion or updates.

Figure 2.17 illustrates the difference between a simple data table and its normalized form using an example of a district database. In the first instance, the information for the provinces is repeated for each district in the province. This not only wastes storage space, it also makes it more difficult to update or change information for provinces. The values would need to be replaced for each individual district. In the normalized database structure, the name of the province has been replaced by a more compact numeric code, which provides the link to a second table. Here, the province code becomes the primary key for the province information that includes the province name, population and total fertility rate. After joining the two databases temporarily by means of the province code, province information can be accessed for each instance in the district table.

Defining a clean database structure is not a trivial task. Some database management programs provide normalization functions that automatically create a relational database structure. However, this is usually not a good substitute for a comprehensive design of the overall database.

Figure 2.17: Relational database tables

Provinces and districts



Original data table

Id	Province	District	P_Pop	P_TFR	D_Pop	D_TFR
101	Merida	Palma	214084	3.2	89763	3.4
102	Merida	S. Maria	214084	3.2	45938	2.9
103	Merida	Veralo	214084	3.2	78383	3.2
104	La Paz	Bolo	397881	3.7	98302	3.9
105	La Paz	Jose	397881	3.7	67352	4.2
106	La Paz	Malabo	397881	3.7	102839	3.7
107	La Paz	Chilabo	397881	3.7	129388	2.8
…	…	…	…	…	…	…

Data tables after normalization

Districts

Id	Province	District	D_Pop	D_TFR
101	Merida	Palma	89763	3.4
102	Merida	S. Maria	45938	2.9
103	Merida	Veralo	78383	3.2
104	La Paz	Bolo	98302	3.9
105	La Paz	Jose	67352	4.2
106	La Paz	Malabo	102839	3.7
107	La Paz	Chilabo	129388	2.8
…	…	…	…	…

Provinces

Province	P_Pop	P_TFR
Merida	214084	3.2
La Paz	397881	3.7
…	…	…

Components of a census database

A comprehensive census GIS database consists of a digital map of census enumeration areas and, in most instances, a series of base map layers that provide the context and orientation in the final enumerator maps. Base data layers might be roads, rivers, buildings, or settlements. Each of these will be contained in a separate GIS database. So, for instance, roads and rivers, although they are both represented as lines, will not be stored in the same digital file.

Before starting data entry and data conversion, the census cartographic staff should design the structure of all GIS data sets that will be produced. This structure definition will be a detailed description of all conventions and guidelines that the cartographic staff

needs to follow to ensure consistency of the final output products. A good planning process will avoid confusion and incompatibilities later in the process.

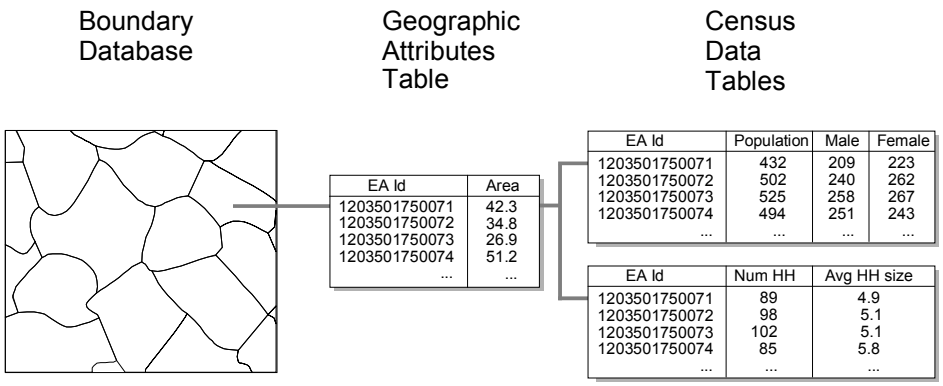
The first step is to think about what the final products will look like. The complete digital enumeration area database, for example, will likely consist of the following components (see Figure 2.):

- The spatial *boundary database* consisting of area features (polygons) that represent the census units.

- The *geographic attributes table*: a database file which is linked internally to the spatial database and contains one record for each polygon. This table contains the unique identifier for each census unit and possibly some additional static, i.e., unchanging, variables such as the unit's area in square kilometers.

- The *census data tables* containing non-spatial attributes—i.e., the census indicators for the spatial census units. Each of these files must contain the unique identifier of the census unit which provides the link to the corresponding polygon attribute table records. There will be one record for each census unit.

Figure 2.18: Components of a digital spatial census database



Definition of database content (data modeling)

Once the scope of census geographic activities has been determined, the census office needs to define and document the structure of the geographic databases in more detail. This process is sometimes termed data modeling and involves the definition of the geographic features to be included in the database, their attributes and their relationships to other features. The resulting output is a detailed data dictionary that guides the database development process and also serves as documentation in later stages.

Note that many GIS databases are created without detailed data modeling. This step requires time and some degree of expertise in database concepts. The additional investment is justified in a comprehensive census mapping project. The process of data modeling imposes a level of rigor and consistency that will ensure a high quality database and easier maintenance. For a census mapping agency that goes through this process for the first time it may be desirable to recruit an experienced GIS database consultant to guide the team through this process.

As discussed earlier, many national and international agencies have already been active in developing generic data models for spatial information as part of a national spatial data infrastructure (NSDI). Often, a census office will be able to simply adapt an NSDI standard to the specific needs of statistical data collection. In cases where such information is unavailable, a data model needs to be developed in house. Templates from mapping or statistical agencies in other countries will provide a useful reference for that purpose.

Annex 3 provides an example that illustrates what a data model description in a data dictionary might look like. Related to the data model are both metadata standards, which are discussed in the following section, and simpler database dictionaries which accompany databases distributed to the general public (see Annex 4).

Data quality issues

Accuracy requirements

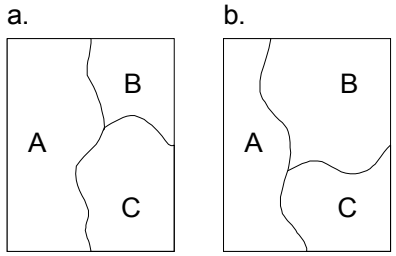
The development of acceptable data accuracy standards is perhaps one of the most important tasks in planning a digital database development project. In many fields such as utilities and facilities management, terrain or hydrological mapping, accuracy database standards exist that can be adopted for any new project. Census mapping, in contrast, has traditionally been done in an ad hoc way using manual techniques and sketch maps with little concern about geographic accuracy. This was adequate as long as census maps were used for the purposes of the census only. With GIS, however, census maps have become an integral part of many analytical applications in the government, private and academic

sectors. This is a major factor that justifies the investment in digital census mapping in the first place. When census maps are combined with other digital geographic data sources, shortcomings in accuracy become immediately apparent. Accuracy requirements for digital census mapping are therefore higher than for traditional census mapping techniques.

Accuracy in GIS refers to both the attribute data — the geographic attribute table and the census data that can be attached to it — and to the geographical data. Issues concerning attribute data accuracy are no different from those encountered in census related data entry and processing activities. They will therefore be discussed only briefly. Geographical data accuracy relates to the points, lines and areas that are stored in the GIS database and that describe features on the earth's surface.

Geographical data accuracy can be divided into *logical* and *positional* accuracy. Positional accuracy is sometimes also called absolute accuracy. Logical accuracy refers to the integrity of relationships among geographic features. For instance, a road in one GIS database layer must connect to a bridge in another layer. A river stored in a hydrological database that defines the boundary between two administrative units should coincide with the boundary between those units. And, a town represented as a point in one GIS database should fall into its corresponding administrative unit in another GIS layer. The same logical relationships can be represented correctly in different maps that have very different appearance. For instance, in Figure 2.19, the two maps correctly represent the neighborhood relationships between three administrative units.

Figure 2.19: Logical accuracy



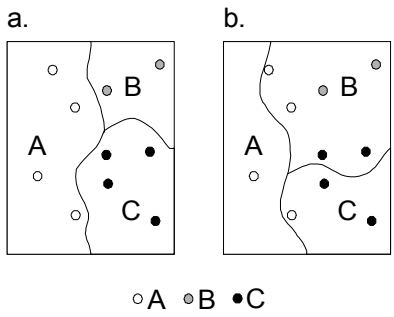
Positional accuracy, in contrast, maintains that the coordinates of features in the GIS database are correct relative to their true positions on the earth's surface. That means that cartographic measurements must be conducted with a sufficient degree of precision using accurate measurement devices such as global positioning systems. Of course, a data set that is free from positional error will also accurately represent the logical relationships between geographical features.

In some applications, logical accuracy is more important than positional accuracy. For a census database, it may be more important to know that a certain street defines the boundary of an EA, than to know that the exact coordinates represent the real-world location of the road to a very high degree of accuracy. In fact, sketch maps produced in traditional census mapping activities are typically logically accurate, but have low positional accuracy. This is not a problem when maps are only used to support census enumeration as long as the distortions do not make orientation in the EA impossible. But, if the census maps are subsequently used for other purposes, significant problems can occur.

Figure 2.20a, for instance, shows a set of sample survey sites that have been determined using a very accurate global positioning system. The underlying base map has a high

degree of positional accuracy, so that the points fall into the correct administrative unit. The base map in Figure 2.20b, in contrast, while logically accurate, has a low degree of positional accuracy. Some of the accurately measured GPS points therefore fall into the wrong administrative units. This will lead to incorrect results when survey responses are aggregated by administrative unit.

Figure 2.20: Problems if positional accuracy is not maintained



A sufficient degree of positional accuracy should therefore be the goal of a digital census cartographic process, if the resulting boundaries are used beyond the actual enumeration. Of course, few geographical data sets are 100 percent accurate. In any mapping effort, manual or digital, there is a trade-off between attainable accuracy and the time and funds required to reach this level of data quality. Typically, an incremental gain in accuracy above 90 or 95 percent requires a greater than proportional input in time and other resources. In fact, some estimates claim that increasing accuracy from 95 to 100 percent would require 95 percent of the total budget of a project (Hohl 1998).

It is common practice in topographic mapping to define accuracy standards based on the position of point locations. Elevation spot heights, for instance, are required to be within *x* meters from their true position in *y* percent of all cases. The acceptable error increases

as the cartographic scale decreases. For instance, on a 1:25,000 scale map, the error should be smaller than on a 1:100,000 scale map. Since census maps will to a large extent be based on available topographic maps, accuracy standards for census mapping should be developed in close cooperation with experts from the national mapping authorities. This will also ensure compatibility between the quality of the products of the census mapping project and that of other national digital map series.

Although a high degree of positional accuracy is desirable, accuracy standards that are too limited will lead to increased costs, exaggerated user expectations and possibly frustrations among cartographic staff who may not be able to attain goals that have been set too high. Accuracy standards that are too low may lead to products that are of insufficient quality. Users may either reject the product if they are aware of its limitations, or they may use it with an overstated level of confidence which may lead to serious errors in the results of analyses. A popular concept in GIS database development is "fitness for use." This takes account of the fact that digital spatial data are never perfect. While they may be appropriate for one task, they may be of insufficient quality for another.

When determining quality standards, the census organization must consider not only its internal needs but also the needs of the outside users of the digital census maps. Data accuracy guidelines should thus be developed in collaboration with all stakeholders as part of the user needs assessment. Standards will also be affected by available resources, the quality of the source materials — information for different data layers may be of varying quality — and the technology chosen for field data collection.

Quality control

Quality control is the set of processes and conventions that ensure that the databases that are developed in the census cartographic process conform to the defined accuracy standards. The revised *Principles and Recommendations* (United Nations, 2006) stress the importance of quality control and contain an overview of these issues in the census process. These general concepts also apply to census mapping.

Tests and error checking procedures form the core of the quality control process. However, quality control is also a matter of attitude among the census cartographic staff to limit errors at every step of the data conversion process. Census staff should be encouraged to report problems in the output products. Recurrent problems may point to inadequate procedures or training deficits, and may require changes in staff members' assignments or a modification of equipment or techniques. It is therefore very important that staff are not afraid to report problems with their own work and that they clearly understand the overall objective of quality control procedures.

While specialization in different tasks among staff members may improve overall data quality in most cases, many tasks in GIS database development are quite repetitive. A monotonous work assignment can cause an increase in errors as concentration diminishes. Rotation of work assignments can help prevent this. This will also expose staff members to different aspects of the overall data conversion process which should improve understanding of their tasks and therefore overall product quality. Staff members should also be asked to suggest changes in procedures that lead to improved data quality. Such suggestions should be evaluated in a controlled environment—not in the regular

work process—before changes are implemented. Achieving the highest possible data quality thus becomes a continuous process.

Quality control procedures consist of automated and manual methods. Automated procedures are preferable since they are fast and reliable. However, many aspects of data conversion can only be evaluated through visual inspection and comparison. Automated techniques for geographic attribute data are similar to those used in census data entry. Range and code checks ensure that attribute fields only contain allowable values. The number of administrative or census units in the digital database must match the corresponding number in the geographic area master list. The geographical area identifier is the single most important field in the census GIS database, since it ensures the match between the digital base maps and the aggregated census data. The largest resources in attribute data checking — automated as well as manual — should thus be committed to ensuring that no errors exist in this attribute.

Automated quality control options for the geographic data are relatively limited. Some GIS packages will check the accuracy of database topology: for instance, whether all areas are closed and all lines connect. A village database can be combined with an administrative unit boundary data set of known quality to ensure that the administrative identifiers in the village database are correct (a point-in-polygon operation). Some errors are obvious, such as when the boundaries of two administrative units that were digitized separately do not match. Others are less easily spotted, for instance when some internal boundaries or roads are missing from a GIS data set. For the most part, therefore, quality control for map products must rely on visual comparison of source materials (maps, air photos, etc.) with digitized data. For this purpose, the digital maps are printed, ideally at

the same scale as the source maps. The source material and product are then compared either side by side or overlaid on a light table. Any systematic error points to a problem in data conversion procedures that should be addressed immediately. Manual error checking should never be conducted by the staff member who produced the data.

Quality control steps should be documented thoroughly. A hardcopy log form is generally the most appropriate means of documenting data quality, although automated, digital forms can also be used. The log form should specify the quality control procedure performed, when and by whom it was carried out, who produced the data that are checked and the results of the tests. Logs should be created for manual as well as automated tests. These logs not only document the accuracy of a data set and its lineage. They can also point out which staff members may require additional training.

A consistent set of quality control procedures should result in an end product of acceptable accuracy. However, in most projects, a final step termed quality assurance is usually added which consists of another round of checking and a last process of problem resolution. Quality assurance will be discussed in a later section.

Tiling of national territory into operational zones

A complete digital enumeration area database will consist of thousands of units. For larger countries, it is not usually practical to store all EA polygons in the same physical data layer. Instead, the national territory can be divided into operational zones. In a decentralized census administrative structure, different regional offices and different operators within each regional office can thus work on separate parts of the database simultaneously. Provided that consistency between the boundaries of the subsections of

the national database has been enforced, the separate pieces can be combined at a later stage to produce district, province or national level maps. This process will, however, require some edge-matching, which involves the manual linking of connected features that cross two or more tiles.

For larger countries, it is likely that cartographic work is decentralized. In that case, operational zones are naturally defined by the area of responsibility for each regional census office. For example, a country may assign census cartographic work to four regional offices with the head office functioning simultaneously as the overall coordinating body and as one of the regional offices. Within each regional office, the databases can be further divided into smaller zones. Working on smaller size databases is usually less computationally demanding. Division into smaller parts also allows several operators to simultaneously work on separate parts of the database.

The digital administrative base map

If a decentralized approach is chosen, the national census office should first create a national boundary template for the major administrative levels in the country. For example, the census office should create, obtain or commission a set of digital spatial boundaries of provinces, districts and ideally also sub-districts. These boundaries should be of high accuracy and should show an amount of detail which makes them useful for EA mapping at larger cartographic scales (e.g., at least at a scale of 1:250,000). These boundaries should be used throughout the census mapping process as well as for the distribution of spatially referenced aggregate census information at these administrative levels.

Such boundaries may have already been produced in digital form by the national mapping agency. In that case, they will represent the officially recognized digital administrative base map for the country (see the discussion about National Spatial Data Infrastructures). Codes used in the administrative base should obviously correspond to the codes used in the census database.

The official district boundaries for each operational zone should be distributed to the offices in charge of delineating enumeration areas. The EA boundaries are then entered into these official administrative unit polygons. This will ensure that in any subsequent aggregation, the boundaries of neighboring districts will match perfectly. If district boundaries were digitized by each local office separately, it is unlikely that boundaries would coincide perfectly. Significant further editing would then be required. Furthermore, there would be considerable duplication of work, since the same boundaries would be digitized twice—once by each neighboring regional office or operator.

Dealing with disjoint area units

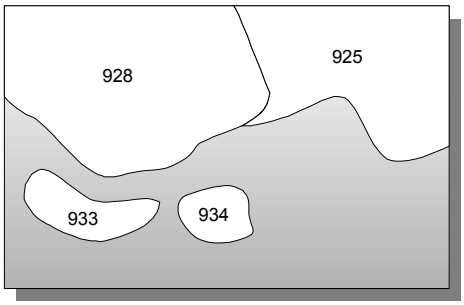
Administrative units are frequently split into separate, distinct spatial units or polygons. For example, a district may consist of an area on the mainland and a number of islands. For census data processing this is not a problem since there will be only one record in each census data table that applies to the district. In the geographic attribute database, however, this district will have two or more records — one for each polygon. This will cause problems when census attribute information is linked to the polygons via the geographic attribute table. In a relational database system, the census data record is linked to each polygon in the GIS database that has the same district identifier. Mapping

average values or densities presents no problem. Average income or population density are the same in the entire district. Count data, however, such as total population or number of households present a problem when a user wants to sum the total population of all districts. Since the records are repeated for each polygon belonging to the same district, some double counting will occur and the final total will be exaggerated. There are two approaches for dealing with this problem.

Some advanced GIS packages allow the definition of *regions*. Regions can consist of one or more individual polygons, but there is only one record for each region in the geographic attribute table. The system keeps track internally which individual polygons belong to which region. In some packages, regions can even overlap, although this is not a useful feature for census applications, where enumeration areas have to be mutually exclusive.

Many lower-end GIS software do not provide this option. In this case, a simple solution is to add an additional data field (a "flag value") to the geographic attribute table. This field will assume the value of one for the largest polygon belonging to the district and zero for the smaller ones. Before summing or averaging any attribute value, the user can first select only the polygons with a value of one in this field. An additional field could be added that contains the number of polygons belonging to the same unit. This information can be generated quickly using the frequency or cross-tabulation feature of the GIS package.

Figure 2.21: Dealing with administrative units consisting of several polygon



Internal identifier	Admin. identifier	Admin. name	Flag
925	02015	Kalana	1
928	02012	Bagor	1
933	02012	Bagor	0
934	02012	Bagor	0

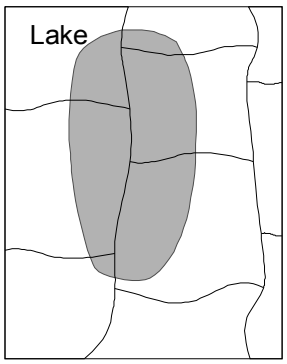
Computing areas

The utility of census databases will be enhanced if a number of standard geographical variables are included. The most important of these is the area of each enumeration area or administrative unit. Any GIS package will compute the area of a polygon provided the database is properly referenced in an equal area reference projection. However, depending on the resolution and accuracy of the digitized boundaries, there may be considerable error in the GIS measurements due to highly generalized boundaries and missing islands that may have been too small to be included on a small scale map. If available, it is therefore preferable to use more exact area figures produced by the national mapping agency.

Area figures are used to produce density estimates, most importantly population densities. Published area figures usually refer to the extent of the total legal boundary of the administrative unit—i.e., its total area. Sometimes this can lead to somewhat

misleading density estimates. In one instance, for example, a national census publication reported the area of several districts that neighbored a large lake. The reported areas included the portion of the districts that extended from the lake shore to the center line of the lake (see Figure 2.22). This inclusion of the lake area doubled the total area of some districts. Consequently, the actual population densities were underestimated by a factor of two. Where official statistics on population density are used, for example, as a criteria for allocation of resources or to determine eligibility for government programs, the definition of population density can have severe consequences.

Figure 2.22: A lake covering a large area in several administrative units



In countries where this is a problem, the census office may decide to report two area fields: one which is the *total area* of an administrative unit, and one which is the *land area* — i.e., the total area minus the area covered by water bodies and possibly other uninhabited areas such as protected conservation areas. Some countries also report the area of agricultural land. This allows the users to compute agricultural population densities or vice versa, the number of hectares of agricultural land available per inhabitant in the district. These area figures can be computed quite easily in a GIS using appropriate geographic data layers subject to the caveats relating to map generalization

mentioned above. In any case, it is important that the definitions of the net areas are well documented.

Since most GIS packages treat every polygon in the database as a separate record, GIS computed area figures for administrative or census units that consist of more than one polygon will not be useful for density calculations. Instead the areas of all polygons belonging to the same administrative or census unit need to be aggregated. This can be done in the GIS using appropriate cross-tabulation functions.

Metadata development

This handbook argues that census mapping should be considered as a long term process, not as a one time effort. Over a long period of time, elements of a database will be accessed repeatedly, sometimes after a considerable interim. The possibility of frequent staff changes means that the institutional memory needs to be based on more than the recollection of the GIS analysts involved in initial data development. The detailed documentation of all steps involved in developing the digital spatial census database is therefore mandatory.

Information about data quality, formats, processing steps, and all other information pertaining to a data set are termed metadata, or "data about data." Metadata have several objectives:

- To support the maintenance and updates of digital data sets held by an organization;
- To support data distribution by providing information about a data set's fitness for use to outside users; and

- To support the integration of externally produced data sets into an organization's data holdings.

What different data producers consider essential metadata can differ widely. Many countries have therefore started the development of general geographic metadata standards. These aim at unifying the conventions for documenting spatial information. They therefore support the development of a national spatial data infrastructure by facilitating spatial data exchange and integration. At the international level, several organizations attempt to coordinate the development of spatial metadata standards among groups of countries. Among these are the International Organization for Standardization (ISO/TC 211) Working Group on Geographic Information/Geomatics (www.statkart.no/isotc211/), the European Commission's Open Information Interchange Service (www2.echo.lu/oii/en/oii-home.html), and the Permanent Committee on GIS Infrastructure for Asia and the Pacific (www.permcom.apgis.gov.au).

Because spatially referenced census data is an integral part of a national spatial data infrastructure, the development of digital census maps should be integrated as much as possible with other digital mapping efforts in the country. Concerning metadata, that means that a national or regional metadata standard, if it exists, should be adopted by a national census organization. Close cooperation with the responsible national authority — usually the national mapping organization or an inter-departmental advisory board — will facilitate the introduction of such standards. If a national standard does not exist, the census organization will save time and resources by adapting a suitable standard from another country rather than developing a metadata standard from scratch.

An example of a well-developed and widely used metadata standard is the Content Standards for Digital Geospatial Metadata (CSDGM) developed by the National Geographic Data Committee in the United States (www.fdgc.gov). It serves as an illustration for the types of information contained in a metadata database. The complete standard is very comprehensive, and various specialized committees develop guidelines for specific types of data. The Subcommittee on Cultural and Demographic Data, for instance, is housed at the US Census Bureau (http://www.census.gov/geo/www/standards/scdd/; also see FDGC 1997b). Here, we will only discuss the main components of the metadata definition.

The CSDGM standard consists of seven main sections and can be thought of as a database template with fields describing different aspects of a spatial data set. Some fields will contain one of a predefined set of codes or attributes. But many elements are text fields in which the data producer describes database features such as quality or lineage information. The most important elements are considered mandatory, so they have to be entered for each data set. This mandatory set of fields is a good starting point for the definition of a census organization's metadata template. Others are labeled 'mandatory if applicable' or 'optional'.

The main components of the standard are:

- *Identification Information* including the data set title, area covered, keywords, purpose, abstract, and access and use restrictions;

- *Data Quality Information* such as horizontal and vertical accuracy assessment, logical consistency, semantic accuracy, temporal information, data set completeness and

lineage. Lineage includes data sources used to produce the data set, as well as processing steps and intermediate products;

- *Spatial Data Organization Information* refers to the way the data are stored such as point, raster, vector, and digital map sheet tiling information;

- *Spatial Reference Information* includes the map projection and all relevant parameters that define the coordinate system;

- *Entity and Attribute Information* contains detailed definitions of the attributes of the data set including the attribute data types, allowable values, and definitions. This is largely the same information that is contained in a data dictionary as described earlier;

- *Distribution Information* including the data distributor, file format of data, off-line media types, on-line link to data, fees and order process; and

- *Metadata Reference Information* provides information about the metadata itself, most importantly who created the metadata and when.

In addition to the seven major sections, the content standard includes three minor elements. These are frequently referenced in the main sections. Instead of repeating these elements many times they only need to be stored in one location. The three minor sections are:

- *Citation Information* ensures consistent referencing of the originator, title, publication date, publisher;

- *Time Period Information* includes single date, multiple dates, or range of dates; and
- *Contact Information* such as contact person and/or organization, address, phone, email.

One advantage of standardizing metadata information among government and other data producers is that generic systems can be developed that manage and use metadata. For instance, a range of tools exist for managing the CSDGM. These include entry forms in text, database or web browser format (via the internet or an intranet) and metadata readers that can be used by libraries or Internet data distribution systems. Commercial software vendors have also added documentation tools to their software that facilitate the development of metadata in the CSDGM format.

Definition of the metadata template that is used for the census mapping project is only one aspect of metadata management. The other is the implementation of metadata maintenance procedures. The census organization must decide when and by whom metadata are entered, in what format they are stored—paper forms or digital files—and who supervises the completeness, accuracy and usability of the resulting information. Metadata development should accompany every step of database creation and should not be considered just a final documentation step. For the benefit of future or outside users of the data, metadata should be considered as important as the spatial databases themselves.

Summary and conclusions

Chapter 2 has provided technical content on the step-by-step process of assembling a digital enumeration area (EA)-level database, including geodatabase basics, the mechanics of data input, geographic coding ('geocoding'), and EA delineation.

By the end of this operational stage, the NSO has created a seamless comprehensive coverage of EAs based on the previous census, ready for field updating. At a certain point, the NSO has done about as much as it can do in the office and needs to get out into the field.

The last point is to shore up the decision to produce an EA level database for census canvassing operations – it will make your census more accurate and provide far better inputs for subsequent analysis and dissemination. This will foreshadow Chap 5, the post-censal dissemination of results, especially for humanitarian applications (e.g., disaster preparedness and management).

Chapter 3: Integrating fieldwork using GPS and remotely sensed data

This chapter continues the step-by-step process of constructing an EA-level geodatabase that was introduced in the last chapter. Here we recognize the value of new data sources made possible by satellite technology – namely global positioning systems (GPS) and remote sensing (R/S) (including aerial photography) – by addressing the new tools and data sources directly.

The main topic for Chapter 3 is using GPS and R/S for enumeration area delineation. This is done to field-validate the enumeration area boundaries that were created in the NSO's GIS lab from the prior census's maps. Or, in the case of un-existence of accurate maps, it is done as a basis for EA delineation in the office before conducting fieldwork for completion and validation. With the help of remotely sensed data, geographic analysts can identify territory most in need of updating, and distinguish it from areas requiring minimal updating. Some GPS basics will be introduced along with some guidelines for use of GPS data in the census, including some examples of GPS use for EA and administrative boundary delineation, and the location of housing units and group quarters, and extraction of other features. A section will cover mobile computing and handheld computers. Another section on remote sensing will address both satellite imagery and aerial photography, presenting some basics and guidelines for NSO use.

Ideally, by now in the process, the statistical agency has scanned the EA maps from the previous census and transformed them into a digital spatial EA database. However time-consuming and seemingly comprehensive, the database is effectively a rough draft, since the EA maps

delineated at the census main office have not been field-updated yet. The degree of collaboration between headquarters and the census field offices in this effort will be determined by the level of centralization of census operations, as well as the communications structure and the accessibility of the country. This handbook assumes that headquarters and field operations will be integrated through field-based operations as well as the sharing of data.

Globally the overriding goal of geographic digitization is to harness new technologies to make better maps faster. Once again we retain our focus on the country's particular conditions, paying attention to how territory is divided for administrative areas and for the EAs that the statistical agency produced for the census. Through integration of satellite imagery, analysts and census planners can identify areas that require additional fieldwork, for instance to account for new growth in areas surrounding cities. For planning and logistical purposes, it makes sense to identify these priority areas ahead of time to locate areas of rapid change since the last census and focus on them. This is what is meant by a "change-detection" approach, and it can be most effectively done using a synthesis of field, lab, and remotely sensed data.

Remote sensing (R/S) is defined as the use of imaging sensor technologies for gathering information on a given area or object. R/S is a powerful tool to "see" the census landscape in ways that can dramatically improve accuracy of the enumeration, and it is also potentially a big investment for NSOs. Remote sensing can do more than augment field mapping; in fact it is best used in conjunction with other data sources such as field maps, boundary descriptions, and EAs from previous censuses. Because R/S is a big investment, NSOs need to decide their plans carefully before they commit to obtaining imagery and training. Such a plan will specify data analyses and products beforehand, so that the NSO does not create extraneous data with little or no use to the public. This is especially critical when technology is regarded as a costly add-on

rather than a means to an end. Once the NSO can answer the question of the specific use of the data. Only then can impacts be assessed in terms of resources, especially human resources.

Global positioning systems (GPS)

Once an exotic gadget, the global positioning system (GPS) has revolutionized navigation to the extent that it has become commonplace. GPS in recent years has also transformed census field mapping. By recording latitude-longitudes in an easily used format, GPS makes the addition of location to any application trouble free. As the prices of GPS receivers have dropped – a reliable model can be purchased for less than US $100 – GPS has been integrated in many areas and have gained widespread personal use, with big gains in the consumer market in cars, boats, construction and farm equipment, and built into handheld computers and laptops. The largest professional user groups are in the fields of utilities management, telecommunications, surveying and navigation. But GPS has also contributed to improved field research in areas such as biology, forestry and geology, and also finds increasing application in epidemiology and population studies. GPS is also becoming a major tool in census cartographic applications.

Most of the discussion refers to the U.S. system commonly referred to as GPS, which is the most widely used system and one for which a very large commercial market of receiver manufacturers and surveying services has developed. Other satellite positioning systems, including the Russian GLONASS system, the European Union's Galileo system, and the proposed Chinese Beidou system, are briefly introduced and discussed.

How GPS works

GPS is a worldwide radio-navigation system using orbiting satellites. GPS receivers collect the signals transmitted from a constellation of 24 satellites—21 active satellites and three spares—and their ground stations. The system, which is called NAVSTAR, is maintained by the U.S. Department of Defense. The satellites are circling the earth in six orbital planes at an altitude of approximately 20,000 km. At any given time, five to eight GPS satellites are within the "field of view" of a GPS receiver on the earth's surface.

The position of the GPS receiver on the earth's surface is determined by measuring the distance from several satellites, in three (X, Y, and Z) dimensions. The GPS satellite and the receiver each produce a precisely synchronized signal (using what is called a pseudo-random code). Synchronization is made possible by very precise clocks on the satellite and in the receiver. The receiver measures the lag between the internal signal and the signal received from the satellite. The lag is the time it takes for the signal to travel from the satellite to the receiver. Since the signal travels at the speed of light (299,338 km/sec), the lag time simply needs to be multiplied by the speed of light to obtain the distance.

Once the distance from several satellites is known, position can be determined by triangulation. If we obtain a distance measurement from a second satellite, we can narrow our position down to the two points, x and u, at which the two circles intersect. Normally, these two measurements should be enough, because one of the candidate positions is likely to be very unrealistic. However, to confirm our exact position, we should determine the distance from a third satellite. The distance circles around all three satellites intersect at only one point, which is our true position. Of course, in reality we live in a three dimensional world. With only one satellite distance measurement, we could be anywhere on the surface of a sphere surrounding the satellite. With two distance measurements, we could be anywhere on a circle that is formed by the

intersection of two spheres. Finally, the sphere surrounding a third satellite intersects this circle at two positions. Again, only one of these is usually realistic. However, to improve the position estimate a fourth measurement is taken. This fourth measurement also helps to correct for any existing imprecision in the receiver's internal clock. The satellite's atomic clock is by contrast very precise.

Global positioning system accuracy

Inexpensive GPS receivers can provide reasonably accurate information about the latitude, longitude and altitude of the users position at any place in the world and at any time. According to most vendors of low-cost "autonomous" GPS receivers, the recorded position is normally accurate to around 15-20 meters for civilian applications. Altitude information is somewhat less reliable than latitude and longitude.

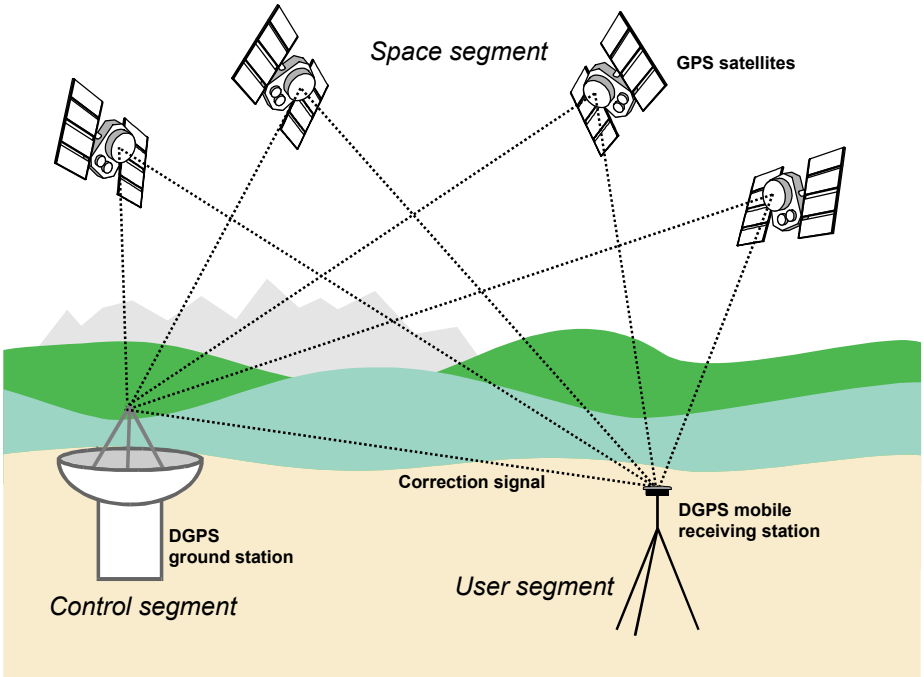
Accuracy is influenced by several factors. One of these is the number and position of the satellites. Ideally these are spread out over the sky to allow optimal geometric computation. Positional dilution of precision (PDOP) refers to the spread of satellites in the sky and can be quantified with a number. Another factor that affects the quality of signal is orbital irregularities. These can be accounted for through ephemeris tables. Lunar influence is another factor, as are atmospheric disturbances that modify the signal as it travels through the atmosphere. "Multi-path error" is caused by scattering of the signals from buildings or other solid objects. Such errors represent more or less random noise—random, short-term fluctuation of the position. Until the practice was phased out in 2000, the greatest source of error was selective availability (SA), instituted by the US Defense Department to reduce accuracy in the signal. Though SA has been discontinued, it could be reinstated during a time of war.

Repeated readings of GPS coordinates will not necessarily improve the coordinate estimate. To obtain more accurate positions one would need to average coordinate readings over a very long time period – i.e., more than 24 hours. In practice there are now better options for improving GPS coordinates.

Differential GPS

For applications requiring higher accuracy, differential GPS (DGPS) systems use correction information transmitted from a base station with precisely known coordinates to correct the satellite signals (See Figure 3.1). DGPS systems are becoming increasingly popular and accessible. DGPS requires the cooperation of two receivers, one stationary and one roving. The signals received by the DGPS base station and the mobile GPS unit are subject to the same errors. The base station receiver measures timing errors then provides correction information to the mobile GPS units. The reference station receives the same GPS signals, and then calculates the travel times of the GPS signals and compares them with what they really are, producing an error correction factor. The accuracy that can be achieved with DGPS depends on the system and coordinate collection procedure. Accuracies of 2m can be achieved with quite affordable hardware and shorter observation times, with even better results in stationary situations. More expensive systems and longer data collection for each coordinate reading can yield sub-meter accuracy.

Figure 3.1: Differential GPS



There are a number of options for implementing real-time GPS correction. Government agencies in many countries are now installing DGPS base stations that continuously broadcast correction information. Such stations are usually located near coastal areas where they support navigation at sea. Relatively inexpensive DGPS base stations are sometimes set up by groups of users, for example in precision farming. Also, some portable high end GPS units can be converted into DGPS base stations that broadcast correction information, increasingly using the Internet. The user needs to find a precisely known location in the vicinity of which precise mapping is then possible. Finally, correction information is also broadcast via geostationary satellites, for example for aircraft navigation.

If real-time data are not necessary, then post-processing of GPS coordinates may be a useful and a less complicated option. Here, the user collects coordinates with a standard GPS receiver. For each coordinate the time and satellites used are recorded in the receivers memory. Back in the

office, the user can download correction information for that time period, applying the correction factors to all collected coordinates. Correction data files are available from a number of commercial or public sources in many countries of the world. Where such information is not available from secondary sources, a DGPS base station can be set up in a central location. To support census mapping, for example, a DGPS station could be set up in the capital, so that coordinate data collected in the field using inexpensive standard receivers can be post-corrected later. In larger countries, multiples of base stations will have to be set up.

Some new national systems worthy of mention makes use of new satellite technology and communications over the Internet. WAAS, or Wide Area Augmentation System, is a continental DGPS system instituted by the US Federal Aviation Association. It includes a geostationary (the term refers to the fixed position of the satellite, as opposed to an orbiting satellite) satellite that broadcasts correction information on a GPS frequency using 24 stations operating in the US. CORS, or Continuously Operating Reference Stations, were also designed for aviation but have many other applications. Each site provides GPS measurements allowing for error correction and enabling positioning accuracies that approach a few centimeters. This service is at the time of publication available only in North America.

Governments in other regions are developing similar satellite-based differential systems. In Asia the Japanese are developing a MSAS, or Multi-Functional Satellite Augmentation System, which will work throughout Asia with accuracy to 3m. Europe has EGNOS, or the Euro Geostationary Navigation Overlay Service, consisting of three geostationary satellites and a network of 34 ground stations. EGNOS began in 2005 and will be certified in 2008 as having better than 2m accuracy. The service area includes Africa and South America.

Other global satellite navigation systems

Several alternatives to the US Navstar GPS system exist. The Russian counterpart of GPS is the GLONASS system, which is operated by the Ministry of Defense of the Russian Federation. Begun in 1976, GLONASS completed its constellation by 1995 but then fell into disrepair, with only 7 satellites in orbit in 2007. The Russian Space Agency plans to restore GLONASS to fully deployed status by 2011, with 24 satellites. A cooperative agreement with the Indian government will see two satellites launched from Indian territory in return to access to high precision signals.

The European Union's Galileo system is being built by the EU and the European Space Agency, with the plan to be fully operational by 2013. Galileo will feature a constellation of 30 satellites to be launched between 2006 and 2010 and two ground stations (in Munich and Rome). One improvement over the US GPS system will be the inclusion of an "integrity message" informing the user immediately of signal errors; another is that Galileo will work in extreme latitudes. Galileo will be interoperable with the US GPS system at the user level.

The Beidou system proposed by China will eventually have 35 satellites, including five geostationary and 30 orbital satellites. Free service with an accuracy of 10m will be offered to those in China; outside subscribers will get more accurate service for a fee. Two additional Beidou II satellites were launched in 2007.

Commercially available GPS receivers vary in price and capabilities. Technical specifications determine the accuracy by which positions can be achieved. The more powerful a receiver, the more expensive it will be. The user needs to decide whether the additional gain in accuracy will be worth the additional cost. In many mapping applications, the accuracy of standard systems is quite sufficient. Receivers also vary in terms of user-friendliness, tracking capabilities which are

useful in navigation — many receivers can now plot maps — and in terms of the map projections and geographic reference systems that are supported. Additional considerations in choosing GPS receivers are the robustness of the units, power consumption (since batteries are expensive, AC adapters for cars are useful), coordinate storage capacity, and the ease of transferring stored coordinates to a laptop or desktop computer.

Most vendors offer integrated products which combine a GPS receiver with a handheld or notebook computer so that the captured coordinates can be plotted on the screen immediately, either in isolation or on a digital base map. These technologies will be covered in a later section on integrated field mapping systems.

GPS in census mapping applications

GPS technology offers many applications in mapping activities, including the preparation of enumerator maps for census activities. With DGPS, accurate geographical positions of enumeration area (EA) boundaries and the location of point features such as service facilities or village centers can be obtained in a cost-effective way. Coordinates can be downloaded or entered manually into a GIS or other digital mapping system, where they can be combined with other georeferenced information. The following section provides some examples of specific census operations that can make use of GPS.

Enumeration areas are the operational geographic unit for census data collection. They can also be used as a unit for the dissemination of census data, but they are mainly for collecting data. Features of EAs include their comprehensive coverage of the country's territory. EAs are designed to represent areas of *equal population size*. They use features such as roads and water bodies as boundaries that can be observed on the ground. Using remotely sensed data in addition

to maps from the previous census will save countless hours of work. Field checking can be kept to a minimum, and resources can be directed to addressing rapidly changing areas.

When delineating enumeration areas, census representatives must understand that ideal EA size includes areal size and size of population contained. One EA represents the amount of territory one enumerator can cover during the period of census data collection. The population threshold for each EA is based on the plan that was developed by the NSO using results from a census pretest to determine the number of days needed for the enumeration.

A most critical element for EA delineation are population estimates. Since EAs are both area and population based, then a system for estimating the number of people in each EA needs to be established that does the most accurate job consistent with resources. Not getting good estimates for EA populations in advance will hamper enumeration and threaten the quality of the results.

Population estimates for EAs can be obtained through the cooperation of local officials. Although the EA units themselves may be unfamiliar, local officials in rural areas can estimate the size of hamlets and villages. Sometimes housing unit estimates are easier than population estimates for small land areas. Failing the participation of local officials, estimates can be made by NSO staff in field visits, or through the use of existing information such as aerial photos, satellite imagery, highway or planning maps, administrative records, population registers, utility company records, or results from the previous census. In the last case, the numbers may have to be adjusted to reflect population growth in the area.

Guidelines for delineating EAs are predicated on the overriding goal of complete coverage, which means overlaps and gaps are minimized. Usually a special symbol such as a squiggly line is highlighted for visibility. EA boundaries follow visible features such as roads, streams, lakes,

railroad tracks; all of which are visible both on the map and on the ground. Some landscape features such as ridgelines and forests make poor boundaries, and these should be used if nothing else exists. "Offset lines" can be used to indicate the need to include housing units on both sides of a road.

EA delineation overall should balance population size with land area and ease of travel. If travel conditions require an enumerator to spend a disproportionate amount of time traveling between housing units, then the size of that EA should be reduced. Natural features such as cliffs, rivers, swamps, and woods can act as barriers to travel, as can human-made conditions such as dispersed settlement patterns, fragmented road system, and poor overall infrastructure.

Some specific GPS-related mapping tasks

1. *EA boundary delineation* – Enumeration area (EA) boundaries are polygonal and are based upon natural landscape features. Given the small unit size of EAs, full ground delineation using GPS is likely to be quite impractical without a triage approach that focuses efforts on recently changing areas. In remote sensing science this is known as a "change-detection" approach. If a country with a population of 20 million requires 40,000 EAs with about 500 people in each, then the sheer effort to record boundaries with GPS units would likely take many years all by itself. A preferred approach is to digitize EA boundaries from the previous census and use a ground-based GPS approach only where needed, particularly when boundary changes such as the creation of new districts or land annexations have occurred.

2. *Administrative boundary delineation* – in most countries administrative boundaries (such as provinces, districts, and subdistricts) have already been delineated in a small-scale (i.e. generalized) way. It is very likely that these units lack the necessary precision for census

work. NSOs should weigh carefully the potential benefit of conducting detailed administrative boundaries at the time of the census against the cost in time and labor. If at all possible, NSOs should gain access to existing digital files for administrative boundaries. These are likely to be available from the national mapping agency. Be sure to enquire about accompanying metadata including datum and projection information before trying to use these data in a GIS project.

3. *Housing unit location* – some countries have gone so far as to record a latitude-longitude for every housing unit in the country, sometimes even photographing each dwelling. Performing such a task for the entire country requires a major expenditure of resources. If done by the census taker in the course of enumeration, it may not necessarily require a large expenditure of time, but the storage and indexing of the files must be done comprehensively to avoid costly re-do's.

4. *"Group quarters" locations (GQs)* – GQs are any kind of communal or institutional housing including hotels, military barracks, orphanages, workers' camps, monasteries, convents, homes for the aged, hospitals, dormitories, and penal institutions. Institutional populations can sometimes be the most vulnerable to natural disasters, and increasingly humanitarian planners are requesting the geographic locations of GQs in order to plan disaster response effectively. Moreover locating GQs with GPS units may be less taxing since they are fewer in number than measuring all housing units in the country.

5. *Other relevant features (including roads)* – Features such as roads can be useful for delineating enumeration areas or for providing navigational information. Water bodies are also useful for orienting census takers, and landmarks can be used as control points for

georeferencing satellite images or auxiliary maps. The NSO should enquire across the government for digital versions of base data such as roads, so that time and money can be saved for the actual enumeration.

For census applications, use of GPS on a wide scale basis should be carefully considered. For many tasks including the recording of readings for every housing unit, and for delineating enumeration areas, the equipment required for a large number of field workers would likely be beyond the resources of a census project.

The exact way in which GPS coordinates are used in census mapping will vary depending on the chosen census mapping strategy. GPS can be used in point mode to collect a coordinate, for example, for each building in a village or for each intersection in the street network of a town. Available maps or sketch maps drawn during data collection will help to interpret the coordinate information back in the office. A second possibility is to collect GPS coordinates in stream-mode, where the system records coordinates at regular intervals. This way, line features can be recorded automatically by walking along a road or traveling on a vehicle or bicycle. If carefully planned, this can be a cost-effective way of creating a street or road network database, although it will depend on the chosen data quality standards whether the accuracy of the resulting lines is sufficient. For safety's sake, as well as providing a backup in areas without regular power supply, retrieving coordinates from the GPS unit and manually recording them on data sheets provides a lower-cost alternative as a backup.

Training requirements for GPS use

For a GPS project to be successful, the NSO must manage equipment purchases carefully, arrange for training, and personnel needs as well as develop point collection protocols (Montana

and Spencer 2004). A GPS coordinator is recommended for such activities. At minimum such a person should understand the GPS units and relevant information related to collecting and storing points. A coordinator should also monitor field workers to make sure they work consistently. A training program for field workers can including understanding the operation of the GPS units, as well as how a receiver calculates a position, and troubleshooting problems with the units.

Ideally, GPS use should be planned years in advance as a part of overall census planning. At a minimum, GPS related projects should be planned out for six months prior to implementation, with data collection strategies developed and posted. At this point in the planning process, hardware needs can be identified. Data tables can be designed with consistent naming conventions for various units and geographies, so that in the database EA or feature IDs are aligned with GPS waypoint codes.

In the event of catastrophic loss of GPS hardware or backups, GPS coordinators should implement a system of backups. This can be as simple as writing latitudes and longitudes on paper copies of survey forms and copying them later into a spreadsheet.

For the purpose of uploading GPS waypoints to a laptop or desktop computer, shareware such as GPS Utility, EasyGPS, and freeware GPSBabel along with manufacturers' software can be used to upload points. Operators can reformat uploaded files in Excel or other spreadsheet for inclusion in a GIS project, at which point additional attribute data can be added. In some programs, GPS can be added to a GIS project by importing a correctly-formatted table or as "event data."

Summary: advantages and disadvantages of GPS

Advantages:

- Fairly inexpensive, easy-to-use field data collection. Modern units require very little training for proper use;
- Sufficient accuracy for many census mapping applications—high accuracy achievable with differential correction;
- Collected data can be read directly into GIS databases making intermediate data entry or data conversion steps unnecessary;
- Worldwide availability; and
- New GPS systems will be coming on line in the next 5 years.

Disadvantages:

- Signal may be obstructed in dense urban or wooded areas (multi-path);
- Standard GPS accuracy may be insufficient in urban areas and for capturing linear features making differential techniques necessary;
- DGPS is more expensive and may not be available in many remote places. DGPS requires more time in field data collection and more complex post-processing to obtain more accurate information;

- A very large number of GPS units may be required for only a short period of data collection, making widespread implementation of GPS potentially very expensive;
- Widespread GPS requires comprehensive planning, including determining what products will result from extensive GPS use; and
- The more complex GPS unit measured, the more training is required.

In the application of GPS, problems can, of course, arise. In dense urban settings, possible multi-path error may make defining adjacent EAs all but impossible. High rise buildings or streets lined with dense trees can make it difficult to receive signals from a sufficient number of satellites, since the satellite signal cannot penetrate solid objects. A trained data collector can still obtain coordinate information by walking to a more open location and applying an offset to the recorded coordinate. In some cases, DGPS must be used or GPS readings have to be cross-checked with additional data sources such as published maps, aerial photographs or even sketch maps produced during fieldwork. Some countries have developed systems of GPS base stations that support very high accuracy mapping using DGPS. In some developing countries such networks do not yet exist.

Integrated field mapping systems using handheld computing

Field data collection in the utility sector and other mapping applications now relies heavily on GIS. In many of these applications, GPS is integrated with a portable computer or personal digital assistant. Coordinates are captured and immediately displayed on the portable computer screen. If a digital base map is available, the coordinates can be displayed on top. Field staff can add any required attribute information and store these data in a GIS database. This GIS

information can then be incorporated in a GIS at the home office. Given that notebook computers and other portable computing devices are becoming less expensive, integrated field mapping systems may soon become a viable option for field data collection for census purposes.

Textbox case study of FIJI experience with GPS:

Fiji's 2007 Census was the first in the Pacific Region to use GPS technology to link census questionnaires with geo-referenced locations for all households in the country. In Fiji, as is the case in many developing countries no exact location database of dwellings existed, and a lack of street names and numbers meant no address lists were available. GPS had a distinct advantage because it increased the accuracy and coverage of household locations, and was used as a management and monitoring tool, (for checking and data validation) streamlining census activities for census managers and enumerators. GPS locations allowed the aggregation of census data into a range of administrative units such as those for health, education and environmental-based applications.

200 Garmin eTrex GPS units and 20 Laptops were purchased for these activities. 10 officers were trained with the GPS and ESRI Arcview, and tasked to carry out validity checks and download waypoints. 200 GPS operators were trained over a three week period. GPS operators and supervisors were given a "cheat" sheet (see photograph), a step-by-step summary of how to: set up the gps unit;record a waypoint;download waypoints from a GPS to the computer;display the waypoints over the image; and export and print JPEG image files.

Field operations were carried out such that GPS operators followed the enumerators and gathered waypoints. These operators were visited regularly by supervisors to check and download this data. Where imagery was available, waypoints were overlaid in Google Earth in order to validate locations. A "three times waypoint number" system was devised to enable household locations and questionnaires to be linked. The enumerator firstly puts a sticker with a unique 6-digit code on the house where the interview takes place and also one on the gate for fenced houses. A small sticker with the same number is placed on the form for that household. A reserve sticker is placed on the gate for households where the front

door and gate are far apart. This facilitates waypoint taking. Secondly the GPS operator comes past the household and must key in the same number as waypoint identification and to later link it to the questionnaire.

Once field work was completed, two databases were created for the waypoints with their latitude and longitude coordinates, and another with the questionnaires. Both databases could be joined using the unique waypoint/questionnaire number. At this stage checks were performed to determine questionnaires which were missing and incorrect waypoints. A team then went out in the field to rectify these issues.

There were a few difficulties encountered, where lessons can possibly learnt for future censuses:

- In early stages of the GPS point collection, operators were taking waypoints before acceptable precision levels were obtained - to solve the problem, waiting time was increased;
- Loss of stickers due to two religious festivals (where houses are repainted for the occasion), meant operators needed to go back to the office to collect names and addresses of residents before returning to the field to gather the waypoints;
- The 6-digit codes were sometimes incorrectly entered into the GPS as a waypoint and into the database from questionnaires. One way to remedy this in the field could be the future use of barcode readers attached to the GPS receivers.
- Another source of error is the fact that the GPS waypoint taking lagged behind the enumeration, sometimes by months. If the enumeration team is also taking the waypoints then these errors will be eliminated.

One of the foreseen benefits of GPS information of this kind for Fiji, is the ability to prepare for natural disasters and manage them when they occur. Overlaying basic household information on a digital terrain model shows very clearly populations affected by any given disaster.

Advances in technology including GPS, wireless communication, and computer miniaturization have made possible numerous new applications for handheld GIS, particularly the development of GIS software for census fieldwork. A host of palm and pocket devices support many display, query and simple analytical applications, with programs and data stored in memory because handhelds contain no hard disk drives. Other features here include communication via Bluetooth and/or WiFi wireless connectivity; synchronization with a PC to allow quick uploads of data and updates, also ensuring backups to prevent loss of data. GIS software has been developed for use on smartphones. Ruggedized PDAs can be used for mobile data collection in extreme environments. Software for handhelds includes 'thin' versions of popular office applications, and some examples of GIS software used for handhelds include Autodesk OnSite, ESRI ArcPad, and Intergraph Intelliwhere.

One caveat must be addressed for NSOs contemplating the use of handheld computers for either precensal mapping or the actual enumeration, and that is cost. GPS-equipped handhelds can easily cost US $750, or more for ruggedized computers. For those managing geographic operations, the operative question is: what is the extra value of using a $750 (or more expensive) unit as opposed to a $100 GPS unit? Some factors include on-screen map readability, power requirements especially in areas without dependable electricity, and other adverse environmental conditions for computers, even ruggedized ones.

Satellite remote sensing

Using imagery to field-verify enumeration area maps produced at census headquarters

Since the publication of the first edition of the GIS Handbook in 2000, remote sensing data have gained in volume, popularity and ease of use. Particularly since the advent of satellite imagery

with high (1m or better) spatial resolution, R/S has revolutionized mapping. It's about time to make use of this valuable resource for census work. One challenge for NSOs is the sheer amount of territory to map. Satellite imagery if used pragmatically can save countless person-hours by allowing the NSO to focus attention on critical areas. We advocate an approach that performs triage on the surface area of a country, partitioning it into areas needing more and less attention. This is what is meant by a "change-detection" approach. Following from the schema introduced in Chapter 2, the NSO has digitized old census EA maps and overlaid other geographic data and has entered all of it into a geodatabase, but it has not yet corrected the results in the field. With the provisional EA boundaries superimposed on R/S imagery, population settlements can be quickly located and priority areas identified. Planning for such activities must however be detailed and realistic.

Principles

In geographic information science, satellite imagery represents one of many forms of geographic data that can be used for analysis and presentation of results. Imagery is formatted as a continuous field, also called a raster, with rows of data corresponding to pixels representing values. R/S is considered primary data capture; in fact it is the most popular form of primary raster capture. Since the advent of orbiting satellites, great potential has been realized for taking readings of the earth surface, with numerous current and planned applications for census work. One of the main strengths of R/S for census work is that it can cover dangerous or inaccessible areas and thereby save countless hours of fieldwork, if done properly.

Satellite images are collected from space based systems, most of which use passive optical sensors to measure radiation reflected from objects on the earth's surface in the visible and

invisible electromagnetic spectrum (Figures 3.2 and 3.3Figure 3.3). Most satellite data collection is considered passive in that it receives emitted energy from the earth, contrasting with active sensors such as RADAR. Satellite systems do not use photographic film to record the reflected energy. Instead, an electro-optical detector array—similar to a CCD (charge-coupled device) camera—measures the intensity of electromagnetic radiation and records it digitally as a regular raster or image of rows and columns.

Satellite sensors operate in multi-spectral or panchromatic mode. Multi-spectral means that the satellite collects several images (or bands) each of which measures reflected energy in a different part of the electromagnetic spectrum, usually in the visible and near infrared range. The ability to separate an image into different spectral bands and to combine specific bands in image analysis facilitates the classification of features on the ground according to their reflectance properties. For example, rice fields may show a strong signal in one particular band, while built-up areas will appear most clearly in another. *Panchromatic* satellite sensors capture reflected energy across a wide range of the spectrum. The resulting images are similar to black and white photographs. They also usually provide higher resolution than multi-spectral images and are therefore the preferred basis for census mapping applications.

Figure 3.2: The remote sensing process

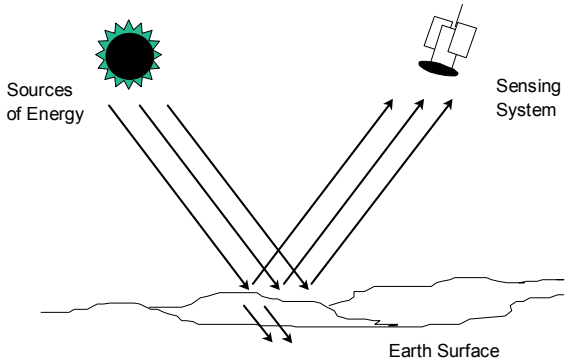
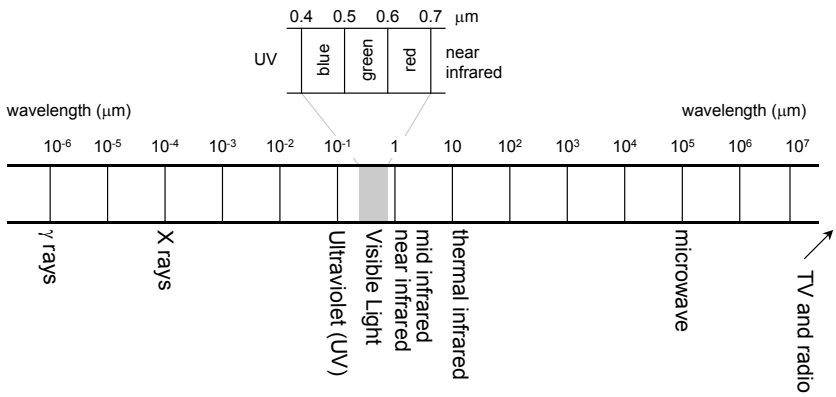


Figure 3.3: The electromagnetic spectrum



The digital data produced by the sensor systems consists of an array of numbers that indicate the level of energy reflected at the corresponding location on the earth's surface. The satellite sends these data to one of a system of earth receiving stations, where they are geometrically corrected and georeferenced. The resulting digital or printed images can be interpreted visually similar to air photo interpretation, or they can be analyzed using geospatial techniques or combined with other layers in a GIS project. Digital satellite images can be displayed in a GIS where features on the image can be delineated by a skilled operator. For many applications such as land use surveys or natural resources management, multi-spectral images are classified using statistical techniques. These predict land cover classes based on a calibrated relationship between control sites of a known category and their spectral signature.

Resolution

1.1. The spatial resolution of a satellite image is measured by the size of a pixel on the ground. There are also other measures of resolution including temporal, radiometric and spectral. Pixel size for commercial satellites varies from the sub-one-meter of the most popular high-resolution systems such as Quickbird and Ikonos; Indian Remote Sensing (IRS), SPOT's panchromatic sensor and Landsat multi-spectral imagery are also considered high resolution systems which allow mapping at cartographic scales of 1:25,000 to 1:50,000 or smaller. At the time of this writing in March 2008, the remote sensing company GeoEye plans a launch of a sensor with a pixel size of 0.41 meters.

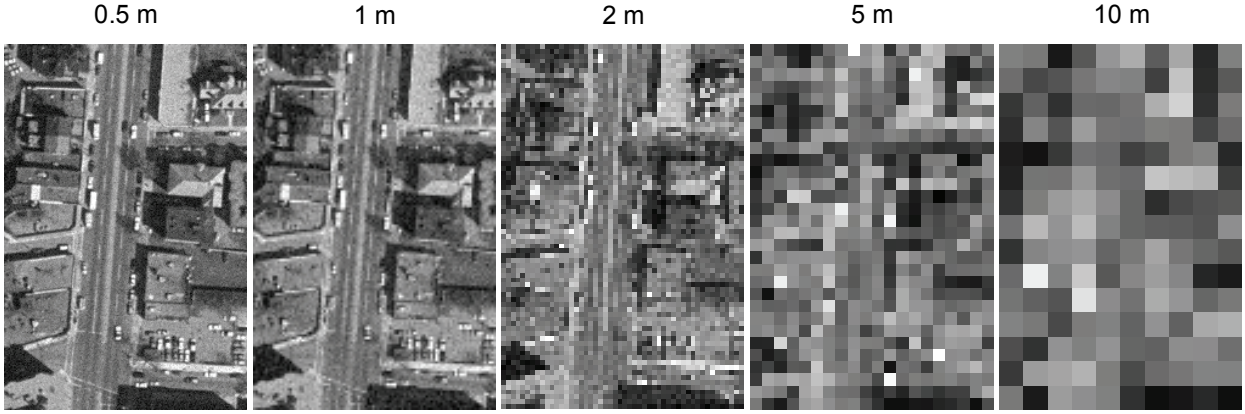
Figure 3.4 compares pixel sizes that were simulated from a 0.5 meter resolution digital air photo by aggregation. The image covers an area on the ground of 100×150 meters. Individual houses and even cars are distinguishable at a resolution of 2m but not with larger pixel sizes. More information can be extracted from remote sensing data by using advanced image processing methods including edge detection and special filtering algorithms. Such techniques have been used successfully for mapping and change detection of newly built-up areas in some fast growing cities in the developing world.

Among the varieties of imagery available, from low spatial resolution/high temporal resolution METEOSAT to radar to moderate spatial and temporal resolution AVHRR imagery, satellite imagery serves many purposes. Only a relatively narrow range of R/S products serve census work well, namely high spatial resolution data such as Quickbird (0.82m), Ikonos (1m), IRS (5.8 m pan), Orbimage 3 and 4 (1m), and SPOT 5 (2.5m). Ikonos was launched in 1999 and Quickbird was launched in 2001.

For most census applications, 5m or better spatial resolution is needed to identify housing units and the spread of population settlements, with multi-spectral imagery less absolutely necessary. The downside of these data sources is the relatively small footprint of high resolution imagery, meaning that numerous scenes would have to be procured to cover even a fraction of a country's territory.

In some cases, Landsat (with a spatial resolution of 30m, or 15m for ETM) and ASTER may be useful for documenting land cover change and human impacts; these data sources are less useful for census work. In deciding which R/S products to use for enumeration, if any, an NSO should carefully evaluate objectives and costs. One option to consider, especially for large countries, is using a mosaic containing 1 or 5m imagery for densely populated areas like cities, with Landsat coverage for surrounding rural areas. In such cases, rural areas will still need to be mapped on the ground in order to best plan logistics during enumeration.

Figure 3.4: Illustration of pixel size in aerial photographs and satellite images



Most of the commercial operators provide several options for acquiring satellite images. Generally pricing is based on whether images are *archived* or collected for the first time. The most expensive option will be special requests for urgent image acquisition for a particular area. With their higher resolution, these satellites cover a smaller area on the ground so that they only cover selected regions along the flight path. A less expensive option is to obtain images on a less timely basis, using image archives, parts of which can be purchased at significantly lower cost. The price of imagery will also depend on the degree of processing of the raw data. This may include radiometric correction, geometric correction, and geo-referencing without or with ground control points. For example, archived imagery available from Digital Globe (prices are from 2008) is US $16 per square kilometer, with a minimum order of 25 square kilometers. Ikonos imagery tends to be less, sometimes around $8 per square kilometer. Bear in mind that prices will depend on the amount of area being covered in the imagery purchase, with larger areas costing less per square kilometer. Raw image data will be considerably less expensive than a digital orthophoto map produced from satellite images. Normally however, images are purchased now fully processed. See Table 3.1 for a listing of very high spatial resolution civilian satellite remote sensing products.

Product	Company	Launch	Mode	Pixel Size at nadir	Height (km)
Quickbird	Digital Globe	2001	Pan/4ms	0.61/2.44	450
Ikonos 2	GeoEye	1999	Pan/4ms	0.82/3.28	680
OrbView 3	OrbImage	2003	Pan/4ms	1.0/4.0	470
Spot 5	SpotImage	2002	Pan/4ms	5(2.5)/10	830
Cartosat-1	NASDA, Japan	2004	Pan	2.5	617
Cartosat-2	NASDA, Japan	2004/5	Pan	1	630

Table 3.1 Very high spatial resolution civilian satellite remote sensing products

Imagery can be ordered online or through a local or regional reseller. Another and perhaps lower-cost option is to use the spatial data infrastructure of the country to gain access to a common archive of imagery. NSO representatives can contact other agencies, particularly the national mapping agency, to enquire about using existing R/S archives for census work, signing a use agreement when necessary which specifies exactly how the imagery can be used.

Online data sources

Arguably as important for NSOs are the new internet-based satellite imagery resources, which can serve as a visual aid for census work while eliminating the cost and trouble of purchasing imagery. Remotely sensed data can now be accessed online using applications such as Google Earth, GlobeXplorer, ArcGIS Explorer, and Microsoft Virtual Earth and other online data

sources. The advantage of using online data retrieval versus buying imagery is the NSO can test the applicability of imagery without the upfront investment. The downside is that the resolution and overall image quality may not be suitable for detailed EA mapping.

Google Earth is a virtual globe program that maps the earth by cataloguing and displaying satellite imagery. Google Earth has taken the geospatial community by storm and is driving public interest in satellite technology and maps. The application was developed by Keyhole, which then was acquired by Google in 2004. Most imagery included in Google Earth is from Quickbird of Digital Globe, although some aerial imagery and 3-D buildings are now included. Google Earth offers 15m spatial resolution or better for most parts of the world, using a geographic projection and the WGS84 datum. Images on Google Earth are protected by copyright.

Currently, Google Earth offers three levels of licensing: the free Google Earth viewer, a Plus version for a subscription of US $20 per year, and a Pro version for commercial applications for $400 per year (prices are from 2008). The Plus version includes GPS integration, allowing the user to read tracks and waypoints from a GPS device. Additionally, Google Earth Plus provides direct support for the Magellan and Garmin GPS product lines, higher resolution printing, customer support via email, and a data importer that can read address points from a spreadsheet using comma-separated values, but this is limited to 100 points/addresses.

The functionality of the Pro version includes add-on software such a movie-making and allows the user to represent location-based data using 3D drawing tools. Users can also transfer up to 2,500 point locations from a spreadsheet. A GIS data importing module lets the user add geographic data including demographic data in shapefile and .tab formats. Measurement tools

available in the Pro version allow the user to calculate areas and linear distance, and images as large as 11 by 17 inches or 4,800 pixels can be exported. Google Earth Pro is not for sale online but must be purchased through a sales representative.

In any version of Google Earth, geographic data in the .kml (Keyhole markup language) format can be exported to Google Earth. A free downloadable script can convert points and polygons to the proper format, although polygons may be more difficult to import. Rudimentary topology, i.e. no polygon attribute tables or node locations, means that problematic data may not import. The Pro version can import but not export shapefiles. Scanned maps for census use, for example EA maps, may present problems if imported into Google Earth, since there is a fairly severe limitation on file size (18,000 by 18,000 pixels). This can be overcome by using a "regionator" Python script, but this might create file management challenges.

Google Earth, particularly the Pro version with its image exporting capabilities, may perform some useful tasks for NSOs, although functionality is severely limited compared with standalone imagery. Strengths of Google Earth include its low cost and ease of use for some low level tasks. The 15m spatial resolution of most imagery can allow EA delineators to view the landscape in *some* detail, although not enough to count housing units. Weaknesses of Google Earth include the lack of necessary resolution to use with EA delineation, the difficulty in transferring imagery into a GIS program, the need for a high-speed Internet connection to download imagery, as well as metadata and authenticity. Other online satellite data sources such as those obtained through ESRI's free ArcGIS Explorer application may circumvent some issues with Google Earth by allowing direct import of imagery into GIS projects. But the spatial resolution of the available imagery might still not be sufficient for some census applications.

High resolution satellite images show a level of geographic detail that is similar to digital orthophoto maps created from air photos. One major complication is that it is more difficult to obtain cloud free images from satellites than from low-flying airplanes which operate on a flexible schedule. Cloud free high-resolution images allow counts of housing units, population estimation and EA delineation. Aerial photography is often done on ad hoc basis, and may be more suitable for detailed surveying and mapping projects. New digital aerial photography is gaining in popularity and may offer superior image quality over even high-resolution satellite imagery. This will be covered in the section on aerial photography.

Applications

Population analysis using remote sensing is in its infancy, but gains are rapidly being made. A National Research Council report (NRC 2007) characterized proxy methods for estimating populations as still insufficiently robust for humanitarian response. High spatial resolution sensors such as Ikonos and Quickbird do not have the depth of archival data that Landsat has, and coverage for large areas can be very expensive. But as was discussed, Landsat is problematic for estimating population size and also has long-term problems with its sensor, in addition to funding variability. Programs to replace Landsat with other high-resolution data such as SPOT are likely to be available to researchers in the near future.

Characteristics of population that can be assessed using satellite imagery include counts of dwelling units, measurement of urbanized land areas (settlement size), and estimates of land cover/land use as proxies for residential spread and population density (Jensen and Cowen 1999). In some disaster prone areas, fixed-wing aircraft has an advantage over satellite imagery in its

ability to fly below clouds. Radar has yet to play a significant role in population analysis, although it has the advantage in that it can penetrate clouds.

A "change-detection" approach can be used to measure population change spatially, and particularly urban growth, using two or more images of the same place over a time period of 5 or more years. For quantifying the spread of urbanization, analysts categorize each image using a hard classification technique, so each pixel is considered either urban or non-urban depending on the spectral signature. Changes in land cover can then be calculated by overlaying the images and measuring the growth of area transformed by human occupation.

A case study (Yankson 2004) cited by Antos used Landsat TM data from 1985, 1991, and 2002 to calculate annual rates of areal spread of Accra during that time period. Yankson found that between 1984 and 1991, Accra grew by about 10 square kilometers per year. This increased to 25 square kilometers per year between 1991 and 2001. This study measured the growth of areal extent alone. To get at significant internal changes that might be development related, one has to adopt a soft classification approach which measures urban buildup as a continuous variable. Instead of classifying a pixel as urban or not, the analyst can use additional sub-pixel data to classify each pixel as having a degree of urbanization as represented by a percentage. This pleads for a continuous approach in using GIS for census mapping operations.

Health studies have used remotely sensed imagery to focus on intra-urban disparities in phenomena such as disease prevalence. Castro (2004) used aerial photography and neighborhood polygons to identify potentially harmful malaria breeding sites. Detection of informal settlements and neighborhoods can be delineated based on pattern recognition of the settlements' particular signature, which can include dense rooftops, low vegetative growth, mostly dirt roads and paths.

In general, informal neighborhoods tend to have minimal texture, meaning low variability in brightness, and high concentrations of impervious surfaces (Weeks 2007).

Finally, the most advanced population applications of remotely sensed data use object-based rather than pixel-based analysis. Pellika (2006) illustrated the use of an automated application to segment a high-resolution image into areas of similar size, shape and color, then labeled each area as a particular surface type, for example, "rooftop." Such a technique would then allow all rooftops of a particular type to be grouped and labeled as part of an informal settlement. Applying the technique to several structures, Pellika was able to achieve 97 percent accuracy. Some drawbacks to object-based image analysis (OBIA) are that while automated, it is still time consuming; additionally, it is site-specific and dependent on having detailed high-resolution data. At the moment, analysis requires expensive software and expertise that is all too often beyond the reach of most NSO staff.

As with air photos, acquisition of satellite images — though normally less expensive than aerial photo surveys — can still be quite expensive. If possible, high resolution satellite data should thus be obtained in a cost-sharing arrangement with other agencies or it could be employed selectively in areas with insufficient map coverage.

Advantages and disadvantages of remotely sensed data –

Advantages:

- Up-to-date coverage of very large areas at relatively low cost with lower spatial resolution images;

- High spatial resolution images offer the ability to cover large areas at a level of detail sufficient for EA delineation, provided population estimates exist for the areas delineated;
- Imagery can permit mapping of inaccessible areas;
- Imagery has multiple uses and once purchased can be used in other applications;
- Online imagery sources may offer a degree of functionality for a low (or no) cost; and
- Update of topographic maps in rural areas is possible; e.g., identification of new settlements or villages that are missing on maps.

Disadvantages:

- The spatial resolution of many systems, especially low cost ones, is not sufficient for census applications;
- Cloud and vegetation cover restricts image interpretation;
- The problem of low contrast between features — e.g., dirt roads and traditional building materials in rural areas — makes their delineation particularly difficult in developing world contexts; and
- Image processing requires a large amount of expertise – which may not be available at the NSO. Amidst other human resource challenges at NSOs, agencies may decide to leverage expertise from elsewhere in order to perform its mapping tasks.

Aerial photography

Air photo overview

Even with the rise of high-resolution satellite imagery, aerial photography continues to be useful for mapping applications that require high accuracy and a fast completion of the tasks. Photogrammetry — the science of obtaining measurements from photographic images — is used to create and update topographic base maps, carry out agricultural and soil surveys, and for many aspects of urban and regional planning. Census projects have also frequently made use of air photo surveys to quickly create maps for areas for which up-to-date maps are not available or which are difficult to survey using traditional field methods. An aerial survey flown shortly before a census will provide the most complete basis for the delineation of enumeration areas within a reasonable short time frame.

Air photos have been used for mapping shortly after the invention of airplanes. Early applications made use of standard cameras. Very soon however, specially constructed camera systems that minimize geometric distortion were mounted on specially adapted airplanes that allows the camera system to face straight down to the ground through a hole in the aircraft's floor. Equipment for interpreting air photos and for converting information extracted from such photos into maps quickly became very sophisticated. For instance, interpretation of stereo pairs of images became the dominant method for producing maps of elevation contours.

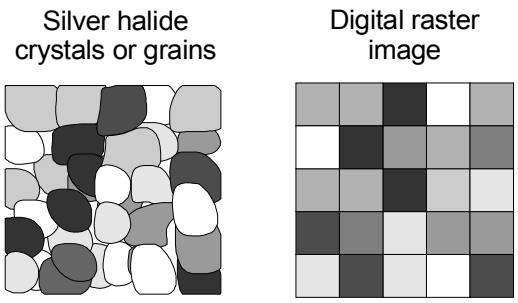
Aerial photography is obtained using specialized cameras on-board low flying planes. The camera captures the image on photographic film, or digitally. In comparison to digital sensor systems, film traditionally has provided a superior resolution (i.e., the ability to distinguish small details), although recent developments in the area of digital imaging have changed this.

According to literature produced by the aerial survey company MJ Harden, the newest state-of-the-art sensors can capture 12-bit imagery with a ground resolution as small as 1 1/2 inches per image pixel, with 4,096 grayscale shades versus 256 from film. South African company Rob Wooding and Associates compare digital aerial imagery favorably with 1m satellite imagery, finding the aerial imagery to be both less expensive and more precise.

Traditionally, the end products of an aerial photography project are printed photos of an area on the ground. The air photo survey is designed so that the resulting photos overlap by between 30 and 60 percent. The photogrammetrist can combine these photos to produce a seamless mosaic covering the entire region. Printed air photo mosaics can be used in the same way as maps. They can be annotated, provide a reference for field work, and allow digitizing of features to create or complement GIS databases.

Recent advances in digital image processing have changed the format in which aerial images are transformed into useful products. In analog systems, the photo is usually an intermediate product. The most common approach is to convert the photo negative into a film transparency which is scanned using a very high resolution scanning device. The result is a digital image that can be displayed and processed further on a computer. The microscopic difference between the photograph and the digital image is shown in Figure 3.5. Black and white photographic film, for example, consists of a layer of gelatin in which tiny, light sensitive silver halide crystals are embedded. These crystals or grains are irregular in shape and size. This scanned image, in contrast, is a regular array of pixels (picture elements).

Figure 3.5: Photographic film versus the scanned image

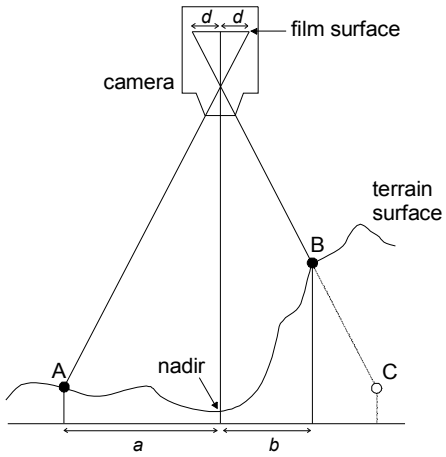


Aerial photographs are similar to maps and satellite images as they provide a top-down view of features on the earth's surface. They are different from maps in that they only show features that are actually visible on the ground. Artificial boundaries, thematic information and annotation are, of course, absent. Without further processing, air photos also do not provide the geometrical accuracy of a map. Camera angle and terrain variation distort the view of an air photo. Additional processing is therefore required to produce orthophoto maps which combine the geometrical accuracy of a topographic map with the large detail of a photograph (see Box).

Box: Development of digital orthophoto maps

To produce map-like digital orthophotos, distortions in the image that are due to camera angle and terrain need to be removed. The distortion introduced by terrain variation is illustrated in Figure 3.6 (after Jones 1997). The photograph is essentially a perspective projection of the earth's surface. Point B is at a higher elevation compared to point A. In reality B lies at a distance *b* from the nadir, which is the point vertically beneath the perspective center of the camera lens. However, the perspective projection in the camera gives a misleading impression. B appears to be located at point C, and therefore projects at the same distance *d* from the center of the film's surface as point A.

Figure 3.6: Distortion due to terrain



To correct for the distortions in the aerial photo, we therefore need to know the elevation at every point on the ground. Elevation can be determined from stereo pairs of air photos. These are photographs that cover approximately the same area on the ground, but which are displaced by a small distance. Analytical stereoplotters allow the operator to correctly co-register the stereo-pair of images and to extract feature locations in three dimensions. State-of-the-art softcopy mapping systems support a high degree of automation for registration of images and removal of distortions. All relevant parameters such as camera tilt during the flight and lens distortions can be considered. The operator can thus extract correctly georeferenced digital data from the air photos. Output products include vector GIS data directly generated from the air photos, wireframe maps showing the terrain, or digital elevation models (DEM) — a raster image corresponding to the air photo, where each pixel value indicates the elevation of that point on the ground. While a DEM is only moderately useful for census mapping applications, such data sets have considerable utility in environmental and natural resources applications, especially in hydrology.

After this process of registration in a proper geographic reference system and distortion removal, the initial air photos have been converted into digital orthophoto maps. These are usually produced at map scales of 1:2,000 to 1:20,000 depending on airplane altitude and processing. Neighboring orthophotos can be digitally combined to create seamless image databases for an entire city, region or indeed a whole country. Mapping technicians can extract or delineate features on these orthophoto maps through on-

screen digitizing. Or they can simply be used as a backdrop to provide a context for existing GIS data layers.

Implementation and institutional issues with aerial photography

The construction of digital orthophotos requires considerable expertise in photogrammetric methods, which is not usually present in a census organization. The census organization therefore needs to establish a collaborative agreement with another national agencies, most likely the mapping department or an air force reconnaissance unit. Alternatively, the work can be contracted out to a commercial aerial mapping company. There are several internationally operating mapping companies who provide the airplane, camera and processing equipment.

These services are not cheap, however. Fortunately, air photos are useful for many different applications including planning of service provision, updating of town maps and land titling projects. Cost sharing among interested government departments and possibly with the private sector can considerably reduce the expenses for the census organization. Where complete national coverage of air photos is not possible due to resource constraints, they could still be produced for specific areas. An example is the use of aerial photography to estimate the number of people living on boats by the statistical office of Hong Kong (NIDI 1996). This illustrates the use of these techniques for counting populations that are hard to enumerate. Other examples are nomadic or refugee populations, rapidly growing urban areas, or regions that are seasonally inaccessible.

As described in the previous paragraphs, the development of orthophoto maps requires considerable technical expertise and specialized equipment. The use of orthophoto maps, in

contrast, does not require significant additional training. A database for a city, for instance, may simply consist of a mosaic of several images on a portable storage medium such as a DVD that can be displayed seamlessly in a standard GIS or desktop mapping package. The digital orthophoto maps can be obtained in standard graphics formats (such as Tagged Image File Format or TIFF). The user therefore does not need specialized image processing software. In fact, any graphics package could be used to extract features from the images, although the georeferencing information will be lost. This information consists of the dimensions and real world coordinates of the digital image and is usually contained in a small header file. With this information most desktop mapping packages are able to register the images with any other GIS data sets that are stored in the same geographic reference system.

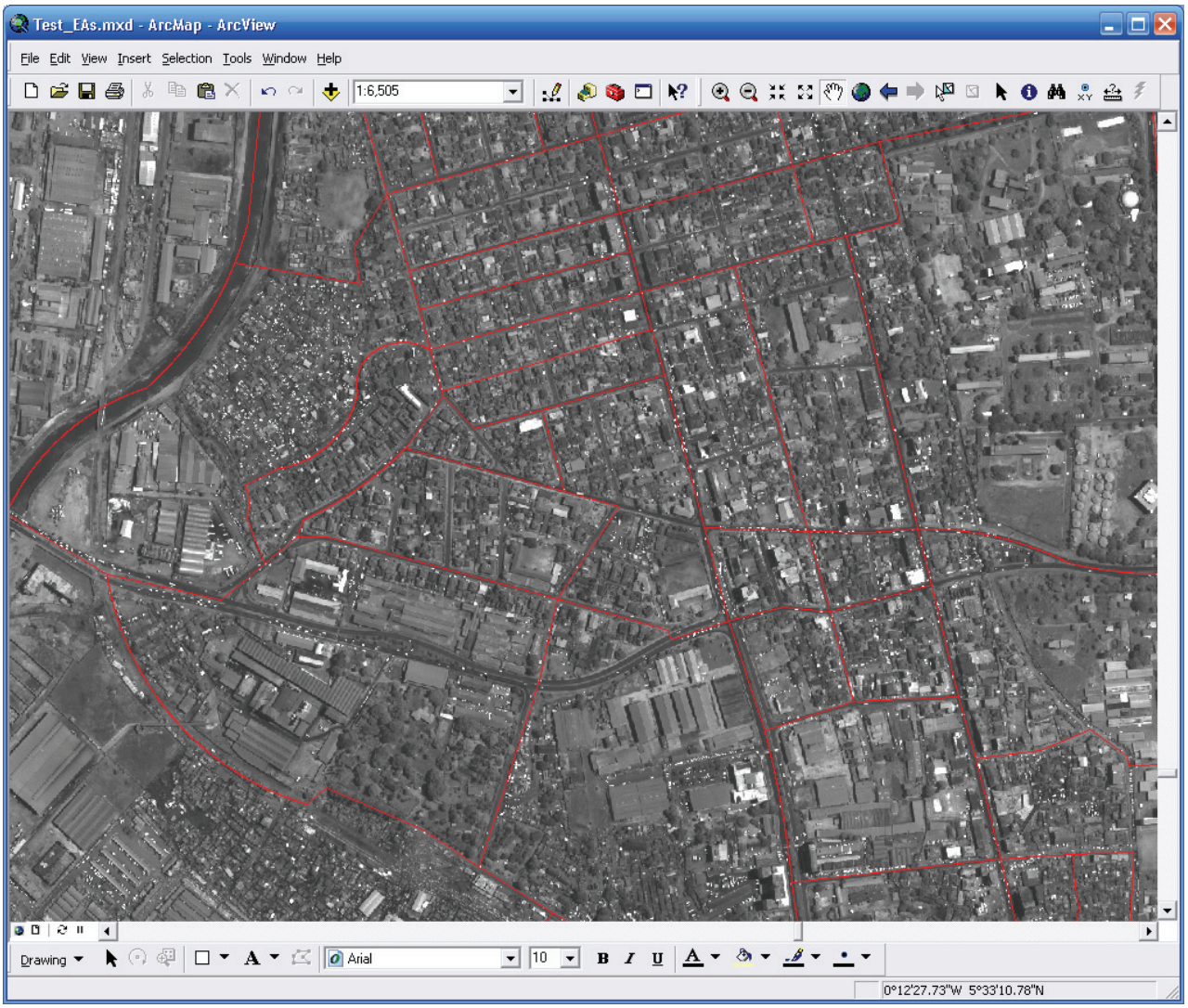
Application of air photos for census mapping

Orthophoto maps are well-suited for dwelling unit counts and population estimation. Dwelling or population counts by means of air photos are sometimes called rooftop surveys. In a rural setting, where settlements are clearly distinguishable on the aerial photo and houses are more or less scattered, the number of dwelling units can be determined fairly easily. A reliable estimate of the average number of persons per household then allows a sufficiently accurate estimate of population for census purposes. In urban settings, houses may be very close together. The number of families living in multi-story homes may also be difficult to determine. Even so, with some training and knowledge of the area, it will still be possible to achieve a sufficient degree of accuracy in the population estimates. Census staff can then delineate enumeration area boundaries that include a specified number of housing units. Since the orthophotos are correctly georeferenced, the resulting enumeration areas will also be registered in a proper map projection

with known parameters. This means that possibly tedious georeferencing to make the digital boundaries compatible with other GIS data will be unnecessary.

Air photo interpretation is most often based on visual interpretation. Census cartographic staff therefore do not need to be trained in advanced image processing techniques. EA boundaries can be delineated on the air photo. Additional geographic features that provide the geographic reference for the enumerators can also be extracted from the photos. These features can be delineated interactively on the computer with the mouse or a similar pointing device (see Figure 3.7). Alternatively, census staff can print the photos and trace features on clear (acetate or mylar) plastic film sheets. These can then be scanned and vectorized. This process requires an additional step and more materials, but often improves the accuracy of the resulting output product (see also the sections on digitizing and scanning).

Figure 3.7: Interactive delineation of EA boundaries on a panchromatic satellite image



Orthophoto maps are also useful as a backdrop to provide a context for the display of point locations collected using GPS or digitized features such as health facilities and transport networks. In the past, in addition to the EA maps, enumerators could have been issued prints of digital orthophotos that show the EA boundaries to support orientation in their assigned area. Today however the remotely sensed imagery is more likely to be built into a GIS project than included as a separate map.

One problem that inhibits the application of this technology in census offices is the large data volume involved in working with high resolution digital orthophoto maps for large areas. For a census office it may thus be better to obtain coarser resolution digital air photos, which show sufficient detail for census applications and will be easier to process and store. Digital orthophotos often have very high resolution, with pixel sizes on the ground in the centimeter range (usually 5-30 cm). Resampled digital orthophoto images with pixel sizes between 0.5 and 2 meters are sufficient for delineating EAs in urban areas.

The future of aerial photography will be a fully digital process, thus eliminating the need to produce intermediate printed photographs. Systems that use in-flight GPS control and digital frame cameras are operational. Digital frame cameras use arrays of charge coupled devices (CCDs) that can create images of 9,216 by 9,216 pixels with a positional accuracy of 1-4 centimeters. Since the intermediate steps of producing photographic prints and subsequent scanning will be removed, this technology is considerably cheaper and faster than traditional photographic technology. Digital camera resolution will continue to steadily increase as will computer processing speeds. Accurate, real-time, and

fully digital aerial mapping is therefore likely to replace conventional aerial photography in the future.

Summary: advantages and disadvantages of air photos

Advantages:

- Air photos provide a large amount of detail and can be interpreted visually. Information about many types of features—roads, rivers, buildings—is shown concurrently;

- Data collection is faster and map data can therefore be produced much more quickly than using cartographic ground surveys. Recent air photos are therefore a more reliable basis for census mapping compared to maps which are updated infrequently;

- Air photos can be used to produce maps for hard-to-reach areas or areas in which field work is difficult or dangerous;

- Topographic mapping using aerial photography can be cheaper than mapping using traditional surveying techniques. However, since the accuracy requirements for census maps are lower than for topographic mapping, the considerable costs are not necessarily justified if the products are used for census mapping only; and

- Printed air photos are useful in field work to provide the "bigger picture". Field staff can see the terrain that is visible from their viewpoint in the wider context of the surrounding area. Digital air photos are useful as a backdrop in the display of GIS data sets.

Disadvantages:

- Aerial photo processing requires expensive equipment and specialized expertise. Census offices therefore need to rely on outside support;
- Air photos still require information on the names of features which need to be extracted from possibly outdated maps. Aerial photography does not necessarily make field work unnecessary;
- Air photo interpretation may be difficult where features are hidden under dense vegetation or cloud cover, or where limited contrast provides no clear distinction between adjacent features (for instance, between homesteads made of natural materials and the surrounding ground); and
- Digital air photos consist of very large amounts of digital data and therefore require fairly powerful computers for display and further processing.

Summary and conclusions

Chapter 3 has discussed the process of utilizing tools such as global positioning systems and remote sensing in census fieldwork, integrating them with ground-based work. Making use of these geospatial tools can allow the NSO to focus efforts on rapidly changing areas within the country. By this point in the census process, the NSO has completed a geodatabase of enumeration areas. The next step is to design, print, and distribute maps from the geodatabase for use in the actual enumeration.

Chapter 4: Use of Geographic databases (maps) during the census

This chapter will cover process flow for enumeration maps, especially how to use the versatility of the geodatabase to get the right information in the hands of enumerators. Covered here will be map compilation, determining the relevant layers for enumerators and supervisors, and printing and distribution basics.

Overall, the chapter will adopt a project management approach to enumeration, which is plan-centric and asks census personnel to make contingencies in the event of delays. Above all, the census is a territorial exercise, whereby the NSO divides the country into operational units that can be canvassed. For the purpose of mapping, the task is to get the maps out of the EA database and into the hands of enumerators. The philosophy here is that the database that has been edited continuously will become the basis for the maps that enumerators take into the field.

A digital enumerator map has the virtue of being modifiable for the specific context. Building EAs digitally allows the NSO to have a living document, "building on the shoulders of giants." With map compilation from previous censuses and all fieldwork finding a place here in georeferenced form, census planners can focus on the areas in need of the most attention, following the change detection approach introduced in Chapter 3. If it has done its work right, the NSO will have everything it needs at this point except the count. With the logistics planned in detail, error is minimized and slowdowns averted.

Quality assurance, EA map production and database maintenance

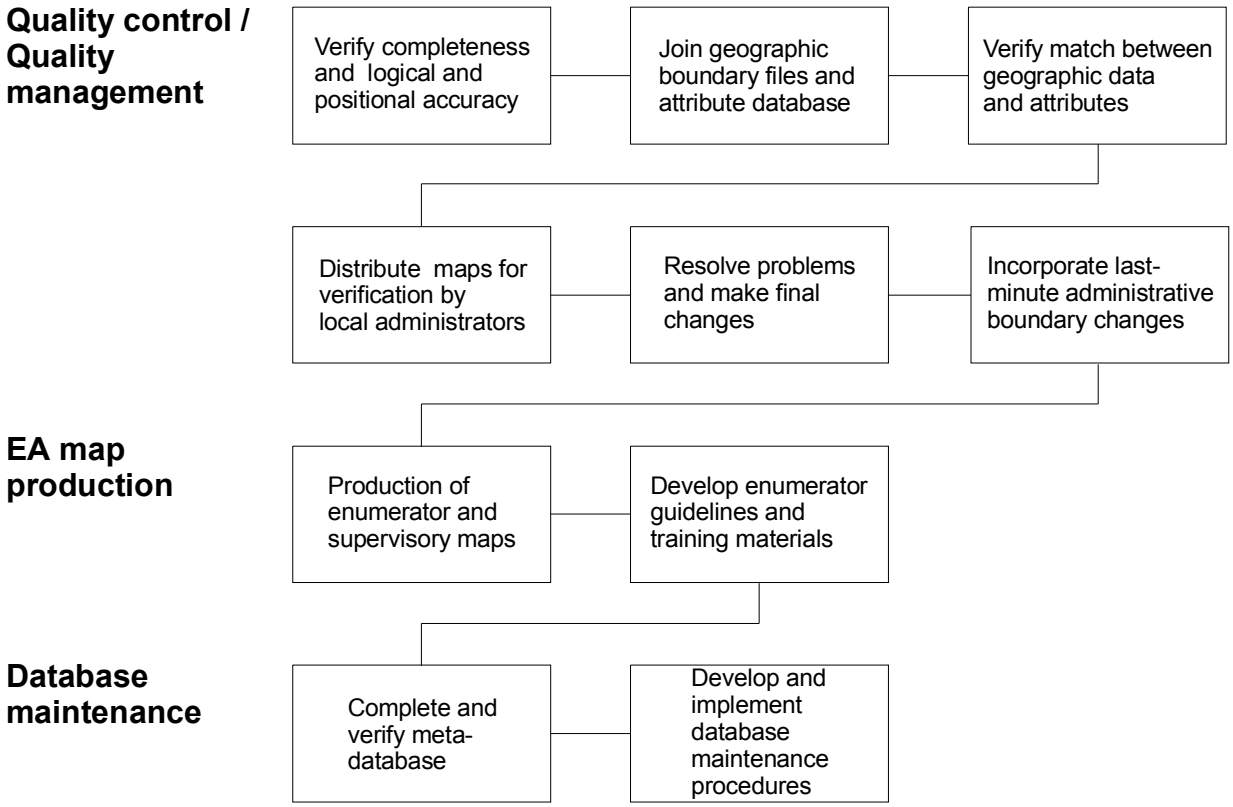
Overview

The accuracy and completeness of census data depend substantially on the quality of the cartographic base maps used by enumerators. In addition to a continuous process of quality control and quality improvement during data conversion, a final step before EA maps are distributed to the enumerators is a thorough review of all map products. This will also involve verification of the correctness of administrative boundaries by local administrators. Any remaining problems and inconsistencies must be resolved before the final products can be generated.

At this point field offices for the census will have been established. The level of centralization of the census organizational structure will have a direct influence on the procedures for distributing EA and crew leader maps into the field. In the field office, a smaller-scale map compiled from EA maps can be displayed to show progress. Other planning efforts of the field office include estimation of workloads and travel costs, arrangement for distribution and receipt of materials, identification of trouble spots, and arrangement of visits by office personnel to field locations.

Production of EA maps is conceptually straightforward provided the quality of the digital database is satisfactory. This step is more of a logistical challenge since thousands of maps must be distributed together with map reading instructions and other guidelines.

Figure 4.1: Stages in quality assurance, output production and database maintenance



Draft map production and quality assurance procedures

Match boundaries and attribute files and print overview maps

In preparation of final map design and printing, the boundary data sets and geographic attribute files need to be matched if they are not already integrated in one consistent database. This step also involves checking the correctness of the match between boundary data and geographic attribute data. If both are correct, there should be at least one map feature (a point, line or polygon) for each record in the geographic attributes file. If this is not the case, there is either an

error in the map database—i.e., an EA is missing—or the geographic attributes table contains a duplicate or erroneous record. If there are two or more polygons for an attribute record, the quality assurance staff must confirm that the conventions defined for such cases are followed.

Those drafting EA maps should bear in mind the need for thorough coverage of the country's territory by EA maps. Treating the EA boundary as a 'fence' around the enumerator's assignment area helps ensure complete coverage. Errors are minimized when EA boundaries use the center-lines of streets and roads, so that enumerators canvass the housing units on one side of the street or road and leave the other side for the enumerator of the next EA. Other rules for EA delineation can be found in Chapter 3.

Once the geographic data and the attribute information are correctly matched, labels need to be added to the map, and map symbols must be chosen that will identify features on the base maps (see also chapter 5 on thematic mapping). Labeling can be done interactively, semi-automatically or automatically using a GIS package or a more specialized cartographic design software. In a very large census map production project, the labeling of features will be a time consuming and tedious task. Especially when EA map design is quite complex—for example, many digital map layers are combined to produce each EA map—the resources required for proper label placement in terms of staff time and computer resources may be very large.

Most GIS and desktop mapping systems provide functions for automated label placement. The user simply specifies the field in the GIS database's attribute table that should be used for labeling, for instance, a street name or building identifier. The system will then use some simple rules to place labels on or near each feature. The user can usually determine the size of the labels and whether labels should be drawn on top of each other if features are too close together.

However, in all but the simplest cases some manual modification of the labels will still be required.

For very large EA mapping programs, the census office might consider to purchase a specialized name placement software package. These programs have more sophisticated algorithms to ensure that the most important rules of label placement are observed:

- No or minimal overlap between labels.
- No or minimal overlap between features and labels.
- Clear assignment of labels to features (i.e., no ambiguity).
- Pleasing overall appearance, for example with regard to font type and size.

The packages base label placement on a number of heuristic rules which can be modified by the user for special purposes. The user can save the labels designed for a specific GIS data layer in a separate annotation data layer and overlay these on geographic features layers as needed.

Quality assurance

Although much consistency checking can be done interactively on computer screens, final quality assurance is best performed using printed hardcopy maps. Large format maps should therefore be produced that contain all information that will also be present on the final EA maps. These maps are produced for final quality assurance and verification and should be organized by administrative unit. If they are printed at the same scale as the final EA maps, several map sheets will be required for each district.

Quality assurance refers to a final check of the digital map database before the products are released for the census operation. Quality assurance is very similar to quality control which was discussed earlier in this chapter. It will consist of software and manual checks. Some of the checks will be performed on all products, while more complex and time consuming checks are done on a subset of products using an appropriate acceptance sampling strategy.

Quality control during the process of data conversion concentrates on the topological and positional correctness of boundaries and coordinates. It is important to ensure that there is a seamless match between boundaries that were digitized and stored separately. For instance, the boundaries between neighboring districts must be identical if district maps are stored in separate digital map files. The emphasis in quality assurance is on the suitability of the final map products to the task of enumeration. This involves verification of several aspects of database integrity described in the next paragraphs. Quality assurance is not a trivial task. It requires considerable time and resources and the census office needs to schedule and budget accordingly, but if done well it will ultimately result in a more accurate census.

Verification by census cartography staff will involve the inspection of the following acceptance criteria:

- Legibility – all annotation on the map must be clearly legible. Sometimes too many features drawn on a map make it hard to read street names or other text information. Some non-critical text labels can be omitted to improve the clarity of the map. Also, it must be clear what feature a text label refers to. In some cases, arrows may be necessary to clarify the assignment;

- The sequence of data layers drawn on a map is important, since layers on top might obscure important features on a lower geographic data layer;
- Map scale – for instance, an EA that is very large but contains a relatively small crowded area, might require an inset or a separate map to ensure that all details can be identified; and
- Source and copyright information – each map needs to list any proprietary data sources that were used to create the digital database used to produce the EA map.

Verification by local authorities and final administrative unit check

As a critical consistency check, the printed EA maps can be sent to local authorities for a verification. Local administrators—inside and outside the census administration—can confirm that all settlements and parts of larger towns and cities are included in the geographic database. Involving local authorities in this process has the advantage that maps are reviewed by persons familiar with the local area. Naming and spelling conventions may vary in countries where several languages or dialects are in use. Approval of the maps by local officials will thus reduce the risk of errors of map interpretation by locally recruited enumerators.

Another part of the verification process is the confirmation of the administrative unit boundaries included on the EA maps. These boundaries change often, with new states, provinces, and districts added on a regular basis irrespective of census-taking efforts. This often poses problems for the NSO, which needs to produce summary statistics for these units. Several options are available to handle this problem:

- Ideally administrative boundaries are frozen by government decree several months before the census. This provides stability of the reference framework for the duration of the census. The boundary structure that is current for this period is the one for which census tabulations will be produced;

- Continuous tracking of administrative boundary changes before the census is a second option. As changes occur, they are immediately committed to the digital map database. That way, the boundaries will be current at the time of enumeration. However, constant monitoring of changes and modification of boundary databases require additional resources. The NSO can consider the use of a spatial database engine to record the vintage of various administrative boundaries;

- In some countries, boundary changes are announced in advance. The census mapping agency can thus schedule work on those areas for a later stage in the census mapping process; and

- The final option is for the census mapping agency to determine a freeze date and to revise all boundaries at a later stage, possibly after the census has been taken. If modified administrative unit boundaries cut through existing EAs, the household questionnaires for these units must be reassigned to the correct units. This introduces an additional step after enumeration and may therefore delay dissemination of census results.

EA map printing

After completion of verification and quality assurance procedures for all base maps and EA delineations, census cartography staff will print the final supervisory and EA maps. This can be done at the main NSO location or in regional field offices, depending on the type of NSO

organizational structure. Supervisory maps will show several EAs and will be printed at a smaller (i.e., more generalized) cartographic scale. Defining the map layout for individual EAs is similar to the cut-out procedures in pre-digital census mapping approaches. EA maps should be simple, because they will be used by enumerators who may have limited experience with maps. On the other hand, they must contain enough information to allow easy orientation. They should contain the following information:

- The entire enumeration area defined by a clearly indicated boundary line;
- Some parts of the neighboring areas (i.e., the peripheral area) to facilitate orientation;
- Any geographic and text information contained in the census cartographic database that will facilitate orientation within the EA: streets and roads, buildings, landmarks, hydrological features, etc; and
- A consistent map legend, including the exact names and codes of the administrative and enumeration zones, a north arrow, a scale bar, and a legend explaining the symbols used for geographic features.

Figure 4.2 shows the components of a hypothetical urban EA map. All features are stored in separate map layers in the same spatial reference system or as graphics templates. The main components are the street network, buildings and EA boundaries layer. In addition, annotation, symbols, labels and building numbers are stored in separate data layers, although these could also be added dynamically. The last component is a template consisting of neatlines and a legend which is consistently used for all EAs. Figure 4.3 shows the complete EA map with all components overlaid on one map display. Depending on the scope of census mapping activities

and the complexity of the enumerated area, EA maps may contain less or more information than this sample map.

Figure 4.2: Sample components of a digital EA map

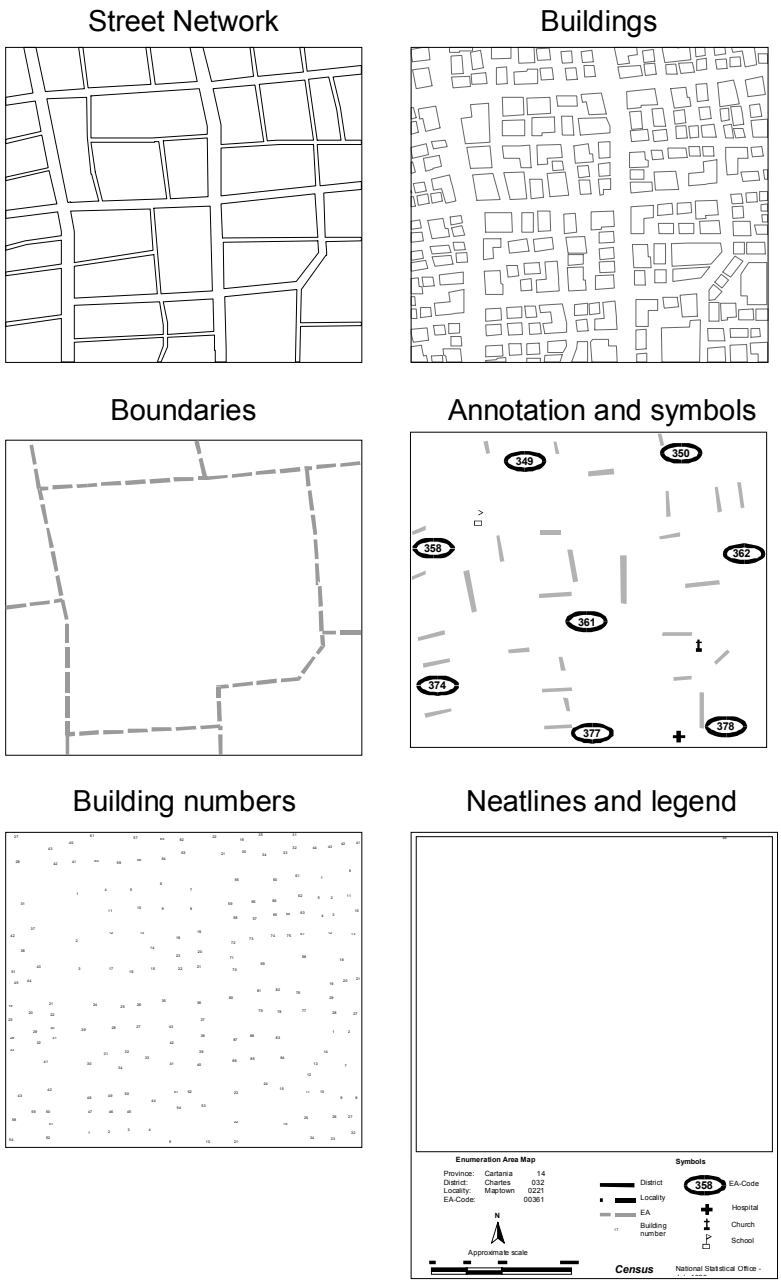
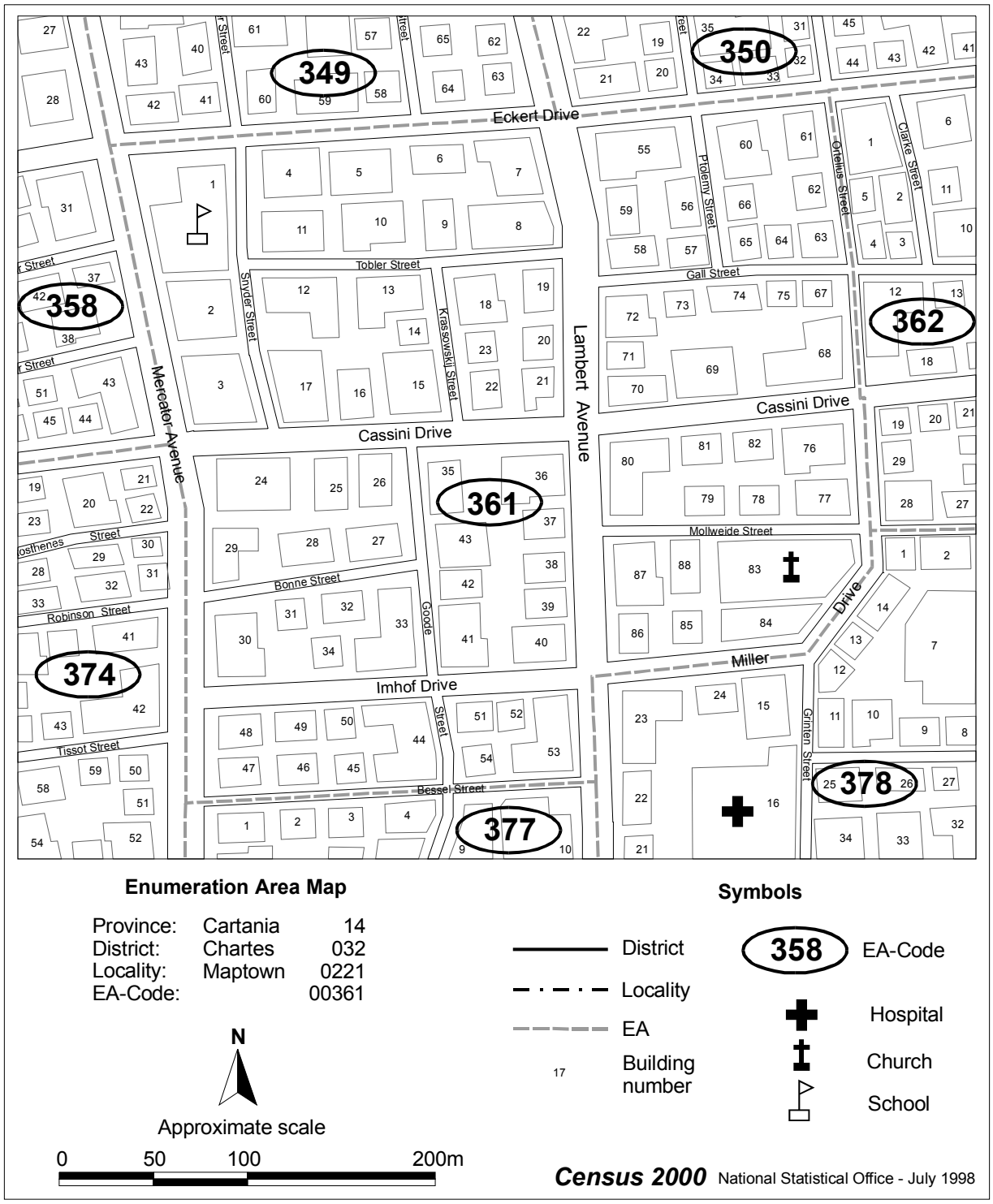


Figure 4.3: Example of an urban enumeration area map



In many countries that are not fully digital for the upcoming census round, EA map design may be simpler than in this example. For example, instead of a fully integrated digital base map in vector format, rasterized images of topographic maps may be used as a backdrop for EA boundaries. In some instances, map features may be more generalized, for instance by using only the centerlines for the streets and polygons for entire city blocks rather than for individual houses.

Decisions must be made concerning format and color of the printed EA maps. Given the high resolution available on laser printers, EA maps can usually be produced on A4 (210 x 297 mm) or letter sized (8.5 x 11 inches) paper if the EAs are not too large and complex. Compared to larger format printers or plotters, standard sized printers have the advantages of lower cost and higher output speed, and when calculated on a per-page basis including ink or toner costs, laser printers are far less expensive than ink-jet printers. Since thousands of EA maps need to be produced, these are important considerations. Problems may occur in areas where a very large EA contains some small crowded areas. For these areas, larger format maps must be printed, or the map design must include insets to show detail in the dense parts of the EA.

For keeping track of a paper map inventory before the start of the enumeration, NSOs should consider barcoding its maps. Barcodes are an easily implemented technology using a laser reader and simple database software. Barcodes contain no descriptive data, but simple a random reference number that the computer uses to look up associated records. The barcode reader senses light and dark from reflected light, which it converts to an electronic signal (high for black and low for white). Many barcode readers now

come with a USB interface with a decoder either integrated into the unit or separate. The NSO can use preprinted barcode labels or write a piece of software that creates a barcode for each map produced. For more information, see www.barcodehq.com/primer

A well-designed EA map will usually work well in black and white. Although color printers have become relatively cheap, their throughput is limited and supplies are often quite expensive. Good black and white maps can also be photocopied without loss of information which allows the local staff to produce additional copies of EA maps as required. Where resources are available, however, color can contribute to the clarity of map design. For instance, the EA boundary can be indicated by a brightly colored line on the map. The same effect can be achieved on a black and white map by going over EA boundary lines with a bright-colored highlighter.

If possible, NSO officials should consider usability testing for readable maps. Design decisions such as map size, color, scale, and map elements can then be made on the basis of actual users' experiences rather than those of an in-house design team alone.

Several copies of each EA map must be produced in addition to backup copies that are kept in the central census mapping office. Each EA map will be made available to the local census authorities, supervisors, and enumerators requiring perhaps 4-5 copies. If mapping activities are centralized in one or a few census offices, a preferred approach is to distribute digital map files rather than hardcopy maps. These files can be transmitted to local census offices on CD/ROM or DVD or via the Internet. The local office will not need to have access to mapping software if the maps are exported to a generic file format such as the portable document format (PDF) or as a graphics file embedded in a generic

word processing format. Such files can be printed on any generic computer system. This approach enables the local office to produce as many copies of EA maps as required and allows for quick response to problems such as lost hardcopy maps.

If the database is consistent and well-organized, EA map production should be fast, and there is usually no need to shop the work out to a contractor. Printing of EA maps will not require a high-end GIS package, but can be done using relatively inexpensive desktop mapping packages. Some of the process can be automated using the built-in macro language of the software. For instance, a list of EAs can be accompanied by the bounding coordinates of the EA (the so-called map extent) in map units. The software can then be instructed to go through this list, include the content of the data layers into a pre-prepared template showing the legend and other marginal information, and print a specified number of copies.

Use of GIS during census enumeration

The main contribution of digital mapping to a successful census is prior to and after the actual enumeration. Mapping, however, also has a role during census enumeration by supporting logistical planning and monitoring census progress. At the same time, the enumeration process gives the census office a chance to perform another round of quality control of the digital census database. Both of these aspects will now be discussed.

Use of digital maps for census logistics

Maps are needed for many purposes in the census process. Among these, GIS can also play an active role in planning preliminary work and the logistics of the enumeration.

Assignment of administrative units to operational areas, location of field offices and planning the travel of field workers and enumerators are some of the tasks where GIS is potentially useful. If digital maps are to be used for these purposes, the census cartographic unit should develop a coarse resolution GIS database at the beginning of the process. This system could consist of small-scale digital maps (at 1: 500,000 or 1: 1,000,000) on settlements, roads, rivers and administrative divisions. These datasets can be obtained from existing sources in most cases. The NSO can further develop its own datasets, or make use of existing resources from the country's national mapping agency or using 'canned' data that came with the GIS software.

Many GIS packages offer network analysis features that allow the planning staff to determine the distances and cost of travel along a road network. The quality of road data varies dramatically by location, and some caution should be used in cases of small-scale (i.e., highly generalized) data. In urbanized areas, planning travel to enumeration sites will not be a major problem. But in rural areas, large distances and natural features that make travel difficult will increase the cost of field based activities. This will also be a factor in determining the location of field offices, which are responsible for a number of supervisory or crew leader areas. Field office locations should be chosen to minimize travel time and thus to facilitate the supervisory functions of the regional census administrators. The area aggregation features of a GIS can be used to determine and display possible regional assignments.

The use of GIS for logistical purposes is not quite as critical as the use of digital techniques to carry out the actual census cartography tasks. Many of the tasks can be done equally well by studying published maps, although doing so reduces the utility of

the geographic database. The advantage of using a GIS for these purposes is that distance and travel time estimates will be more accurate and that the census staff can quickly produce maps showing various aspects of the census planning process. Furthermore, the development of a small, coarse-resolution GIS database for the country is a good precursor for the much more challenging task of producing a detailed georeferenced census database.

Monitoring progress of census operations

During the census and the activities immediately following the enumeration, headquarter staff will monitor the progress of enumeration and data compilation. Typically, regional census offices will compile information about completion of enumeration activities and first results. The headquarter will collect this information and assess where operations are running smoothly and where problems may be encountered.

Some countries implement a 'quick-count' strategy in which total population figures are rapidly compiled and compared with prior estimates. Areas in which the reported figures are unusually high or low may need immediate attention. Traditionally these assessments are compiled in tabular form. If a detailed digital census map database exists, however, this information can also be displayed geographically. This makes it easier to spot problem areas.

In practice, any suitable summary statistics can be compiled in a standard relational database system. Examples are an indicator that shows whether or not the enumeration has been completed in the reporting area or the percentage of enumeration areas in each district that has been completed. Census staff can then regularly link this information to

the GIS database and prepare map output for evaluation by the overall census supervisors.

The key to this rapid quality control procedure is the fast flow of information from the supervisors to regional offices and on to the headquarters. The quickest way of exchanging this information is the Internet. If local and regional supervisors have access to the Internet, information can even be submitted through a password protected database interface on the Web.

Guidelines for map use by enumerators during the time of the census

Canvassing techniques during enumeration begin with the basics. Enumerators should have located his or her EA both on the ground and on the map. Secondly, the enumerator should make sure to use his or her map, even if it seems obvious where they are. Orienting the map means laying the map flat so all parts of it are visible. The enumerator should begin orientation at a major street intersection, finding the location both on the map and on the ground, identifying features – houses, houses of worship, railroads, rivers – and checking them against symbols on the map.

Enumerators with the help of crew leaders should plan the route of travel around the EA to minimize backtracking. An EA can be divided into sections, blocks, or sectors. The enumerator should mark an X at the starting point of the day's canvassing, and mark the map with the date at the stopping point at the end of the day. Enumerators should follow a "keep to the right" rule, proceeding clockwise from the starting point so the houses being enumerated are on the enumerator's right, and orienting the map so it points in the direction of travel. As each corner is turned, the map should be re-oriented so the features

line up with those on the ground. Enumerators should identify housing units on the map for purpose of return visits if necessary. Obstacles in the terrain such as landforms, water, and poor roads should be noted on the map so that future enumerators can make use of this information.

Updating and correction of EA maps during enumeration

Even if a thorough quality control program has been carried out during the preparation of enumerator maps, it is likely that many of the maps contain errors, sometimes significant ones. For example, during initial field work, buildings or streets might have been overlooked or registered incorrectly on the maps. Since cartographic field work needs to be conducted several months or even years ahead of census taking, new construction and infrastructure developments may not be considered in the enumerator maps.

In addition to training in data collection and basic map reading skills, the census office should also instruct the enumerators to annotate the EA maps during enumeration to point out any errors or omissions. The census cartographic staff should collect the EA maps after the census and follow-up on any suggested revisions through an established procedure of incorporating edits into the master geographic census database (this topic is addressed in the following chapter). This may simply require making the corresponding corrections in the digital census database, or it may require some additional field checking. This process will ensure that the census office holds the most current information on the enumeration areas, which will reduce the workload for cartographic activities before future censuses or surveys.

Box: Field map production for a 2000 round country: **BRAZIL**

Summary and conclusions

Chapter 4 has covered the process of creating the maps needed for enumeration, with an operational focus that picks up where the geographic database discussion left off. Included have been sections on process flow for enumerator maps, compilation, determining relevant layers for enumerators and supervisors, printing and distribution basics, a project management approach to enumeration, and a plan for contingencies in the event of delays.

Chapter 5: Geographic databases for postcensal dissemination of results and analysis

At this point in the process, the results are in and the geographic database used for the census is being updated with corrections from the field in preparation for dissemination of results and analysis. This chapter will cover the process of aggregating data to dissemination units, some decisions about products and formats, as well as privacy and disclosure issues, spatial analysis techniques and sample products, including those using the internet.

All through this handbook we have stressed that the key to an NSO's investment in a geographic database is its ability to be used for all phases of the census process. Now that the enumeration is complete, results can go back and improve that database further. Using the geographic database to create products to inform and educate the public will raise issues about aggregation and scale that will require input from those outside the NSO.

The previous chapter has discussed the use of GIS to support census enumeration. The following sections will deal with tasks that the geographic unit of the census office needs to carry out after the census and between censuses, and with the dissemination and use of geographically referenced census information.

The handbook has argued that all geographic plans must align with an overall plan for the census. If a complete digital census GIS database has been created, then statistical GIS databases for administrative or statistical units can be produced simply through aggregation. For the countries that do not use digital techniques for the production of EA maps for the 2010 round, options still exist to develop a digital georeferenced census database for producing publication-

quality maps to accompany census reports, for distribution to outside users who want to analyze census data spatially, or for internal applications. This database can be compiled for a suitable level of the administrative hierarchy or for other aggregate statistical regions. At that level of aggregation, the resources required for producing a digital database are much lower than those necessary for a complete digital EA map database.

For the most part, however, this chapter assumes that a complete digital enumeration area or dwelling unit database has been created for the purposes of census enumeration. To justify the investment necessary for developing such a database, the census office needs to adopt a long term perspective. Immediate tasks after census taking are thus only the first steps in the preparation of cartographic materials for the next enumeration.

The main sections in this chapter discuss management tasks relating to geographic databases after and between censuses and the development and dissemination of output products. The final sections discuss some advanced topics such as urban and rural area delineation and techniques for dealing with incompatible geographic units, ending with some methods for analyzing census data spatially.

Tasks after the census and during the intercensal period

Immediate tasks

Incorporate updates and changes identified by enumerators

Part of the enumeration process was the training of census-takers to notice errors or inconsistencies in the delineation of EAs or base map features the canvassed territory. The census mapping office should encourage enumerators to record errors, which the local

supervisors should then collect after enumeration and forward to the census mapping office. The census geographic unit can then, through a process devised by the NSO, correct the map database that was used for EA map production based on this information. This procedure will have two benefits.

Firstly, it ensures that tabulations and the development of digital and hardcopy map products are based on the EA delineation actually used during enumeration. Secondly, committing the modification of EA boundaries into the master geographic database will facilitate future census or other statistical data collection activities that are based on the same or similar geographic collection units. This has the obvious benefit of keeping the geodatabase as current as possible while also supporting the active role of the field participants in improving data quality overall.

Reconciliation of collection units and tabulation or statistical units

The most important post-enumeration responsibility of the census mapping agency is to support the development of tabular statistical data produced from census returns. Census data are required for many different types of aggregated areas since census users from different sectors tend to use different geographic areas as the basis for planning and operations. EAs therefore need to be aggregated to these various reporting units as required for the development of a wide range of census output products. It is unlikely that all users will require the granularity, not to mention the challenging file management, of an EA-level database that spans the country.

Matching of data collection (EA) and tabulation units at various levels of geography requires the development of *equivalency* or *comparability files*. These files list for each tabulation unit the corresponding EAs that are part of that output unit. Once such lists have been defined, aggregation can be done using standard database operations. The changes in enumeration areas

or other units from census to census can be seen in Table 5.1, which shows that one EA (census tract) from 1990 was split into three new EAs (census tracts) for 2000, accompanied by street-side mileage changes and revised coding.

The development of equivalency files is made easier if a consistent coding scheme has been implemented. Geocoding is covered extensively in the handbook, particularly in Chapter 2. Coding reaffirms the importance of developing intuitive and flexible conventions for assigning numeric or alphanumeric codes to each unique EA in the early stages of a census mapping project.

Field Length Field Description:
1-11 census tract code,
12 part (P) flag
13-18 street-side mileage of 1990 census tract
19-22: percentage of 1990 census tract street-side mileage
23-31 2000 census tract code
32-33 2000 census tract suffix
34 2000 census tract part (P) flag
35-40 Street-side mileage of 2000 census tract
41-44 Percentage of 2000 census tract street-side mileage
45-50 Street-side mileage of the area covered by the record
51-64 14 Land area of the record (1000 sq. meters)
65-66 2 2000 state name abbreviation
67-126 60 2000 county name

The record below shows that 1990 census tract 402 split into three 2000 census tracts: 402.01, 402.02, and 402.03

```
10001040200P    11735029610001040201    34771000    3477     9796DEKent
10001040200P    11735044210001040202    51891000    5189    32059DEKent
10001040200P    11735026210001040203    30691000    3069    59822DEKent
```

Table 5.1 Table comparing old and new EA units

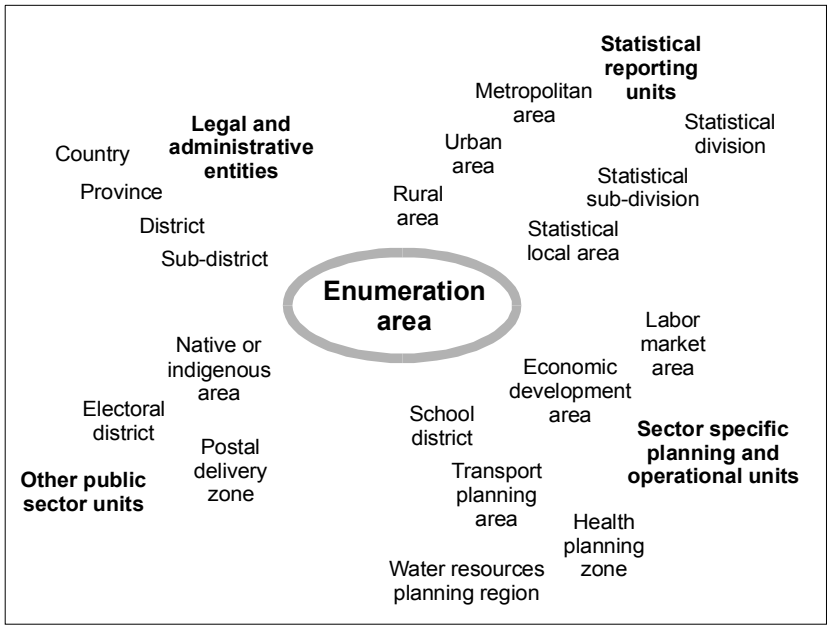
The number of output units for which equivalency files need to be developed can be very large. In addition to legal and administrative units such as districts or provinces, census data may need to be compiled for a range of planning or operational units. Some examples are health units, school districts, transportation planning regions, electoral districts, utility service zones, postal

zones, and environmental planning units (see Figure 5.1). These may in some instances coincide with administrative areas, but often they will be incompatible with standard reporting units. In addition, special tabulation requests are likely to come from the private and academic sectors. Developing a consistent procedure for production and maintenance of equivalency files is thus an important task of the census mapping office. As we have stated throughout the handbook, planning effectively will have numerous seen and unseen benefits for the NSO.

Additional comparability files should be developed to reconcile past with present enumeration or statistical reporting areas. Since both data collection and tabulation units tend to be modified regularly, it is difficult for census data users to determine changes in census variables over time. The geographic unit of the census office should therefore keep track of those modifications in the country's census geography and provide data users with comparability files that allow the harmonization of past and present census data.

This effort will be particularly helpful when the NSO participates in spatial data infrastructure (SDI) activity within the country, since it will provide those requiring census data with the exact specifications needed to locate the correct census files for their particular purposes.

Figure 5.1: Examples of census tabulation and reporting units



Database maintenance

Database archiving

After errors and inconsistencies have been addressed in the master geographic database, benchmark copies of all GIS data sets should be produced and archived. This database, for which the census geography has been frozen to reflect the situation at census time, will be the basis for all cartographic outputs including reference maps, thematic maps of census results and digital extractions from this master database for distribution. All census results that are tabulated following enumeration will refer to the reference units in this database. This also implies that all documentation and meta-data are thoroughly checked, so that the census office can answer any questions concerning the data that may arise in the future. Copies of this reference database should be backed up and archived in a secure place immediately following completion of database work.

For census agencies that have a continuous mapping program, a copy of this database will serve as the basis for regular updating during intercensal activities. The advantages of a continuous mapping program are discussed in the following section.

Database maintenance: advantages of a continuous mapping program

This handbook has argued that the benefits of a digital cartographic census program will outweigh the considerable costs, but only if the resulting census database is used for many applications beyond the core tasks of a census. The full range of benefits can only be realized if the database is maintained so that updates for future census applications will require relatively minor resources. Deploying the census cartographic database for the largest number of uses and

ensuring that maximum use is made of existing digital data in subsequent enumerations is possible only if there is a high degree of continuity in the national census mapping program. Continuity of census geographic activities will therefore ensure that the investment in database development is preserved.

One aspect of this is that the census mapping office should implement database maintenance procedures immediately following a census. This involves a continuous updating of boundaries and other features as new information becomes available. During the intercensal period, a clear system of version control should be implemented that specifies how changes to the database are implemented and documented. For instance, only one or a small group of staff members should have the authority for committing changes to the master database. Granting such special administrative rights to only some staff members avoids the case where different staff members make changes to different versions of the database which later have to be reconciled.

During the intercensal period, the census mapping agency should follow industry trends as well as new approaches adopted by other national census mapping agencies. This will inform decisions about investments in software and hardware upgrades. Given how fast technology changes, periodic investments in these areas may be required to ensure a high quality of census operations in the intercensal period.

Digital cartographic data development requires special expertise in computer use, geographic concepts and specialized software packages. It is expensive to train personnel in all but the most basic GIS concepts and tasks. For a long-term census mapping program to be successful, staff continuity is therefore a critical factor. The census office needs to identify a core staff that will maintain the database in the intercensal period, provide GIS services for other statistical

applications such as sample surveys, and serve as an institutional memory. This will facilitate a smooth operation of census GIS applications in the next enumeration. Core staff can, for instance, carry out the training of temporary staff recruited for digitizing or field work. Retaining core staff will also reduce the start-up costs otherwise required for recruiting GIS experts who would then need some time to be fully integrated in the census cartographic process. Note that human resource issues were covered in Chapter 1.

These paragraphs have again emphasized the importance of a long term view of census cartographic activities. The additional resources required to maintain a cartographic capability in the census office between censuses will be well worth the benefits from pursuing a long term strategy.

Dissemination of geographic census products

Planning data dissemination

The definition of cartographic output products and the scheduling of their release needs to be closely coordinated with the time table for the overall census project. Tabulation of census data may require cartographic information from the census geography unit, and, vice versa, thematic maps and digital geographic databases can only be completed once census data processing has been completed.

The selection of suitable output products should be guided by a detailed assessment of customer requirements—i.e., market research—that should be carried out in the early stages of census planning. These plans for dissemination products should be made very early, and published widely in order to get feedback from the user community.

It is useful to establish an advisory panel of representatives from the most important census data user communities that can guide the census community. The advisory group functions do not need to be limited to the census planning stage, but could be a permanent formal or informal mechanism for exchanging ideas between the census office and data users. The examples of the use of small-area census statistics provided in the introduction to this handbook provide some indication of the wide range of data users that the census office should consider in their user needs assessment.

Past experience of what has proved popular with census data users can only be a limited guide to the definition of output products. Demands change, partly in response to changing technological capabilities among data users. Digital map database products were rarely available after the last round of censuses, while they will be one of the most important outputs of the current census round. While demand for hardcopy maps may be larger in many countries than requests for digital information, this is changing as electronic distribution of map products becomes increasingly ubiquitous. The census mapping agency needs to be flexible to respond to changing customer needs and special requests.

It is advisable to look ahead several years when planning the output strategy. For example, the Internet is becoming a major data distribution channel even in poor countries as communications infrastructures improve worldwide. New user communities emerge as new data products are created. For instance, mobile telephone companies seek georeferenced demographic data to plan the installation of new towers and other infrastructure. To increase the societal benefits from census data collection, the census office can actively search for potential new customer groups and introduce their products to them. One effective way to do this across government and among

nongovernmental data users is through participation in a national spatial data infrastructure (NSDI).

The census office should estimate the volume of possible demand for their products and services which will allow some assessment of required capacity for servicing customer requests. Again this is difficult since demand may increase as new products are introduced and as new users see census products and realize their potential for their own needs. Thus the census office needs to be prepared to serve a growing demand once products are made available. Effective plans are scalable with the level of demand for census products. It is advisable to define clearly and early which census data users' needs *must* (i.e., are legally required to) be served, whose *should* (through notions of good customer service) be served, and whose *will not* (because of resource constraints) be served. A clear set of priorities will also facilitate the development of a time table for census product distribution.

An open data dissemination policy—i.e., low-cost or free access to data—can help reduce the workload of the census office. In countries where census data are freely available, private sector service providers may be able to cater to the special needs of some census data users. This allows the census office to concentrate on data users that they are mandated through legal statute or legislation to serve.

Some census geographic data products will be required for internal and official use. These may include equivalency files and reference map libraries, but also special purpose products such as electoral district maps. In some countries, the census office may be required by law to produce certain map products. These products may have to be generated on a regular basis or upon special request for example from government ministries or parliament.

Other, more generic products will be designed for wider dissemination to government and private sector users and to the general public. The census office should attempt to exploit as many distribution channels as possible and meet multiple users' needs through publication not only of data and results but also analysis. Some examples of such spatial analysis are found later in this chapter.

The following sections will discuss census output products and dissemination options including required products, thematic maps which can be distributed in hardcopy or digital format, digital cartographic database dissemination, digital census atlases and Internet mapping. A thorough background in cartographic techniques for thematic mapping is required for many of these output products. Only the more general issues concerning thematic mapping are discussed in this chapter. Annex 5 provides a more comprehensive overview of thematic map design.

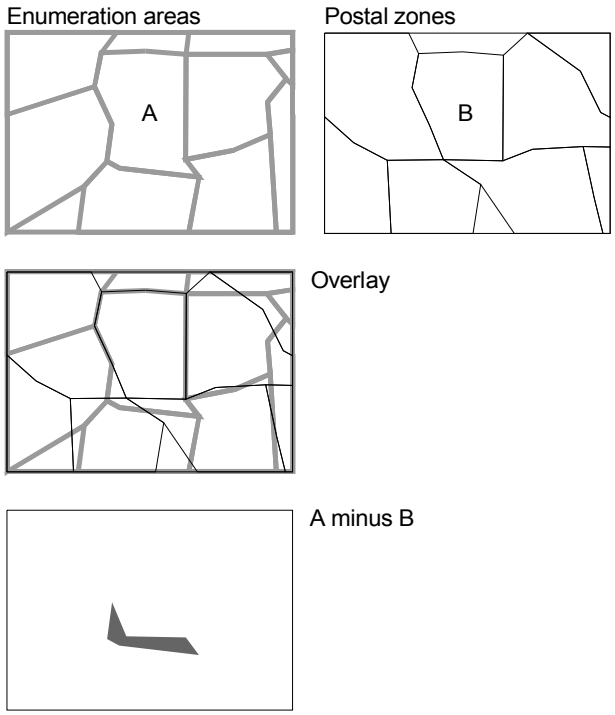
Disclosure and data privacy considerations: The differencing problem

Various government agencies and outside data users may require census data for different sets of small geographic units. For example, some organizations use small postal zones or health areas as their primary reporting units. To satisfy the needs of these data users, the national census office may want to distribute census information for several sets of small geographic areas whose boundaries are independent from each other. If boundaries and data tables are published for two or more sets of areas, a user may be able to use GIS operations and simple data table manipulation to derive census statistics for very small geographic areas. The census counts for these new units might fall below the agency's disclosure threshold. This problem is called the differencing problem in statistical disclosure (see Duke-Williams and Rees 1998).

This problem does not occur if boundaries overlap irregularly, unless one of the overlapping zones has zero values. In most cases, a user cannot be certain that a zero value is actually correct. This is because most census offices use perturbation or broad-coding (giving a data range such as "<10" rather than the exact small value) to prevent users from being able to derive exact characteristics for small groups of individuals in areas with low population. (check with expert group about this)

The differencing problem can occur, however, if a zone from one set of geographic areas nests into a zone from another set, and the user has data tables for both sets of areas. For example, postal zone B in Figure 5. nests into enumeration area A. By overlaying the two sets of boundaries we can determine the geographic area that is in A but not in B. Using the data tables we can now derive census data for the individuals in this small area by simply subtracting the counts for postal zone B from those for enumeration area A. These counts may well fall below the disclosure thresholds even if the counts for A and B do not.

Figure 5.2: The differencing problem in statistical disclosure



To avoid data disclosure problems, the census organization should carefully review the boundaries of alternative census geographies. In instances where differencing appears possible, additional data protection must be introduced. Duke-Williams and Rees (1998) analyze the differencing problem in great detail. Based on their experiments, they give some general recommendations that address the problem:

- Use minimum threshold levels for tables. Some further protection can be given by introducing mild perturbations of data values for very small areas or to use ranges rather than exact values for small counts. This will reduce the risk of publishing census data for more than one set of small area units.

- The primary census geography chosen for distribution should be as generally useful as possible. For example, if most agencies in the country use small administrative areas as their primary reference, census data should be published for those units.

- The risk of publishing alternative geographies whose zones are much larger than those of the primary census units is very small. Even if differencing is possible in these cases, the resulting counts are unlikely to fall below the safety threshold.

- If two census geographies of approximately equal resolution are very similar—that is if many of the boundaries are the same—the risk that differencing is possible will be larger than if the boundaries are very different.

- Differencing problems can be reduced if the NSO settles on standard administrative or statistical boundaries with known coordinates and level of boundary line generalization.

Marketing of digital map products

Given the large and growing demand for small area census data, national statistical organizations should consider ways to streamline user access to results. Strategies can be pursued to develop in-house capacity to meet user needs or leverage outside expertise. Wherever applicable, an NSO should develop capacity in house to distribute its own data and analysis, relying on outside entities such as commercial firms when potential demand exceeds the ability to disseminate.

Countries that aim at recovering some of the costs of developing census GIS databases and in which there is a strong commercial demand for small area statistical data may want to explore the possibility of entering into a marketing agreement with a private data vendor. Potential collaborators include the local distributors of the major GIS software producers, online mapping

and imagery providers, and vendors of geodemographic data. Most of the leading GIS vendors produce and sell GIS data sets on many topics. This is partly an additional revenue source, and partly a way to facilitate the use of their software products by providing data sets in the software's data format. These private vendors sometimes collaborate with national mapping and statistical institutes to produce professionally designed GIS databases.

For the national statistical office, this has some advantages. The software and data vendors can contribute technical know-how and possibly computing resources to the development of the GIS database distribution package in return for a share of the proceeds of database sales. Internationally operating software vendors can also increase the distribution of national GIS data. Demand in other countries may come from internationally operating companies or academics studying the country.

One possible problem in collaborating with a commercial software vendor is that the vendor may want to distribute data only in its own proprietary format. The census office should make sure that data users who want to use another format will be able to access the data also. The disadvantages of commercial distribution have been mentioned before. By assigning marketing rights to a private company, the statistical office cannot distribute data free-of-charge or at very low cost. If the goal is to attain widest possible distribution, in-house development and distribution of databases is preferable.

Other potential marketing partners are universities or other government departments that disseminate information. In all cases, a clear marketing and revenue sharing agreement must be put in place to avoid problems later. The census office should make a detailed evaluation about the market value of their data in relation to the costs of producing, advertising and selling the

data, to ensure a fair and mutually beneficial agreement will be the basis of a public-private or public-public partnership.

Outreach

To ensure broad awareness of data availability and widest possible distribution of georeferenced census data, the national statistical office should develop an outreach plan. Part of that plan could be printed brochures and posters featuring census maps. These can be widely distributed to schools, universities, commercial enterprises and national and local government offices. A marketing campaign can precede the census to raise awareness of the need for up-to-date statistical data and ensure full public participation.

The census office can also organize a series of regional user seminars across the country. In these workshops census staff can introduce the use of free or low-cost mapping packages for the analysis of census data to a wide range of potential users, while also informing the public about the range of products forthcoming from the census effort.

Required products

Equivalency and comparability files

Equivalency files have been discussed previously as one of the first responsibilities of the census mapping office after the enumeration. In addition to their immediate use for census data tabulation, equivalency files are also an output product. Data users may require information about which EAs belong to a given statistical or administrative output region, or which small area statistical units make up a more aggregate reporting unit.

Equivalency files should be made available in both hardcopy and digital format. Most users who work with digital census data—whether geographically referenced or tabular—will benefit from having these files available in computer readable format. This allows the direct use of these files in database operations.

Reference map library

In addition to equivalency files, the census office should also produce reference maps of all reporting units. In some countries, the census mapping office is legally required to produce such maps for use by government officials and the general public.

Reference maps can be disseminated in digital form as simple graphics, postscript or PDF files. However, not all users will be able to use digital files. Complete sets of hardcopy reference maps should therefore also be made available on demand.

Reference maps need to be accompanied by a detailed description of the definitions of each census geographic area. A good example of a comprehensive reference map documentation is the Geographic Areas Reference Manual produced by the U.S. Census Bureau which is available on the Internet: www.census.gov/geo/www/garm.html

Gazetteers and centroid files

It is usually the responsibility of the national mapping agency to produce a gazetteer, a list of place names and their geographic location. However, a large scale national mapping program implemented for census purposes may provide an improved or updated information base for a national gazetteer as well. In some countries where no other source for such data is available, a gazetteer may be one of the required products from a census mapping project. If the census

mapping project has made extensive use of GPS data collection, development of a gazetteer that lists all geographic places should be straightforward. A gazetteer can include the names of all populated places along with variations in names, latitude-longitudes, administrative codes, and populations.

A gazetteer should be stored and distributed in digital form, allowing direct use of coordinates and name information into a GIS. It will also be useful to develop a simple query system, where users can request coordinates of a specific place such as a village in a given province. Such data can be made available via the Internet using a straightforward front-end to access the database.

Thematic maps for publication

The power of maps

Before discussing the types of thematic maps that can be produced for census publications it is useful to review why thematic maps are powerfully effective for the presentation of census results:

- Maps communicate a concept or an idea.
- Maps are often meant to support textual information. Themes and issues that are difficult to explain in words can be illustrated more effectively in map or graphic form.
- Maps appeal to the viewer's curiosity. They provide eye catching anchors on the pages of a report. These will get the readers attention and encourage reading the accompanying text.
- Maps summarize large amounts of information concisely. It would be hard to match a map's ability to represent not only huge quantities of numbers but also information about the spatial

relationship between observations. A map of population densities at the county level in China or the United States, for example, will show more than 3,000 data values. This map can be printed on a letter-size page without major loss of clarity. It would be difficult to fit 3,000 numbers on a letter-size page and this would still provide less information, for instance about where low and high values are clustered in the country.

- Maps can be used for description, exploration, confirmation, tabulation and even decoration. Maps can serve many purposes. Presentation maps in census reports are usually descriptive in nature. They simply present census results with or without analysis. A demographer or geographer using census data, in contrast, might use maps to explore relationships between different variables, say life expectancy and literacy rates. In a final report, maps of these factors might be used in addition to text and charts to support the analyst's results. The map thus becomes a tool for confirmation of results which may or may not be obtained by looking at the map alone. Maps might also be used simply for inventory purposes, for example to show all the schools or health clinics in a country. Of course, inventory quickly leads to analysis, for example, by pointing out areas that are not served sufficiently by public facilities. Finally, maps are popular because they are often eye-catching and attractive. Witness the large number of maps hanging on office walls. Few people hang up statistical charts or tables of numbers.

- Maps encourage comparisons. Whether descriptive or exploratory, the main purpose of thematic maps is to compare things across geographic space. Many types of comparisons are possible:

  - Between different areas on the same map: where are population densities highest?

- Between different maps: is child mortality higher in the districts of province *A* than in province *B*?

- Between different variables for the same area: where and by how much do literacy rates for males and females differ in the districts?

- Between maps for different time periods: did fertility rates decline since the last census?

Thematic mapping of census data

GIS encourages a view of maps that is quite different from traditional cartography. In a computer, maps can be generated quickly on a computer screen. This supports a mode of work that is optimized for data validation, exploration of data patterns and data analysis, an area increasingly known as "*geovisualization"*. Geovisualization grows out of established principles of map production and display. It is defined as the creation and use of visual representations to facilitate thinking, understanding, and knowledge construction about human and physical environments, at geographic scales of measurement (Longley et al, 2005).

Maps created on a computer screen are sometimes called "virtual maps" to distinguish them from printed or drafted hardcopy maps, although increasingly the line is blurring between computer and hardcopy maps. In the census process, relatively little concern needs to be given to traditional cartographic map design in the early stages of a digital census mapping project. The emphasis—as shown in chapter 2—is on database development and verification. Even the production of EA maps, which show the main features of an enumerator's work area, usually employs relatively simple cartographic design.

Once census data have been compiled, however, the census office will usually want to produce publication quality maps that illustrate census results and accompany published census reports. Such maps will be presented to a wider, non-specialist audience. They will therefore have to be designed much more carefully, whether the final product is printed in book form, published on a CD-ROM, or posted on an Internet web site.

Table 3.1 shows a list of possible thematic maps that can be included in a census atlas or a census office's Internet web site (see United Nations 2006). Many other types of maps might be considered for publications on special topics or to highlight interesting aspects of census results in regions of the country. Just as tabulations of census data can be disaggregated by gender, age group or urban/rural areas, census maps can also be divided into population components. Maps that show comparisons over time, if comparable indicators are available from previous censuses, are also informative.

Table 3.1: List of thematic maps for a census atlas

(Source: *Principles and Recommendations for Population and Housing Censuses,* United Nations 2006)

Population dynamics and distribution
Percentage population change during intercensal period(s)
Average annual growth rate
Population density (persons per square kilometer)
Urban population as percentage of total population
Distribution and size of major cities and towns
In-migration, out-migration and net migration rates
Born in country and foreign born
Born in another division of the country

Demographic characteristics

Sex ratio (males per 100 females), possibly by age groups

Percent of population age 0-14

Percent of population age 15-64

Percent of population age 65 and over

Percent female population in child bearing age 15 - 49

Total dependency ratio (percent of population age 0-14 and 65 and over to population age 15-64)

Marital status

Birth rate

Total fertility rate

Mean age at first marriage

Death rate

Infant mortality rate

Life expectancy at birth

Percentage of people with disabilities

Socio-economic characteristics

Percentage of children not in primary school

Adult literacy rate (age 15 and over)

Mean years of schooling (age 25 and over)

Illiteracy rate of population age 15 and over

Illiterate population age 15 and over

Educational level of population age 10 and over

Labour force as percentage of total population

Women's share of adult labour force

Percentage of labour force by economic sector, type of occupation and status in employment

Household and housing

Average number of persons per household

Percentage of households headed by women

Average number of dwelling rooms per household
Tenure status (owned, rented, and so forth)
Type of construction material
Percentage of population with access to adequate shelter
Percentage of population with access to safe water
Percentage of population with access to electricity
Percentage of population with access to sanitation
Percentage of population with access to health services

Publication-quality census maps will usually be produced only for fairly aggregate statistical reporting units. A census agency can produce national overview maps showing the distribution of indicators by province or district, as well as more detailed maps for each province. For major urban areas, very detailed maps can be produced using census block or enumeration area level data.

GIS and desktop mapping packages provide a wide range of cartographic functions, and many commercial map makers have switched to fully digital production techniques. Achieving high quality cartographic output still requires considerable experience and know-how. Tools provided by computer-based mapping systems do not substitute for cartographic training. The availability of easy-to-use mapping packages has led to a proliferation of maps that violate many of the standard cartographic design principles. Initially this was due to the lack of proper cartographic functions in early GIS packages. With improvements in the power and versatility of GIS programs, those with little or no training in cartographic techniques may still have difficulty producing decent maps.

In most census agencies, professional cartographers will be in charge of producing maps for publication and distribution. These staff members can produce high quality maps on the computer after receiving some training in digital mapping techniques.

Due to the widespread diffusion of GIS and desktop mapping software, thematic maps are increasingly produced by subject specialists with little or no training in cartographic design principles. Annex 5 therefore provides a summary of thematic mapping techniques. The information contained in that annex should be of interest to core cartographic staff as well as to people inside and outside the census agency who may produce maps from digital spatial databases only occasionally. Excellent additional references on cartography and thematic mapping are Robinson et al (1995), Kraak and Ormeling (1997) and Dent (1998). MacEachren (1994) produced a very useful primer on thematic mapping specifically targeted at GIS users with little formal training in cartography. Design guides designed specifically for GIS maps are becoming increasingly common (for example, see Krygier and Wood 2005, and Brewer 2005).

Spatial Analysis Techniques

A variety of techniques offer functionality beyond thematic (choropleth) mapping, with many tools now available in both COTS and FOSS software programs. The presupposition behind use of new methods of spatial analysis is the availability of population data with a higher level of *granularity* than previously, at the EA, population cluster, or other small-area levels. If analysts or other GIS users wish to analyze the spatial distribution of population, or map demographic or other variables in relation to others, they can now make use of a variety of techniques.

Spatial analysis is defined by Longley et al. as a set of methods whose results change when the locations of the objects being analyzed change. Spatial analysis is sometimes called “the crux of

GIS" because it includes methods to turn data into information. It can be used to answer such questions as, what is the spatial relationship between *X* and *Y*, or where can similar characteristics be found? Or, can a model be found to describe a general pattern, and where are the anomalies? Techniques range from simple queries to measurements to transformations, descriptive summaries and models. Each of these will be described in turn, with examples provided.

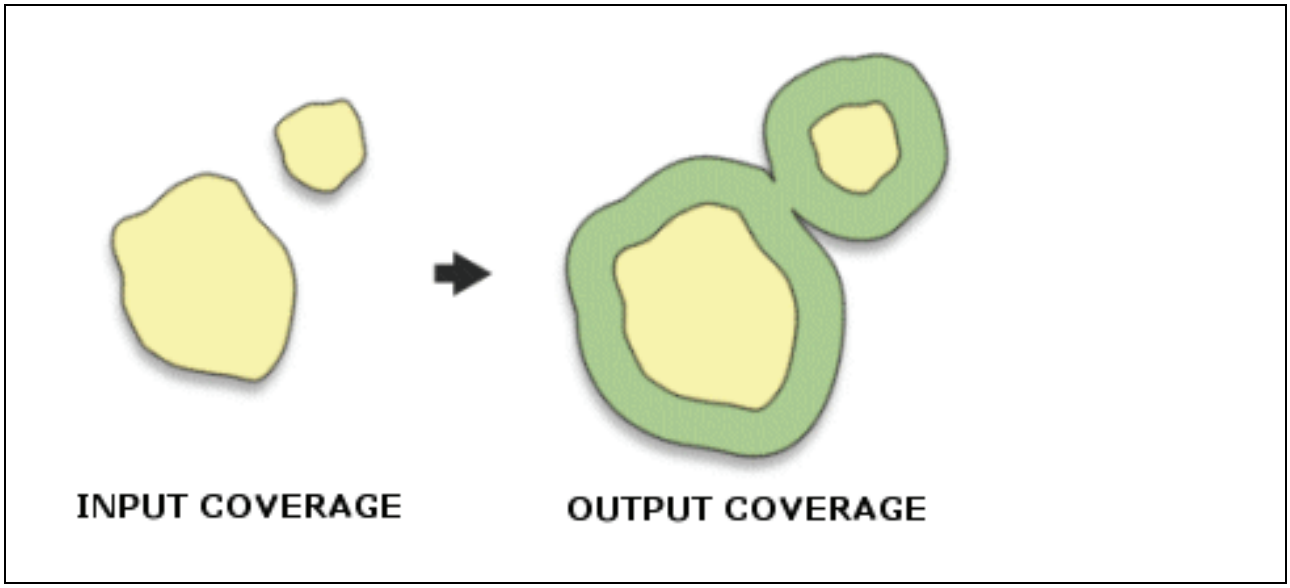
Types of spatial analysis and modeling can be classified into several groups:

Queries are considered the most basic type of analysis operation. They use a GIS program to answer simple questions posed by the user, with no changes in the database and no new data produced. Linkage between a data table and a map, or between a data table and a scatter plot, allow the user to check how typical an observation is against other observations. An example of a query using geocoded census data is, select all towns with a population greater than 1,000 persons. The term *exploratory data analysis* refers to investigations of patterns and trends in data using such techniques as querying.

Measurements are simple numerical values that describe aspects of geographic data, including the simple properties of objects, including length, area or shape, as well as the relationships between pairs of objects, such as distance or direction. The area of polygons in the pre-digital time was quite challenging to measure but can now be easily done by an algorithm on a computer. An example of an analysis using measurement is the selection of villages located more than a kilometer from a school, clinic, or well, which can then be further analyzed using the attribute information for the populated places themselves.

Transformations are methods of spatial analysis that use simple geometric, arithmetic or logical rules to create new datasets, either through combination or comparison. Transformations can include operations that convert raster into vector data, or a stream of GPS coordinates into a route or a boundary. Of all the transformational techniques, buffering is the most well known and important. **Buffering** involves building a new vector (or raster) object by identifying all areas that are within a certain specified distance of the original. Buffering can be performed on points, lines and polygons (see Figure 5.3) and can be weighted by attribute values. Buffering can be used to model travel time, for instance, or estimate the amount of territory that may be susceptible to flooding.

Figure 5.3 Buffering of a polygon object



Another example of transformation is **point-in-polygon** analysis, which determines whether a point lies inside or outside a polygon. This can be used to compare observations from a linear transect created from an aerial flyover to a polygonal dataset. **Polygon overlay** involves

comparison between the locations of two data layers. Comparison of the boundaries of two administrative districts could be used to troubleshoot errors in the field enumeration process.

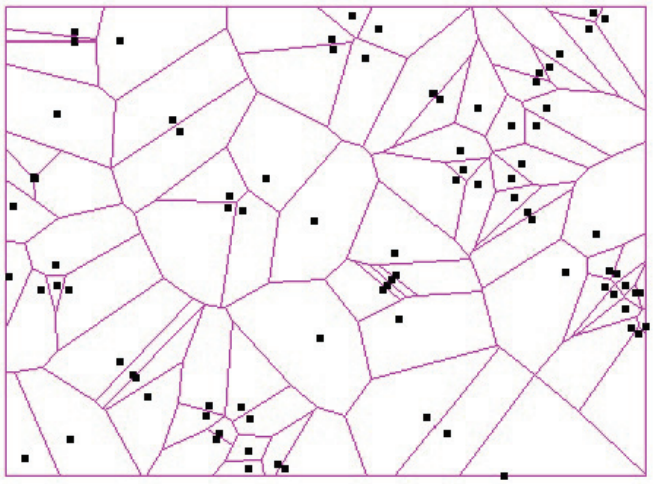
Spatial interpolation is a method designed to fill in values that lie between observations. A variety of methods including inverse-distance weighting and kriging are used to estimate the values of unsampled sites, based on Tobler's first law that all nearby objects are more similar than distant objects. In kriging, the general properties of a surface are modeled to estimate the missing parts of the surface. See Figure 5.4 for an illustration showing how contour lines can be derived through linear interpolation from a satellite image.

Figure 5.4 Example of linear interpolation creating contours

Thiessen polygons are spatial objects used to create areas around point data, based on the distances among points arrayed in two-dimensional space. This method assumes that the values

of the unsampled data are equivalent to those of the sampled points. See Figure 5.5 for an illustration of Thiessen polygons.

Figure 5.5 Thiessen polygons illustrated

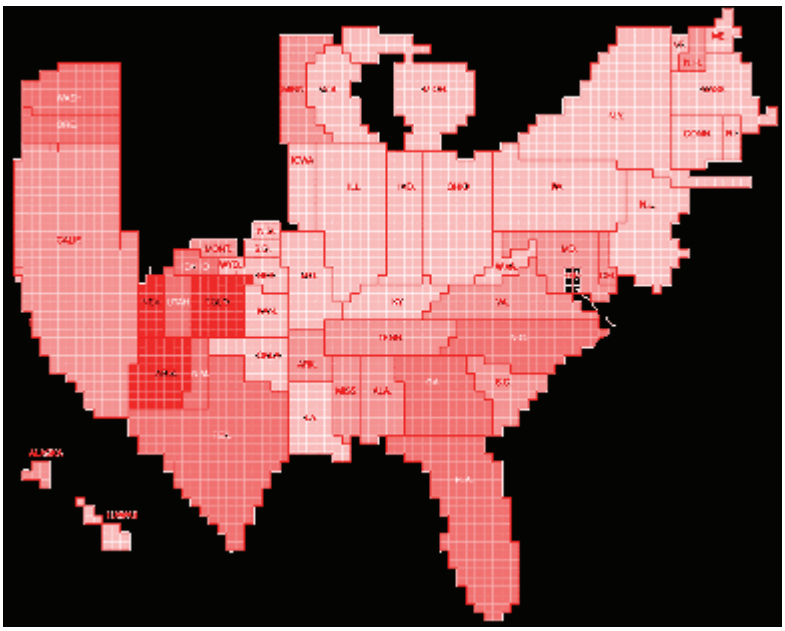


Descriptive summaries are a spatial equivalent of descriptive statistics (such as mean and standard deviation) that represent the essence of a dataset in 1 or 2 numbers. **Centers** are the two-dimensional equivalent of a statistical mean and are often used by population geographers to display the center of population using the weighted average of x and y coordinates of populated points. **Point pattern** or **cluster analysis** regards the distribution of points in space irrespective of their actual locations to determine whether patterns are random, clustered, or dispersed. The Moran statistic (between 0 and 1) indicates the general properties of attribute patterns and can be used to identify hot spots, where high values are surrounded by high values, or cold spots, where low values are surrounded by low values. These re particularly useful for identifying populations at risk.

Cartograms (Figure 5.6) are sometimes used to display census results. In a cartogram, the areas of the original polygons are expanded or contracted based on their attribute values. Analysts can

now obtain scripts and extensions from ESRI and elsewhere (for instance, see MAPresso – www.mapresso.com) to create them using their polygon layers.

Figure 5.6 Example of a cartogram



Map production and publication issues

Types of output

After the census is completed the statistical office will create publication quality cartographic outputs for a variety of purposes. Some examples are:

- Standard reference maps that describe each statistical dissemination unit defined during census data tabulation.

- Maps as illustrations in printed reports on census results or methodology. Here maps are not the main content of the publication. Rather, they complement the text. Often these will be printed in black and white which is easier to produce compared to full-color printing. For wide distribution, the number of copies printed will be relatively large. Printing will therefore be carried out by the census organization's print shop or an externally contracted printer.

- Printed census atlases can range from short brochure type publications to comprehensive hardcopy atlases with dozens or even hundreds of maps.

- Digital census atlases which are a cost-effective alternative to printed versions in countries where computers are widely available. Census atlases can either be based on static pre-prepared maps, or on a simple thematic mapping interface where the user can select the variables to map, the classification scheme, cartographic symbols and colors, and a basic layout.

- Maps can also be published on the Internet. The choice is again between static maps that are no different from other images or photos published on the Internet and dynamic mapping interfaces that give the user control over the thematic design process.

- Special purpose maps in various formats will be generated for in-house or outside census data users by special request. Such products will be printed in small numbers on in-house output devices such as laser or ink-jet printers.

- Presentation materials such as slide shows or large format posters on census topics benefit from the inclusion of maps.

Cartographic tools and software

The first generation of GIS packages did not provide convenient cartographic tools. Map output was created using command line interfaces or macro languages. To put text on a map, the user had to specify the coordinate on the map page for the text location, the text size and style as separate commands. The new generation of desktop mapping packages have much improved cartographic design functions. The user has access to numerous fonts, line and fill patterns as well as clipart that can be integrated in map design. The systems also come with special cartographic symbol sets providing point or line symbols commonly used on topographic and thematic maps. The user interface of desktop mapping packages is generally very similar to standard graphics software packages where the user can select styles from interactive menus and map elements can be moved and resized using the computer's mouse. The on-screen map display shows quite realistically how the map will look on the printed page.

The cartographic design functions of modern desktop mapping and GIS packages will satisfy most users' requirements. For some applications, however, professional cartographers prefer to export the basic map from the GIS and import it to a graphic design or desktop publishing or graphics package. These packages provide sophisticated graphics functions such as 3-D effects, graduated fills or transparency which give the cartographer greater flexibility in design. To copy from the GIS to the graphics package two options are available. One is to use the standard cut and paste options in the Windows environment. The other is to go through an intermediate file in a standard format that can be imported by a graphics package (see the next section on output options).

Output options

Digital files

All GIS and graphics packages allow the user to export the map layout to a number of graphics file formats. This option is useful for a number of reasons. It allows the exchange of files between packages. For instance, a basic map from a GIS and charts from a statistical software package can be exported to a graphics package where the final page layout is designed. The finished graphic can be imported to a text processing software to be integrated in a report or publication. Most of the graphics in this handbook were produced in this way. Graphics files can be incorporated in web sites as static map images and can also be exchanged as file attachments via electronic mail.

Graphics file formats—similar to GIS data structures—can be divided into those that support vector graphics and those that are raster or image files. Raster images represent graphical objects as variations in color or graytones of tiny dots or pixels arranged as a regular grid. Continuous color tones or grayscales are used for photographic-type pictures. Fewer colors are needed to show more discrete objects typically found on thematic maps.

Vector graphics formats represent graphical objects as points, lines and areas using an internal coordinate system that can either be device independent or is tied to an output page size. Some file formats can handle both raster images and vector objects. Such formats are useful for GIS maps that combine, for instance satellite images with line and polygon data layers. Regardless whether a raster or vector graphics format is used, the graphical content needs to be rasterized before the information can be displayed on the screen or printer which are both essentially raster display devices. This is done automatically by the computer's operating system and printer drivers.

Below is a brief description of the most commonly used file formats. This list is by no means complete as there are dozens of different formats. See Murray and van Ryper (1994) for a comprehensive overview.

Raster image formats

Raster images can be created by GIS or graphics packages directly. In some instances two other options for creating images are useful. One is to use the screen-capture command in a raster-oriented graphics package. These "screen grabbers" are sometimes better at preserving the original display colors than the export functions in GIS or graphics packages. A second option is to use a specialized piece of software or hardware to convert graphical objects into raster images. These raster image processors (RIP) can, for example, produce very high resolution images that preserve all the details of a vector format. The resulting output files can become very large, however.

File size depends on two factors: the number of colors available in the image and the degree of image compression. For example, an image format that supports only two colors (black and white) requires only one bit to represent each pixel. Eight bits (one byte) per pixel can store up to 256 colors and high-end displays or image formats which use 24 or 32 bits per pixel can store more than 16 million colors. For thematic maps, a relatively small number of distinct colors is usually sufficient. 16 or 24 bit image formats are more useful for photos or photo-realistic graphical images.

Most image formats use some form of compression that reduces file size. The simplest compression scheme is run-length encoding, a technique that is also used in some raster GIS systems. If there are many pixels with the same color in an image row, the system stores the

number of repetitions and the pixel color only once. For instance, five pixels with color number four, would be represented by a pair of numbers (5,4) rather than as (4,4,4,4,4). The color number actually represents an index to a color table which is contained in a small file header and contains the color specification in a common color model such as RGB.

Some standard raster file formats are:

- **BMP**. The Microsoft Windows device-independent bitmap (DIB) format. It allows Windows to display the bitmap image on virtually any type of display device. This is one of the most basic raster file formats. Run-length encoding is supported, but file sizes are usually bigger than for other image formats.

- **TIFF**. Tag Image File Format is one of the most widely supported raster image formats. It supports various numbers of colors and a number of compression schemes. TIFF images can be imported by most software packages that support graphics, although problems can sometimes occur importing images created on a different hardware platform.

  TIFF has specific importance for geographic applications, since it is often used as a format for displaying satellite images, aerial photographs, scanned maps or other raster data in a GIS or desktop mapping package. The need for a platform independent standard file format for geospatial imagery resulted in the development of the GeoTIFF standard. This standard provides the specification for information included in the TIFF image header that describe all geographic information associated with the image, such as the projection, real-world coordinates, map extent, etc, while still complying to standard TIFF format specifications. GeoTIFF is supported by most major GIS vendors, government agencies and academic institutions.

- **GIF**. The Graphics Interchange Format was designed for interchange of raster image graphics across hardware platforms. It supports a compression scheme that reduces file sizes significantly and is therefore optimal for exchange through computer networks. In fact, the format was developed by CompuServe for use in their early bulletin board services. GIF, which supports up to 256 colors, is one of two raster image formats supported by web browsers. Most non-photographic raster images on web pages are in GIF format.

- **JPEG**. Developed by the Joint Photographic Experts Group, the JPEG format was developed as a compression scheme for images that have a very large number of colors or grayshades such as photographs or photo-realistic graphic images. JPEG format is also supported by web browsers and is used to display photographs on web pages. JPEG has a variable compression option, which is not fully reversible. That means that a photograph that has been exported with a high degree of compression cannot be restored to show all details in the original photograph.

- **PNG**. The PNG (Portable Network Graphics) file format was created to be the free and open-source successor to the GIF file format. The PNG file format supports true color (16 million colors) whereas the GIF file format only allows 256 colors. PNG excels when the image has large areas of uniform color. The lossless PNG format is best suited for editing pictures, and the lossy formats like JPG are best for final distribution of photographic-type images because of smaller file size. Many older browsers do not yet support the PNG file format, however with the release of Internet Explorer 7 all popular modern browsers fully support PNG. The Adam7-interlacing allows an early preview even when only a small percentage of the data of the image has been transmitted.

- **RAW**. RAW refers to a family of raw image formats that are options available on some digital cameras. These formats usually use a lossless or nearly-lossless compression, and produce file sizes much smaller than the TIFF formats of full-size processed images from the same cameras. Unfortunately, the raw formats are not standardized or documented, and differ among camera manufacturers. Many graphic programs and image editors may not accept some or all of them, and some older ones have been effectively orphaned already. Adobe's Digital Negative specification is an attempt at standardizing a raw image format to be used by cameras, or for archival storage of image data converted from proprietary raw image formats.

Vector file formats

Vector file formats are more closely associated with vector GIS data. They can represent line or polygon data more compactly and will preserve the full resolution of the original GIS data layers. Some standard vector graphics formats are:

- **WMF**. Windows Meta File is a graphics file format for use in the Windows environment. It is most often used for vector data, but can also store bitmap images. Enhanced WMF (EMF) files are a more comprehensive variation of WMF format developed for the 32-bit Windows environments. WMF is one of the most stable formats for exporting and importing graphics files among Windows applications. WMF is also one of the formats used by Windows when a graphics object is copied to the clipboard and subsequently pasted in another application.

- **VML**. “Vector Markup Language (VML) is an XML language used to produce vector graphics. VML was submitted as a proposed standard to the W3C in 1998 by Microsoft, Macromedia, and others, but was rejected as a web standard because Adobe, Sun, and others

submitted a competing proposal known as PGML.[1] The two standards were joined to create SVG. Even though rejected as a standard by the W3C, and largely ignored by developers, Microsoft still implemented VML into Internet Explorer 5.0 and higher and in Microsoft Office 2000 and higher. Google Maps currently uses VML for rendering vectors when running on Internet Explorer 5.

- **KML**. (Keyhole Markup Language) is an XML based language schema for expressing geographic annotation and visualization on existing or future web-based online maps (2D) and earth browsers (3D). KML was originally developed for use with Keyhole, the predecessor company that was acquired by Google in 2004. The KML file specifies a set of features (placemarks, images, polygons, 3D models, textual descriptions, etc.) for display in Google Earth, Maps and Mobile, or any other 3D earth browser (geobrowser) implementing the KML encoding. Each place always has a longitude and latitude. Other data can make the view more specific, such as tilt, heading, altitude, which together define a "camera view". KML files are very often distributed as KMZ files, which are zipped KML files with a .kmz extension.

- **CGM**. Computer graphics metafiles are an international standard for storing two-dimensional graphical data. Initially developed as a pure vector standard, later versions also support raster images. There are three CGM format types: one is a character encoding that reduces file size and increases transmission speed, one a binary code for speed of access, and one a clear text mode for file-based editing.

- **HPGL**. Hewlett-Packard Graphics Language is a file format that was initially used for pen plotters. Before the advent of large-format inkjet and electrostatic printers, pen plotters were the most widely used output device for GIS projects that needed to print large maps.

- **DXF**. The DrawingXchange Format was developed by Autodesk, a software producer specializing in computer aided design (CAD) and GIS software. Initially designed to exchange Autodesk native files between platforms, DXF has become a standard exchange format that is supported by most GIS packages and many graphics software.

- **PS and EPS**. Postscript is essentially a programming language for describing vector data in a plain text file. It is the most widely used page layout description. Postscript was developed by Adobe, a graphics software company. Optimized for scale-independent vector graphics, postscript files can also incorporate raster images. The main use of postscript is as an output format for sending documents and graphics to postscript printers. Postscript is thus fundamentally an output format. Many graphics packages support postscript import, but because the postscript codes are not completely standardized, it is often not possible to import postscript files for further editing if they were created in a different computer program. This is especially true when the postscript file travels across hardware platforms. Sometimes, it is not even possible to import a postscript file created in the same software.

  While it is often not possible to modify an imported postscript file, most software packages will be able to incorporate a postscript file in a document. Instead of the file content, only a labeled box will be shown on the screen display. Once sent to a postscript printer, the actual postscript file content will be printed. Since postscript files are scale independent, the imported postscript graphic can be resized to fill the desired space.

- **PDF**. The Portable Document Format was also developed by Adobe. Its initial use was for distribution of complex documents—containing text and graphics—on the Internet. PDF files can be created from any text processing or graphics package using the Adobe Acrobat printer driver has been installed. The PDF reader can be downloaded free of charge from the Adobe web site. Some experts predicted that PDF format will replace postscript files as the main standard for high-level graphics printing. *[Need to verify with EG if this is indeed true.]* The PDF language is simpler than postscript which makes PDF files easier to rasterize. Rasterizing of a graphics file is necessary for display on a computer screen and for high resolution printing.

Personal printer

For small print runs or quality control plots a census office should have one or a number of printers available. The following paragraphs briefly describe the most popular printer types.

- **Laser** printers employ a laser beam and a system of optical devices to selectively discharge a photoconductive surface. Oppositely charged toner is then brought in contact with that surface and is attracted to the areas that retain the charge. The toner is then transferred onto the page and fixed. A process similar to electrostatic photocopying is then used to apply the image from the drum onto the paper. Monochrome laser printers can achieve an output quality that is close to professional typesetting systems. Color laser printers reached a price range to be considered for most graphics application environments.

- **Inkjet** printers produce output by squirting electrically charged drops of color through a nozzle onto the page. Liquid inkjet printers use liquid ink that dries through evaporation. Ink is sent through the nozzle using hydraulic pressure in the so-called pulsed inkjet technique.

Thermal inkjet, in contrast, uses heat to create a bubble of ink in the ink nozzle. The bubble is forced through the nozzle onto the paper when it is large enough. Solid inkjet printers use ink that needs to be melted from its solid state before it can be squirted onto the paper where it solidifies quickly. Solid inkjet printers produce finer dots on the page compared to liquid inkjet technology. Inkjet printers work with plain paper, but to achieve highest possible output quality, specially coated paper designed for use with inkjet printers is usually recommended. Due to their reasonable cost and ease of operation inkjet printers, which are available for a range of output page sizes, are currently the most widely used color output device.

- **Thermal** printers require special paper and ink coated ribbons which are moved across a thermal head. Ink is fused to the paper where the thermal head applies heat. The color ribbons are coated with three (CMY) or four (CMYK) colors so three or four passes of the thermal head across the paper are required. In thermal wax printers the heat causes a layer of colored wax to be fixed to the paper. In thermal dye processes, the dye is diffused into the printable surface. Dye diffusion printers usually achieve higher resolution and more color variation compared to thermal wax printers.

- **Electrostatic** printers use toner that is transferred through electrical charges to a nonconducting surface. Toner is either attracted or repelled. Direct electrostatic printers apply the charge directly to the specially coated paper. Toner for each color is applied in separate passes. Subsequently, the toner is fused to the paper after all colors have been applied. Another electrostatic process is color xerography which uses a drum or belt that is charged when exposed to light.

Printing technology is constantly changing and the range of available products is very large. In choosing appropriate printers, a census office needs to consider the following criteria:

- cost of hardware, maintenance and printing per page,
- throughput (pages per minutes),
- output resolution in terms of dots per inch (dpi) and number of color or gray tones that can be produced,
- media size, and
- supported media types (plain paper, specially coated paper, transparencies, etc.).

Many draft maps do not have to be printed in color. In fact, small-format black and white maps can be more easily photocopied. Laser printers that support A4 or letter size paper combine fast printing with very high resolution (600 dots per inch and more). They are ideal for printing reports and other documents that consist mostly of text with some graphical illustrations and maps.

Color printers are useful for printing complex maps for which monochrome shading and symbolization would be insufficient. Inkjet printers are commonly used color printers—from A4/letter size desktop printers to large format printers (e.g., 60 x 90cm or 24x36 inches)—although color laser printers may supplant them. They produce high quality maps at 600 dpi. Printing speeds are still relatively slow for inkjet printers.

In deciding upon a suitable printer for a GIS project, cost is a major issue. Something to keep in mind is that the purchase price of a printer is only one—often relatively minor—cost component.

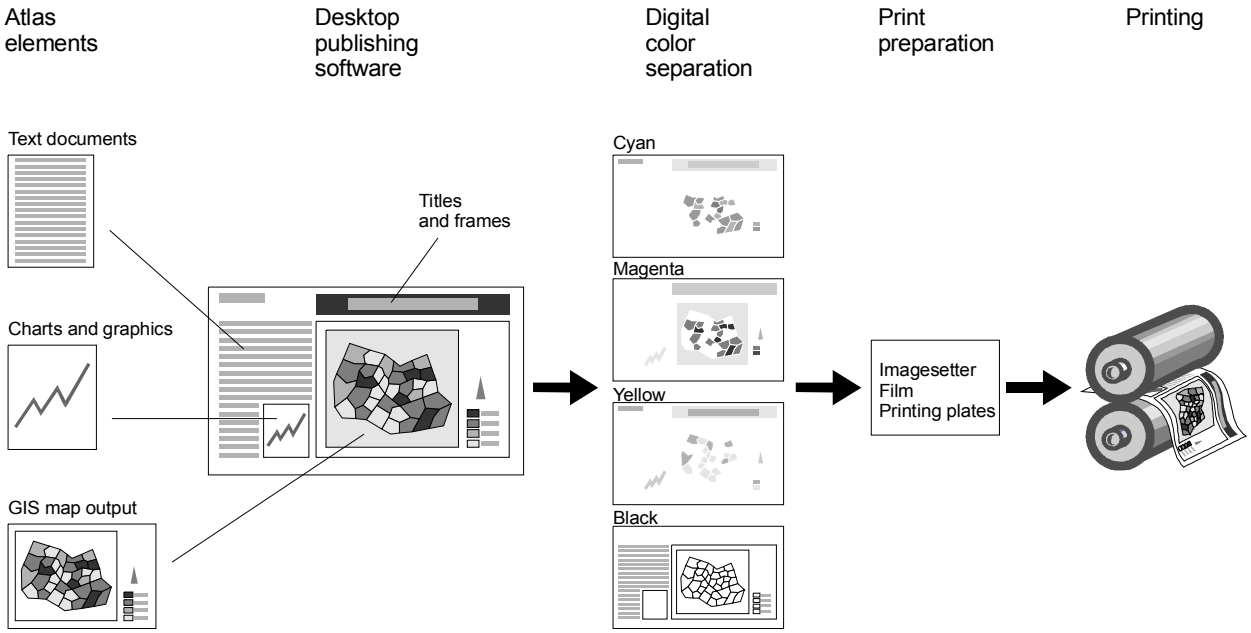
While printer prices have dropped considerably, the cost of ink cartridges and special paper have remained fairly high. In some cases it appears that hardware prices are kept very low by printer manufacturers who hope to profit mainly from selling the hardware specific supplies. In addition to the purchase price, one should therefore also compare the printing cost per standard page (e.g., where 5 percent of the page is covered by ink). Computer trade journals often publish comparisons.

Commercial printing

For larger print runs, personal printing devices are too slow and the costs per page are too high. Brochures, posters, or census atlases will therefore be printed in an in-house or commercial print shop. If print volume is high, analog printing processes, where printing plates are produced and used in lithographic or similar printing machines, are currently still cheaper and faster than digital print processes.

The process up to the production of printing plates, however, is already almost exclusively digital. The typical production process for a digital census atlas may look like this (see Figure 5.7). After an initial planning stage in which the text, graphics and map contents are specified, census cartographic staff produce all maps for inclusion in the atlas. These maps are stored in postscript format ready for printing. For complex map designs that incorporate graphics produced in external packages or photographs, the layouts might be produced in a high end graphics package. Other census office staff members will write the text to accompany the maps, tables, references and other textual content in standard word processing packages.

Figure 5.7: The digital printing process



In a second step, all atlas elements are combined in a desktop publishing program. Text headings, figure captions, pictures, text and graphical elements are formatted and arranged in a visually appealing layout that will match exactly the page size of the printed product. This work might be done in-house or by a contractor or external service bureau.

Once the final atlas layout has been produced it is saved in a digital output file. The most common file format is an encapsulated postscript file, but some software specific file formats can also be used by commercial printers. Most high-end graphics and desktop publishing programs can also produce color separations, which are either stored in separate files or all in the same file. The actual printing machine uses four print plates, one each for the colors cyan, magenta, yellow and black (the CMYK color model). The colors on the maps and graphics are produced as additive combinations of various percentages of these four colors. The digital files are then sent to an image-setter which creates the film from which the printing plates are produced. Using digital files for producing the film will generally produce the best results. Camera-ready copies printed on a laser printer, which are reproduced by photographic techniques might be cheaper, but will not produce the same high resolution. Unless a production line has already been established and tested, it is usually desirable to obtain and evaluate a color proof from the printer before final production. Many vendors of printing hardware and software also provide extensive information and other resources on their web sites.

Digital geographic databases for dissemination

Even with advances in Internet use worldwide, the publication of hardcopy census map products will continue to be one of the primary means of disseminating geographic

census results. Access to computers varies by country, and even where computer use is widespread, users may still prefer a printed product. In parallel to the production of printed census maps, a census office should also pursue a digital data dissemination strategy.

Demand for digital databases that consist of extractions of the census agency's digital geographic master database will continue to increase. Census data are an important input in policy planning and academic analysis in many fields. Health service provision, educational resource allocation, design of utilities and infrastructure, and electoral planning are some applications where government agencies require spatially referenced small area population statistics. Commercial users employ such data for marketing applications and location decisions.

Advances in data processing and GIS mean that producing a digital geographic database at the enumeration area level is increasingly within reach of many national statistical organizations. NSOs should consider the costs and benefits of providing data at such a detailed level.

Some benefits include unsurpassed detail and precision, the potential use of census data in numerous applications--especially when overlaid on other geographic data such as terrain, and the relative ease of management and storage of thousands of units.

Some costs include the expense in processing and data management, possible data disclosure issues, and quality control. Releasing census results at the level of what is a data collection unit may expose unanticipated errors in the process.

One alternative to release of an EA dissemination database is the creation of a derived product at a similarly detailed scale.

Textbox: South Africa's experience with the cluster level of geography.

Digital data dissemination strategies and users

The wide range of potential users of small area census data means that the census organization needs to pursue a multi-leveled digital data dissemination strategy. Broadly, we can distinguish between the following types of users:

- **Advanced GIS users**, who want to combine small area census data with their own GIS data on health facilities, school districts, or sales regions, for example. These users are sometimes called data extractors.

- **Computer literate users** in the government, commercial or private sector. These users want to be able to browse the thematic information in a census database spatially. They will want to produce thematic maps and thus need to be able to perform simple manipulation of cartographic parameters. Simple analytical functions such as aggregation of census units to custom-designed regions should also be possible. These users are sometimes called data manipulators.

- **Novice users**, who mostly want to view pre-prepared maps on a computer and perhaps perform some basic queries.

The first group of users will want access to spatial and attribute information in a comprehensive digital GIS format. The census office needs to supply comprehensive documentation (see metadata section of Chapter 2) on the geographic parameters used for the GIS database as well as on the individual census variables. The spatial information will be distributed in an open GIS format that can be easily converted into any number of commercial GIS formats.

The second group of users is best served with a comprehensive, pre-packaged application that is designed for a commercial or freely available desktop mapping package. Documentation requirements are somewhat smaller, since the users are unlikely to change the geographic parameters of the database or perform more advanced GIS operations.

For the third group of users, finally, the best data distribution strategy is to produce a self-contained digital census atlas. This atlas could consists of a series of static map images, for example, in the form of a slide show. Or it could be a very simple mapping interface with pre-designed map views that allow basic queries. Both, static maps and a simple map interface, can be made accessible through the Internet.

Definition of data content

The first step in preparing the GIS databases for general release is to define the data content. The following questions need to be addressed:

Up to which level will data be released?

To maximize the overall benefits of census data collection, the objective of the census organization should be to release geographically referenced census data at the smallest level that does not compromise data privacy. Even at the enumeration area level, there may be special reporting zones that contain only a few households, and for which census data can not be released. If necessary, data for selected reporting zones must be deleted or re-coded.

One large GIS database or a family of census databases?

A high-resolution census GIS database will consist of thousands of reporting units. Such data volumes will be beyond the computing capacity of average data users. Instead of distributing one large database, the census organization should consider producing a family of census databases. At the medium resolution level — for instance districts — a national summary database can provide a sufficiently detailed overview of socioeconomic conditions in the country. For each major civil division or even each district, separate databases which show indicators at the sub-district and enumeration area level can be constructed. Individual databases can also be useful for major urban areas.

Finally, a point database of settlements in the country with associated census data will serve the needs of users who do not need the spatial resolution of a GIS database of reporting units. This database should at least contain all settlements classified as urban and the aggregate census indicators for each town or city. Ideally, a village level database should also be constructed for the benefit of planners in the health, education or agricultural sectors. A village database can be based on a gazetteer of place names and locations if such information has been collected during census mapping.

Offering databases for sub-sections of the country will increase data use. Many users only need census information for a relatively small region. A subset of the national census database is easier to process by users with moderate GIS computing capacity. Also, in countries where the data access fees are larger than the cost of reproduction, smaller data sets are affordable to a larger number of non-commercial users.

If separate databases are distributed, care has to be taken that the individual parts or tiles are compatible. That means that the shared boundaries between database subsets need to match exactly. Separate pieces of the database should be in the same geographic reference system and have the same attribute database definitions. If the master database used by the census office is very detailed, it may be beneficial for some data users if a more generalized version of digital census maps are available as well. Some countries offer digital census maps at different nominal map scales or coordinate accuracy. Fees can be higher for users requiring very high accuracy and detail.

Many commercial GIS data producers distribute their data in latitude/longitude (i.e., geographic) coordinates, rather than in a specific projection. Geographic coordinates are the most general reference system and can easily be converted into other projection systems, if the user wants to use the census boundaries in combination with other data layers. Specific national projections and coordinate systems, in contrast, may not be supported by the GIS software. Users would then have difficulties to employ the census database for geographical analysis applications.

How tightly should boundaries and database be integrated?

Census GIS databases are characterized by their large number of attribute fields. Census questionnaires provide information that is stored in possibly hundreds of variable fields. Usually it is impractical to store all of these in the same data table. A better approach is to select a small number of most important indicators in the geographic attribute table and provide the remaining information in a series of separate tables. These external tables can be organized by topic — demography, household data, etc. The user can then link tables to the GIS by the common geographical identifier as needed.

Data formats

Coordinate data

GIS software packages differ widely in terms of the data formats that are supported. Each commercial package has its own native data format. In addition, import and export functions allow the user to convert data from a selected number of external data. In some instances, these conversion functions need to be purchased separately. For example, the GIS company ESRI sells a Data Interoperability extension that allows the conversion of most popular GIS and CAD formats.

Despite some efforts by commercial and public GIS groups (see www.opengis.org), there is still no universally accepted and widely used generic or open source data exchange format. Instead, a number of exchange formats developed by leading GIS vendors have become de-facto standards that are also supported by other software systems. The most important of these are described briefly here:

- AutoCAD DXF format (.dxf) originated in the CAD world. It is well suited to transfer the geographic coordinate data, but not as good at converting attribute information.

- Arc/Info export format (.e00) was developed as a cross-platform exchange format for GIS databases produced by the Environmental Systems Research Institute's Arc/Info GIS. Export files can be compressed to support smaller file sizes. However, to ensure maximum compatibility it is usually better to use the uncompressed export format. The resulting files can then be compressed using a standard compression and archiving program such as PKZIP. The .e00 format is not published, but many other GIS packages have developed import routines.

- ArcView Shape files (.shp) are a simpler format used by ESRI's ArcView desktop mapping software. A shape file database consists of several files containing the coordinate data, a spatial index, and attribute data respectively. Their file formats are published and many other GIS systems are able to import shape files.

- The File Geodatabase format is the new recommended native data format used with ArcGIS. File geodatabases are stored as folders in a file system, and are designed to support the full information model of the geodatabase including topologies, raster catalogs, network datasets, terrain datasets, address locators, etc. Each File Geodatabase file can be as large as I (1?) terabyte. Personal geodatabases use the Microsoft Access file system (.mdb) and can be as large as 2 gigabytes.

- MapInfo Interchange Format (.mif) is used for exchange of files produced with MapInfo, a leading desktop mapping system. MIF files are in ASCII format and can be read by many programs.

- MicroStation Design File Format (.dgn) is used by Bentley's Modular GIS Environment (MGE) and Geographics GIS packages. The format does not support attribute data directly but provides links to external database tables. A separate export format combines geographic and attribute files.

All of these formats support boundary and attribute information. Any commercial GIS will have an import function for at least one or two of these formats. Ideally a census office should offer their public-release GIS databases in several formats to serve a wide range of users with varying GIS skills and different software platforms. The choice of distribution formats should be guided by information on which mapping systems are most widely used in the census user communities and by the flexibility and robustness of the data format.

Distribution of GIS data in their native, internal format—for example, a directory containing an Arc/Info coverage or a MapInfo workspace—is not usually a viable option. Data in native formats often cannot be transferred to another operating system, pathname incompatibilities may be encountered, and other GIS packages are usually unable to import native GIS data formats. It is thus always preferable to use a robust data exchange format as implemented by most commercial GIS packages.

Tabular data

New GIS software development has de-emphasized the import of tables in favor of relational databases such as Oracle and Access. Most GIS packages still do support several file formats for attribute data. Some also have functions to connect the coordinate database to an external database management system. For data distribution, however, it is

better to use a simple, widely used file format for data tables. The most widely used format is the dBase (.dbf) format which can be produced by most database management and spreadsheet packages as well as by census tabulation packages such as REDATAM and CSPro.

While tabular data distribution in dBase format ensures wide compatibility with GIS packages, the format has a number of limitations. For instance, field names, which are listed in the first row of the table, are limited to 10 characters. The spreadsheet or database management's software documentation will provide details about compatibility issues. In the table layout, the most important field is the common identifier that is used to link the attribute data to the reporting unit boundaries. This field should be located in the first column of each attribute table. It is usually also good practice to sort the data sets in a consistent order, for instance by their geographical identifiers.

Documentation

Consideration must also be given to the file formats for the data documentation. Simple ASCII text files can be read by any user. However, they do not support graphics, complex tables, or formatting of text. The Adobe Acrobat system's PDF format is now becoming a standard for platform independent distribution of formatted documents. Since the Adobe Acrobat reader is available free of charge, PDF documents are accessible by any user who can download the software.

An alternative is to produce documentation in a format readable by web browsers, which are also available free of charge from Microsoft, Mozilla, etc. HTML files allow a

considerable degree of formatting and when located on a CD/ROM or hard disk can be accessed even when no Internet connection exists.

File naming conventions

Although the Windows, Macintosh, UNIX, and LINUX operating systems all support long file names, it is good practice to use the DOS 8.3 file naming conventions for all data and documentation files that are distributed. Some users may be working under DOS or Windows 3.1 or with older GIS software packages. Short file names can reduce incompatibilities, for example, with older network software, to a minimum. Consistent naming conventions that are explained in the documentation will make it easier for the users to find the data they need quickly.

Compression

GIS files can be very large and together with the tabular data, the set of distribution files may become quite voluminous. Especially for Internet data delivery or for distribution on data diskettes, file compression will greatly facilitate data distribution. The most widely used compression software for the Windows environment are PKZIP and Winzip utilities. They are available on most computers and utilities that can extract files from the compressed archives also exist for the UNIX operating system. Self-extracting files are more convenient for inexperienced users and do not require a decompression utility. However, they are operating system specific and should only be used if the target computer platform is known.

Documentation and data dictionaries

The documentation that will accompany the data set distribution does not have to be as comprehensive as the in-house information that is compiled for all databases (see chapter 2). Data users will usually not need detailed information on data lineage or processing steps, and ease of interpretation is more important for external users. Thus, the documentation should contain a clear, concise and complete description of those aspects of the database that are relevant to a user. Provided that the census office maintains a comprehensive meta-database, the user-targeted data documentation can be compiled very quickly. Data documentation may include the following information:

- Data set names and reference information including all data sources;
- Narrative content of the data sets;
- Description of the hierarchy of administrative and reporting units and their relationship to other features (e.g., settlements). This should include a clear description of the statistical definition used for each type of reporting units. A complete list of all reporting units and their geographic codes is useful;
- Software and hardware requirements;
- General data format, decompression and installation guidelines;
- Geographic referencing information (all geographic data sets should be in the same reference system):

- Cartographic projection with all required parameters such as standard parallel or meridian, false easting and northing, etc;
- Coordinate units (e.g., decimal degrees, meters, feet);
- Source map scale; i.e., the scale of the hardcopy maps from which boundaries were digitized;
- Geographic accuracy information. For instance, any numeric accuracy information available for the source maps can be reported. If a quantitative assessment of data quality is not possible, accuracy could be described in more general terms;
- Printed maps of the GIS data sets are a useful addition to the documentation. For example, it enables the user to verify that import of maps has been performed correctly;

- Conventions for dealing with disjoint reporting units;
- Information about related products, for example more detailed census GIS databases or additional data files that can be used with the boundaries;
- Bibliography of relevant census publications;
- Contact information for user support; and
- Disclaimers, copyright information, etc.

In addition, each GIS data set should be accompanied by a data dictionary that provides information for each individual GIS data layer or data table. This should list the following information:

- File names and file formats;
- Feature types (points, lines or polygons);
- Relationship between coordinate data files and associated external attribute data tables;
- For each field in the attribute table and in additional external tables;
    - Field name;
    - Description of field content (e.g., Total population, 2005) and the exact statistical definition employed. For derived demographic indicators, the formula used can be given, for example using field names of the variables employed as the numerator and denominator;
    - Field definitions including the variable type (e.g., real, integer or character field), the range of acceptable values, and the conventions for dealing with missing values. For classified data, the coding scheme needs to be explained in detail. For example, in a settlements database a numeric field called TYPE may use a "1" for the national capital, "2" for provincial capitals, "3" for district administrative centers, and so on;

- Any available data quality information that allows users to judge the suitability of the data to a given task.

The data documentation and data dictionaries can also be incorporated into a comprehensive users guide. A users guide might contain a more detailed explanation of database content, data lineage and quality. Step-by-step explanations of example applications or copies of census maps created with the database can also be included. A sample data dictionary is presented in Annex 4.

Preparation of deliverables

Quality control is an important step before release of the final product for reproduction. After producing the final version of all databases in the form in which it will be distributed (e.g., compressed), the database should be tested on all target platforms (e.g., Windows environment, UNIX, Macintosh, LINUX).

For many users, the CD-ROM remains the most appropriate distribution medium for large data sets. A CD-ROM can hold up to 700 MB, and most computers are equipped with CD-ROM readers. CD writers are also quite inexpensive, so that digital masters can be produced in-house. This also allows distribution of customized data sets for which only a few copies are required. For wider distribution of large data sets, CD-ROM offers the advantage of low per-unit cost of production, durability, and their readability on multiple hardware platforms.

DVD (Digital Video/Versatile Disk) technology has in some areas supplanted the CD-ROM, and DVD writers are becoming more common on desktop computers. A current, single-sided, standard DVD can hold 4.7 gigabytes of information.

In the longer term, most data distribution will be done via the internet. Currently, limited bandwidth—the amount of data that can be transferred in a given time period—is still hindering distribution of very large files. Download times are often unacceptable due to shortcomings in the Internet infrastructure in many countries. The main bottleneck, however, are modem connections from homes or offices to the main Internet cables. Large files can be transferred to academic, government or commercial users who have dedicated high-speed Internet access.

Internet data distribution eliminates much of the cost of reproduction for the census organization. The remaining costs are for development of the software interface, maintenance of the web site and incremental use of web server resources. Census GIS databases can thus be provided to the user at very low cost or free of charge. Some organizations may, however, decide to charge for on-line data. One reason may be to cross-subsidize a publication program for users without access to the Internet. Another is where the organization intends to recover parts of the cost of data collection and compilation of the census data.

Legal and commercialization issues

Data copyright

A copyright is the exclusive and legally guaranteed right to publish, reproduce or sell a piece of work—in this context a digital geographic database. Because digital data are easy to reproduce, copyright issues concerning GIS databases are a more pressing issue than they have been for paper maps. The census office thus needs to develop a data access policy for tabular as well as cartographic census information.

Copyright covers two areas: moral rights and material rights. Moral rights protect the integrity of the work in prohibiting any alterations to the original product. Material rights refer to the right to any monetary benefits when the product has been released for reproduction, use or transformation. Any rights granted by the copyright holder will be specified in a license agreement.

The copyright issue is related to the pricing policy for digital data products. A census organization has several options in deciding upon a pricing strategy for digital spatial data. The agency can decide:

- to bear the full cost of census data collection and distribution;
- to charge for data distribution cost (cost of media and shipping);
- to recover all or parts of the cost of data collection and compilation;
- to produce revenue beyond the actual cost of the GIS investment and data development.

Trade-offs in the commercialization of geographic data

Copyright laws differ from country to country. On one extreme, some governments have no copyright on information that is produced by public agencies. The rationale is that since tax-payers have already funded data collection, they should not be charged again for the use of the data. As a consequence, GIS data produced by public organizations is distributed free of charge or at the cost of reproduction. Also, any commercial enterprise can use government information, repackage it, and sell it at a profit.

In the United States, for example, free access to public data has led to a large service industry that produces spatially referenced census data in various formats for sale to private, commercial, and public users. Although companies charge for the data, the non-exclusive use of the census data has brought many companies into the market. This competition has kept the price for re-packaged census data low while increasing the range of specialized products. Users who are willing to do their own data conversion still have access to the free data.

The benefit of this development has been a very wide use of census data for geographic applications. The increased number of users has in turn encouraged the commercial development of easy-to-use desktop mapping packages and the provision of value-added services. The overall economic benefits of this development are high as tax revenues increased and improved access to information led to productivity gains and better decision making in the public and private sectors. These benefits have justified the royalty free release of data which was essentially a public subsidy for private companies.

In other countries, shrinking government budgets have increased the pressure on public agencies to generate income to support their operations. As a consequence, prices for

geographically referenced census information are sometimes very high, thereby limiting use of the information . These prices may reflect the commercial value of such data to, for example, financial institutions and businesses. Yet, they may price small companies or non-commercial users out of the market for census information and may limit the overall use and therefore benefits of census GIS data. As Prevost and Gilruth (1997) point out, cost recovery efforts that put census GIS products out of reach of non-commercial users often lead to illegal copying of data sets, time-consuming duplication of data development from original source materials, or use of alternative, cheaper and lower-quality data.

Restrictive licensing agreements also preclude or hinder the distribution of derived census products and services. This lowers the public welfare effects from census data collection. The reduced overall economic impact due to the absence of such spin-offs may well be larger than the increased revenue for the census organization. In fact distribution policies for government produced data in some countries are moving back to a free or low-cost approach due to a realization that the benefits of charging higher prices do not justify the cost of enforcement of copyrights and of lost societal benefits due to the reduced use of vital information.

Access to the data and secondary uses are also often restricted where the census office collaborates with a private data producer or where data from public or private data producers are used to produce census maps. For instance, the census agency may enter into an agreement with a private mapping firm that absorbs part of the cost of digital map production for the census. The firm will only be able to recover its investment if it is awarded an exclusive right to market the geographic data (this will, of course, not be an

issue where the agency simply purchases the services of the company, and all outputs remain the property of the census organization).

If data from other agencies—for instance the national mapping agency or local authorities—are used for producing census maps, pricing, copyright issues and the definition of source and credit information shown on the census maps need to be clarified in detail. Conflicts over copyright issues should be avoided especially because the census agency will likely require the collaboration of those agencies for future census mapping activities.

In most countries the trade-off between widest possible access to census data and the pressures to recover some of the cost of data collection will lead to a compromise between the two extreme positions just described. For instance, special arrangements can be made between government agencies who want to incorporate each others data into their products. The census organization may enter into agreements with the national mapping organization to distribute digital base maps of roads, rivers, etc., to census GIS data users. Also, academic and other non-profit users can be granted discounts. Another option is to provide some generic products free of charge, while charging for value-added products that require more processing.

Liability issues

Courts have ruled in several instances that data producers can be held responsible if errors in geographical information lead to accidents or other damages. Most cases have so far dealt with accidents due to missing or erroneous information on topographic maps. Cases have been documented of plane crashes and accidents at sea caused by erroneous

information on navigational maps. Map design and information content is guided by their intended use, but maps are sometimes used for purposes not anticipated by the data producer. For instance, a census organization might publish data for reporting units together with a street network database. Because the road information is not critical to census data use, quality control of this information may have been much less rigorous than if the road information had been compiled for an emergency services routing system. If the imperfect data are used for such unintended purposes, damages may well occur.

Another example related to liability issues that is very relevant to census data dissemination is the violation of privacy of information. Usually, a census organization publishes only aggregate data at a level that does not reveal the information for an individual, a household or a very small group of persons. If the census organization re-aggregates the micro-data for several small area geographies—e.g., enumeration areas, postcode sectors, health or education districts—there is a possibility that clever GIS operations can isolate information for groups of persons smaller than the lowest disclosure level (see next section). In some countries this may be grounds for legal action by the concerned individuals.

Interestingly, Johnson and Onsrud (1995) argue that selling GIS data and restricting secondary uses of data may increase liability by a data provider. The fee would imply a guarantee by the data provider that the material is error-free and fit for the purposes intended. Placing data into the public domain, in contrast, may shield the agency from such claims.

Before distribution of spatially referenced data, the agency should thus consult with legal experts and draft a disclaimer that is accompanies the data products. The disclaimer may include the following points:

- A statement that the information was believed to be accurate at the time of collection and to have been obtained from reliable sources, but no warranty can be given as to the accuracy.
- Warnings that information is subject to change and notification of actual changes.
- If any parts of the geographic database were created by an external agency, this should be stated clearly.
- Mention that use of data implies acceptance of disclaimers and agreements.

Digital census atlases

While a more generic GIS database is targeted at users who have considerable experience in GIS, a digital census atlas is aimed at the general public, schools and other non-expert users. Two approaches for producing a digital census atlas are considered here. A *static* census atlas consists of a collection of maps and other materials that have been prepared by the census office. It is essentially a presentation in which the user can change the sequence of viewing the content, but cannot change the content itself. A *dynamic* census atlas, in contrast, combines a digital GIS database and census data in a simple mapping package. The user can use the data to produce custom maps which can be printed or copied into other applications packages.

Static census atlases

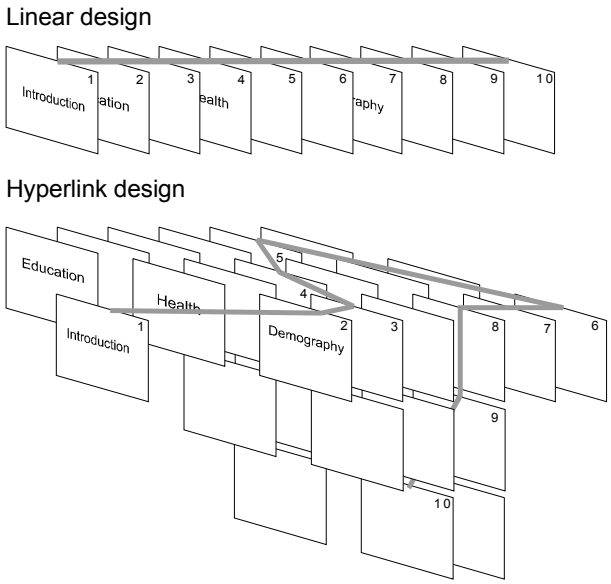
A static digital census atlas can bring together maps, tables, graphs and possibly multimedia products such as photographs or movie clips in a visually appealing, user-friendly environment. The presentation can be put together in a standard presentation software such as a Powerpoint slideshow. Some presentation graphics packages allow the developer to produce a stand-alone version of a graphics presentation which can be distributed together with free viewer software. Most presentations or graphics can also be exported to PDF format which can be distributed on computer-readable media or via the Internet. Maps can be produced in a desktop mapping package and incorporated into the presentation software using a graphical interchange format or simply the cut-and-paste commands in the Windows environment.

An alternative presentation platform is an Internet browser. Most computer users have an Internet browser on their computer which can be used to view files that reside locally on the computer as well as remote content. Maps and other graphical content can be included as graphics images in GIF or JPEG format, which can be produced from GIS map layouts. The presentation design might result in a *linear* presentation. The user is led through a series of maps and graphics that were arranged to reflect a consistent story line. This is appropriate for relatively short presentations. For the presentation of a larger number of maps, the viewers' patience might be taxed when they have to go through many slides with material that they are possibly not interested in.

Most presentation packages provide a better design option which is based on hyperlinks. These links allow the user to jump between different sections of the presentation. They

also allow to integrate additional sources and information that might only interest a small number of viewers. For instance, on a page that shows a map of a population projection for districts, links to a methodological paper explaining the projection's assumptions can be added.

Figure 5.8: Presentation design options for a static digital census atlas



The hyperlink concept is illustrated in Figure 5.8, where it is contrasted with the linear design approach. In the hyperlink design, several parallel topics are presented which are interconnected by links as appropriate. For instance, the three parallel story lines or chapters that follow the introduction page (1) could be on education, health and demographic indicators. The user might follow a path—indicated by the gray line—beginning with the demographic topic (2) where one of the slides (3) shows a map, tables and graphs of the proportion of population under fifteen. From here, links might be provided to maps showing child health indicators (4), educational facilities (5), and so on.

Using hyperlink oriented designs requires a very careful design of the presentation, since users are easily lost after following a number of links. It is important to include clear navigation tools on each page, including 'breadcrumbs,' which are links designed to allow the user to back out of the current page and find his or her way.

Hyperlinks are of course familiar to anyone who has used the World Wide Web. In fact, rather than using a presentation software package, a static census atlas can also be implemented in the standard Internet browser language HTML. Web page design tools give the developer considerable flexibility in the design of the census database. One tool that can make the presentation more interesting, for example, is a clickable map. For instance, the entry screen might show an overview map of the country with instructions to click on the province of interest for more detailed maps at the subnational level. Web technology also allows the inclusion of multimedia content and links to information outside of the presentation, for instance to other parts of the census office's web page or to other government agencies. These can, of course, only be accessed by users with Internet access.

One advantage of using web design tools is of course that the same static census atlas can be distributed on CD-ROM or diskette for stand-alone use, and it can be posted on the census office's web site for viewers anywhere in the world.

Dynamic census atlases

An alternative to a static census atlas is to publish a digital map and database together with mapping software that allow the user to produce custom maps of census indicators. This of course requires some knowledge of cartography on the user side. A dynamic

census atlas will include digital boundary files at a lower resolution than the full census database to allow fast drawing and low disk usage. The closely integrated attribute table should contain only a selected number of census indicators. Densities and ratios that are appropriate for mapping should already be calculated.

This approach will serve the needs of users who do not have the GIS expertise and skills required to make use of the complete digital census GIS database, but who want more flexibility in exploring and utilizing geographical census information than is possible with a pre-packaged static census atlas.

The problem, of course, is that such users may not have a desktop GIS package available that can be used to create maps. The data provider should therefore provide an easy-to-use package together with the boundaries and data. The use of that package should require minimal training and experience. Essentially, the application should be "plug-and-play"—after installation, the user should immediately be able to produce maps.

Some census offices have developed map viewing software in-house and distribute these with their census data products. The maintenance of such programs is expensive, however, and binds resources that could otherwise be spent on data development or dissemination. Some GIS vendors are now selling GIS software tool kits that can be put together to produce custom applications or to integrate GIS functions in other software products (e.g., spreadsheets or database applications).

As an alternative, there are now several mapping packages available that are free of charge and can be distributed with a database. See Chapter 2 for discussion.

Some commercial GIS vendors also make viewing software available free of charge and allow users to distribute these simple mapping systems freely in a database distribution package. An example is the ArcGIS Explorer package produced by ESRI, Inc. (Redlands, CA). ArcGIS Explorer is a mapping interface for data created in the Arc/Info and ArcView GIS packages.

The ArcGIS Explorer interface is very easy to use and the system provides basic mapping functions for producing thematic maps that can be exported as bitmaps or Windows Metafiles. ArcGIS Explorer can read data from the local hard disk or a CD-ROM. On computers with an Internet connection, it is also able to display data that resides on remote web sites. Analytical functions are limited, but the system does support different types of data query—interactive or using SQL-like commands—and address matching.

The documentation for a dynamic census atlas needs to include much of the same information that should accompany a more comprehensive census GIS database. However, the text should be designed with non-expert users in mind. Technical GIS jargon should be avoided. Since the users are unlikely to use the database for more advanced applications, the emphasis in the documentation should be put on the attribute information and less on the technical geographic details.

Internet mapping

Many national statistical organizations have embraced the Internet as a means to disseminate information and data. Web pages range from simple lists and tables of census results to sophisticated query interfaces in which the user can request special cross-tabulations.

The Internet is also suitable for presenting and distributing geographic information. The simplest option is to present static map images that were produced by the statistical office. For instance, a series of maps showing census variables can be produced using a desktop mapping package. Most packages allow the user to save maps in a standard image format such as GIF or JPEG. These images can then be integrated into web pages just like any other graphic or photo. Such web sites can give data users access to useful information. However, they do not allow the user to manipulate the data and to produce custom maps for specific geographic areas. The following sections concentrate on approaches that allow a significant degree of user interaction with the census geographic database.

Most GIS and desktop mapping software companies have developed platform independent tools for Internet mapping that make use of standard data exchange protocols. These tools enable the statistical organization to set up geographic information on a server and allow users to map and query these data interactively using standard Internet browsers. Internet users can thus access GIS applications without having to purchase proprietary GIS software. Any data that can be stored or manipulated with a GIS can be distributed in this way—including vector maps, raster images and data tables.

Internet mapping software is also useful as an in-house tool to make spatial data accessible to statistical office staff on an Intranet. Rather than purchasing site licenses of commercial GIS packages which are run from a central server, each staff member can access geographic information through their browser software.

There are three main options for implementing Internet mapping:

- In server side strategies the user sends a request for a map to the server holding the database. Mapping software on the server processes the request, produces a map—for example, in GIF format—and sends it back to the user.

- In client side strategies in contrast, most of the processing tasks are performed on the user's ("client's") computer locally.

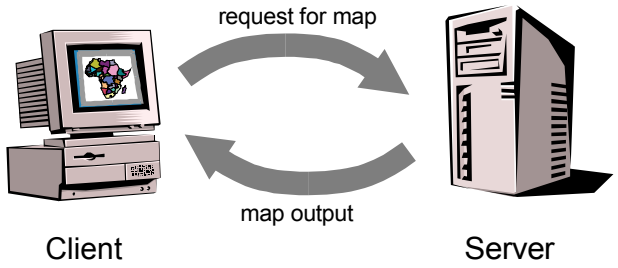
- Hybrid approaches, finally, combine server and client side approaches.

Server side approaches

Sometimes called "thin client/fat server" architecture, these strategies put most of the data processing load onto the server that is located at the data distributing organization. This is similar to traditional mainframe architecture where a powerful central computer handles data management, storage and processing for a number of users that are connected by dumb terminals.

The principle of a server side strategy is summarized in Figure 5.9. The user connects to a web site and enters a request for a map. User defined specifications for the output map include the geographic region of interest—which is specified either through the name of the region such as the district's name, or through coordinates that form a bounding rectangle—, the variable to be mapped, the classification and color scheme, and additional data layers that provide context such as roads, rivers, or administrative boundaries.

Figure 5.9: Internet mapping – the server side approach



The user's request is sent through the Internet to the server and routed to a GIS package. The GIS software can either be located on the web server, or it could reside on a separate computer connected to the server. The GIS package could be a commercial Internet mapping package, or a tailor-made Internet mapping package that is based on commercially sold mapping software modules. The map software accesses the required databases, produces the map and sends this output back to the user as a web page. Maps are usually sent as standard graphics images in GIF or JPEG format, since web browsers cannot handle vector data formats. If the user wants to modify the map design, a new request is sent to the server.

The server side approach has the advantage that the user does not need a powerful computer to access possibly very large spatial databases. Even fairly complex GIS procedures such as address matching or network routing can be carried out quickly if a powerful server is available. All that is required by the user is a basic browser and an Internet connection. Data integrity is maintained since the user cannot manipulate the database itself. The user is also always assured access to the most recent information. The data provider has more control over what users can see and how they can see it. Cartographic design choices can be pre-set to ensure that even non-expert users will

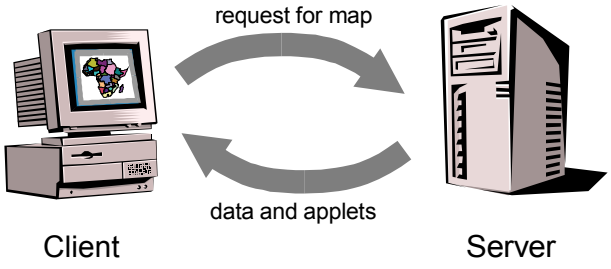
obtain acceptable map output. One disadvantage is the network traffic load on busy servers.

Client side approaches

Client side approaches—thick client architectures—transfer much of the required processing to the user's computer. The server is mainly used to hold the database and send required pieces of the database possibly together with mapping modules to the user. Two variations of the client side approach are available.

In the first, no mapping capability resides on the user's computer. After the user's request has been submitted, the server sends the geographic data as well as a small program or applet that enables mapping or geographic analysis (Figure 5.10). An applet is a platform independent piece of software written in the Java programming language which can be executed by standard web browsers. The user can then work with the data independently from the server. Browsing the map layers or changing the cartographic design does not require new requests to the server.

Figure 5.10: Internet mapping – the client side approach



In an alternative client side approach, a mapping package, applet or browser plug-in resides permanently on the user's computer. A plug-in is a program that extends the

Internet browsers capability, for example, to enable it to display files of a certain format. The advantage of this approach is that the mapping software does not need to be downloaded every time the user accesses the map server.

After data and programs have been downloaded, the user does not need to communicate further with the map server. Mapping or analysis can be carried out off-line. The user's computer resources can be utilized, usually resulting in faster processing. Client side approaches can give the user more flexibility and freedom in the analysis and display of spatial data. However, data and program files may be very large, requiring a fast Internet connection, users with less powerful computers may not be able to execute more complex mapping and analysis tasks. Client-side approaches may allow the user to save the raw geographic data that is requested from the server on their computer. This is a problem if some or all of the geographic data on the statistical agency's server are copyrighted.

Hybrid approaches

Server side approaches are good at providing access to relatively simple maps to a large, non-expert audience. They would thus be most suitable for a census office's presentation of census maps to the general public. Client side strategies, on the other hand, are preferable for Intranets where a smaller number of users with relatively comprehensive knowledge of GIS and mapping access complex databases. They would thus be suitable for in-house GIS data access for census office staff.

Hybrid approaches combine the advantages of client and server based strategies. They provide flexibility to the user in querying and manipulating maps locally, but transfer most of the processing load in demanding analysis tasks to the server. This requires some

degree of communication between client and server concerning the available processing power.

Opportunities for census data distribution; MapServer

Currently available Internet mapping packages are scalable. Data providers can purchase an off-the-shelf package that works with standard data sets. Since mapping of census data is a fairly standard application, national statistical offices should have no difficulties finding a suitable solution. For more complex applications, a tool box of software modules can be obtained which allows the data provider to custom-design the map server interface.

The MapServer concept was conceived as an open source development environment for building spatially-enabled internet applications. Mapserver is the most used Open Source web-based

server technology available today and one of the core components needed to translate GIS data into a map image to be viewed by a web client. Indeed, MapServer supports Open GIS Consortium (OGC) standards, and it features advanced cartographic output including thematic mapping and map element automation, scale-dependent feature drawing and feature labeling, customizable, template-driven output, support for popular scripting and development environments, map projection support, and a multitude of raster and vector formats. For more information, see http://mapserver.gis.umn.edu

With increased network capacities, larger data sets and program modules can be transferred to users, and more users can be served simultaneously. The problems inherent in both client and server side solutions can be overcome with faster Internet connections.

Client computers can have frequent communication with servers without delays leading to near instantaneous execution of user requests.

For census data the best Internet data access and distribution strategy will depend on the capabilities and expertise of the user. A flexible system will provide services for any level of user:

- "Power users" who want to obtain the entire database for use on their own computer using commercial GIS software. These users are served by conventional data distribution methods such as purchase of CD/ROMs or Internet download options of "raw" census GIS data sets.

- Active users with some expertise in GIS but who do not have local GIS capabilities. These users want to download parts of the database together with GIS program modules (applets) that can perform the required tasks.

- Passive users who simply want to obtain a pre-designed map. The user request is executed by the server and the resulting information is sent to the user through the Internet in a suitable format—e.g., raster image or postscript files for maps and spreadsheet or database files for the data.

A flexible census data distribution system on the Internet could look like this:

- Users determine the geographic extent of the region of interest. This could be to download the data or to simply request a map. The geographic region of interest can be specified using any of the following geographic addresses:

- The name of the geographic region such as a city, district or province name.

- A bounding rectangle determined by geographic coordinates.

- A region that is interactively specified by a user through browsing and zoom functions. For instance, the interface may start with a map of the country. The user can then zoom into a region of interest and select the specific geographic area by drawing a rectangle or polygon on the screen. As the user zooms in, more detail is shown on the map interface. At the start the map shows only country and province boundaries. As the user zooms into one province, district boundaries and town locations appear. Selecting a specific town will show major streets and urban enumeration boundaries. Which level of detail is shown is determined by the map scale that corresponds to the current map extent on the user's screen.

- A region that is defined by a geographic query. For instance, a commercial user who wants information about demographic characteristics of potential customers could request demographic information for a circular area of five km radius surrounding a shopping center location. A government planning agency may request data on the population living within five km of a proposed highway corridor.

- The user specifies the variables of interest and the type of output desired. Options may include maps for which the user can specify basic cartographic designs such as the number of categories, type of classification and shade colors. Or, the output could be a simple data table showing the selected variables for the region of interest. The

user also specifies whether a database and geographic query and analysis modules are required, or whether a map or database result is desired.

- The database server interprets the user's request and creates the appropriate subset of the database. For regions specified using geographic names, this will simply involve a logical selection of, for example, all census enumeration areas within a given district. For requested areas which do not match the standard census geographic hierarchy, some further processing is required. In some countries, dwelling unit GIS databases are now available or under construction, where every residence is associated with a geographic coordinate. A GIS on the server can then compile a custom tabulation by selecting all households that fall into the user defined geographic area. In cases where this is not possible, the server based GIS needs to perform an areal interpolation using techniques such as those described in an earlier section of this chapter.

- The result of the query is returned to the user either as base data that can be manipulated further by the user using GIS applets, or as a map or database report that can be used directly by the user. Of course, in addition to the database or maps, data documentation and other relevant information must be available as well.

Depending on the data distribution policies of the country, these services could be free or fee-based. While requests for basic information that has been compiled already could be provided free of charge, more complex requests could be fee-based.

An important consideration if the custom-tabulations are based on micro-data is data privacy. Internet security issues are as significant in the management of census data on networks as they are in commercial Internet applications. The internal network which

may provide access to census microdata must therefore be separated by a firewall from the Internet domain that allows external users access to aggregate census data.

Obviously, the envisioned data distribution interface is very ambitious. It requires fast Internet connections and can only reach a large number of users if Internet access in private households, businesses and government agencies is widespread. In many countries these conditions are not yet present, but given the rapid spread of technology, many countries will be in a position to satisfy the majority of data requests via the Internet in the near future. Some census organizations are actively pursuing data distribution strategies that include elements described here.

textbox: Canada experience with web mapping for dissemination of census data – adapt from EGM paper ESA/STAT/AC.115/19

Summary and conclusions

Chapter 5 has examined the use of geodatabases in the postcensal phase. Sections have included a list of electronic products from the census, the aggregation of census collection units into dissemination units to display results, disclosure and privacy issues (relevant for the release of small-area data) and the option for a national statistical agency to raise revenues selling value-added products such as data CD-ROMs or DVDs. It also includes geographic data products like map viewers, attributed spatial files for use in

commercial GIS packages and internet mapping products. Finally, the chapter addresses data maintenance issues for a continuously updated census geo-database and provides arguments for continuous or updated measurement in the field. The annex section, providing background information on GIS, R/S, GPS and some mapping and geospatial principles, is updated from the 2000 edition, with new sections added on spatial analysis and where to go for assistance.