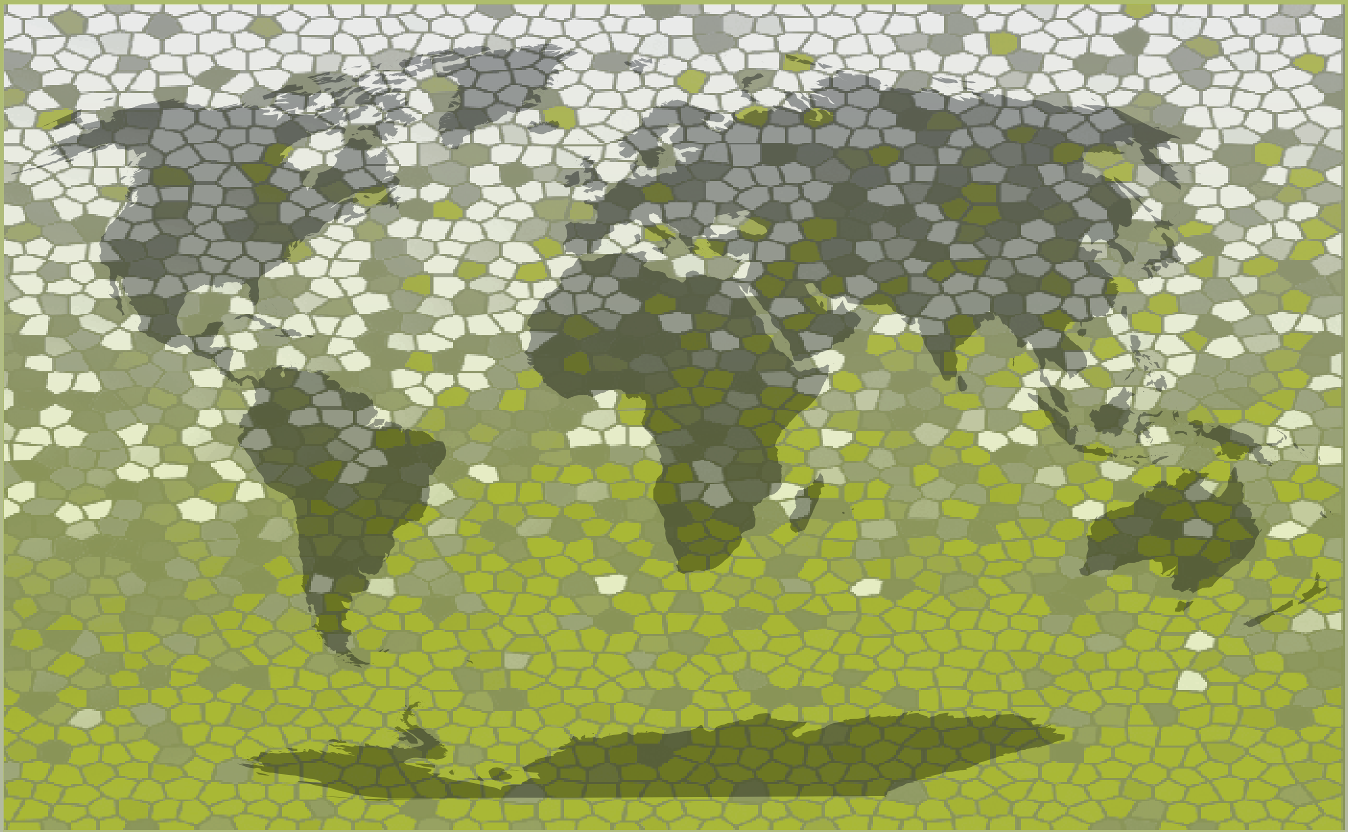
Economic & Social Affairs

Handbook on Geospatial Infrastructure in Support of Census Activities



United Nations

ST/ESA/STAT/SER.F/103

Department of Economic and Social Affairs
Statistics Division

Studies in Methods Series F No. 103

Handbook on Geospatial Infrastructure in Support of Census Activities

United Nations
New York, 2009

Department of Economic and Social Affairs

The Department of Economic and Social Affairs of the United Nations Secretariat is a vital interface between global policies in the economic, social and environmental spheres and national action. The Department works in three main interlinked areas: (i) it compiles, generates and analyses a wide range of economic, social and environmental data and information on which States Members of the United Nations draw to review common problems and to take stock of policy options; (ii) it facilitates the negotiations of Member States in many intergovernmental bodies on joint courses of action to address ongoing or emerging global challenges; and (iii) it advises interested Governments on the ways and means of translating policy frameworks developed in United Nations conferences and summits into programmes at the country level and, through technical assistance, helps build national capacities.

Note

The designations used and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The term "country" as used in this publication also refers, as appropriate, to territories or areas.

The designations "developed regions" and "developing regions" are intended for statistical convenience and do not necessarily express a judgment about the stage reached by a particular country or area in the development process.

Symbols of United Nations documents are composed of capital letters combined with figures. Mention of such a symbol indicates a reference to a United Nations document.

ST/ESA/STAT/SER.F/103

UNITED NATIONS PUBLICATION
Sales No. E.09.XVIII.8

ISBN: 978-92-1-161527-2

Printed in United Nations, New York

Preface

The United Nations published the *Handbook on Geographic Information Systems and Digital Mapping* for use during the 2000 round of population and housing censuses. The 2000 *Handbook* has provided useful guidance in the field of census cartography. It needs updating and reviewing, however, to take into account recent developments in geospatial technologies and their applications for statistical exercises, in particular for population and housing censuses.

For the 2010 World Programme on Population and Housing Censuses, the Statistical Commission, at its thirty-sixth session, requested that the United Nations Statistics Division proceed with its work on the revision and update of Principles and Recommendations for Population and Housing Censuses. The Commission requested that the Statistics Division address some related specific issues, including the application of geographic information systems (GIS) in both the collection and dissemination of data. These developments have been reflected in the *Principles and Recommendations for Population and Housing Censuses, Revision 2*, which were adopted by the Statistical Commission in March 2007.

Subsequent to the adoption of the *Principles and Recommendations, Revision 2*, the United Nations Statistics Division initiated a series of activities to promote and support the 2010 World Programme on Population and Housing Censuses and review the 2000 *Handbook*, taking into account recent geospatial technological advances. In May 2007, the Statistics Division organized an expert group meeting in New York on contemporary practices in census-mapping and the use of geographical information systems. The meeting was held to provide input into the revision of the 2000 *Handbook*. The Statistics Division also conducted, in the fourth quarter of 2007 and early in 2008, five workshops on the use of geospatial technologies in census-mapping operations. Two workshops were held in Africa for English-speaking and French-speaking countries, respectively, while one workshop each was held for the Asian, Caribbean and Pacific island regions. The Statistics Division hired a consultant, David Rain, to assist in the preparation of the draft revised *Handbook*. In April 2008, a second expert group meeting was organized in New York to review the draft revised *Handbook*.

The newly revised and renamed *Handbook on Geospatial Infrastructure in Support of Census Activities* reflects the recommendations of the expert group meetings and regional workshops on GIS and census-mapping. Those bodies emphasized the need for countries to approach the use of census geography programmes as a continuous process rather than merely a sequence of mapping and dissemination operations. It was emphasized that the *Handbook* should demonstrate how the use and application of contemporary geospatial technologies and geographical databases are beneficial at all the stages of the population and housing census process. For instance, the *Handbook* should show how those technologies improve efficiency in the preparatory, enumeration, processing and dissemination phases of the census.

In this regard, it is important that the *Handbook* put into the hands of census planners and related personnel a technical guide on the contemporary methods, tools and best practices that would enable them to better articulate their needs and deal with census-mapping operations more efficiently. In short, the *Handbook* covers both managerial and operational needs in considerable detail. It addresses organizational and institutional issues that concern statistical agency heads and other managers; and it explicitly addresses technical and practical issues that concern census cartographers and takers.

The present *Handbook* is divided into six chapters and seven annexes. After a brief introduction in chapter I, chapter II covers managerial issues for statistical agency heads that should be considered when reorganizing national statistical offices to permit the full use of geospatial infrastructure. Chapter III provides technical content for the data-processing manager or cartography/GIS chief to use for the practical establishment of a digital enumeration area (EA)-level geographic database. Chapter IV continues the technical focus, spelling out the process of constructing an EA geodatabase and using such technical advances as global positioning systems (GPS) and remote sensing to make corrections, integrating them where necessary with ground-based work. Chapter V covers the process of creating the maps needed for enumeration, with an operational focus that picks up where the geographic database discussion left off. Chapter VI covers the use of geospatial infrastructure for dissemination of census results. Annexes I to VII provide a handy reference for those planning and implementing geospatial solutions to census projects.

During the revision process, the United Nations Secretariat consulted census cartographers and GIS experts representing all regions of the world to review and finalize the *Handbook*. The *Handbook* also contains some examples of country practices in the application of GIS, GPS and digital mapping used in censuses contributed by some of these experts. The *Handbook* was prepared by David Rain, a consultant for the United Nations Statistics Division, in collaboration with the Statistics Division team.

Contents

Page

Page

Boxes

Figures

Page

Tables

Chapter I
Introduction

A. Foreword and rationale for the Handbook

1.1. The *Handbook on Geospatial Infrastructure in Support of Census Activities* seeks to recognize and build on the achievements of its predecessor, the *Handbook on Geographic Information Systems and Digital Mapping*, published in 2000.

1.2. Major technological advances include the widespread availability of personal computers, hand-held computers, global positioning systems (GPS), geographic information systems (GIS) software and low-cost aerial and satellite imagery. These advances have put new tools in the hands of national statistical organizations (NSOs) to collect more accurate, timely and unbiased information about their populations. The emergence of new technologies is indeed a driving force behind the new *Handbook*. At the same time, it is recognized that adopting such new methods will challenge the leadership of NSOs and bring about changes in their organization. The implementation of new geospatial capabilities is taking place throughout the census programme.

1.3. The *Handbook* argues that some of the biggest challenges for NSOs are not only technical but also organizational, institutional and managerial. Most member countries have begun to make use of geographic information science and technology appropriate to the scope of their programme needs. One purpose of the *Handbook* is to respond to those needs. The legal or constitutional mandate for conducting a national census of population in many countries has remained in effect. At the same time, changes include the increased use of census data for disaster management and many other purposes (for a more complete listing, see Chap. II). Other changes are the increased availability of data in a variety of useful formats; and new technology to make data collection, analysis and storage easier than before.

1.4. In the production of data for these uses, countries are discovering that they can leverage the strengths of other government agencies through what it is referred to as a National Spatial Data Infrastructure (NSDI).[1] An NSDI is an institutional arrangement to permit data-sharing and collaboration across government at a variety of levels, including the national, regional and local levels. Additional 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. The principle "Create once, use many times" applies to the construction of a national census geospatial database, because once it is created it can be used in different national contexts for many purposes.

1.5. The present *Handbook* provides constructive options for the reorganization of census-mapping and analytical tasks around digital geographically referenced databases. Structuring the statistical organization around such "geodatabases" calls for careful planning, not only because of upfront investment required for GIS but also for the work required to develop the capacity to analyse census data and deliver products

1 An NSDI is a combination of the technology, policies, standards and human resources necessary to acquire, process, store, distribute and improve the utilization of geospatial data. Conceptual parts of NSDI comprise (*a*) an institutional framework that defines the policies, legal and administrative support to create, maintain and apply the standards to fundamental data sets; (*b*) standards that define technical characteristics to fundamental data sets;(*c*) fundamental data sets that require geodetic framework, topographic and cadastre databases; (*d*) a technological framework that allows users to identify and receive the access to fundamental data sets (see *GSDI Cookbook*, 2000).

to the public in a timely way. The success of GIS as an industry is built on the power of geospatial information to enable problem-solving and decision support.

1.6. Developing GIS capacity may mean altering the traditional organizational chart of an NSO, and specifically 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 reorganization may require a continuously funded operation, staffed with expertise to help the NSO perform its tasks decade-round. A dedicated geographic core will require a team with not only geospatial skills but also an operational goal of modernizing census-taking. Trained personnel will follow a strict timetable that ensures that 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 management, 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.

1.7. Planning for the operational implementation of a geographic programme cannot happen soon enough. Due to the 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 be needed for enumerators and crew leaders. Producing detailed geographic data at a fine enough scale to use with other data layers in a GIS, with dissemination units composed of enumeration areas or population clusters, 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 require organizational restructuring. 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 NSDI. Indeed, for NSDI efforts, the contribution of NSOs is likely to be:

(*a*) A spatial geographic database, with polygonal and attribute information for the enumeration areas of the country (i.e., the units for which the territory is allocated to canvassers during the census). A common digital base can assist with censuses of agriculture and population, as recent country experiences suggest. Census data can be released at the EA level or aggregated into new small-area dissemination units, such as population clusters;

(*b*) 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 by avoiding the re-surveying of territory for other purposes, such as electoral redistricting;

(*c*) A national gazetteer, including geographic names and coordinates of population settlements (referred to as "P-Codes" within the United Nations humanitarian community). When paired with population estimates, this can be used by humanitarian organizations for development and emergency response purposes. Additional data on housing units in vector format can be used in a similar way.

1.8. For geographic and demographic data in GIS format to be shared among other organizations, NSOs will follow standards for geographic referencing and metadata established at the national level. Special consideration needs to be given to the administrative classification system to be used to organize territory for the census. The present *Handbook* will refer to this as "geo-coding", since it serves as a linking point for demographic information and its location on the Earth's surface.

1.9. Above all, the *Handbook* highlights the need to develop realistic plans in order to harness the power of GIS and other geospatial technologies for modernizing census operations and obtaining better results and analysis. The need to extend the scope of census cartography to the national spatial framework of the country is also emphasized.

B. Scope, purpose and outline of the *Handbook*

1.10. 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 motivations for the present *Handbook*. Any country embarking on a census project will need to plan extensively to minimize the costs and maximize the benefits from the required geospatial activities. Besides informing key supporters of geographic initiatives within NSOs, the present *Handbook* aims to provide technical and methodological information to support the choice of a suitable set of tools and procedures for a given country.

1.11. 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. The present *Handbook* is therefore designed as a practical reference document, a "cookbook" that illustrates the role of geospatial technology in each step of the census process. Each country needs to evaluate how available mapping options fit into the context of its own census programme and national planning. Such issues as existing geographic resources in the country, 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.

1.12. The present *Handbook* is not a general-purpose GIS manual. It does not contain the commands and routines needed to operate specific software, nor does it make recommendations for specific software. It is also not a general guide for conducting a census. The *Handbook* respects the grand tradition of census cartography and argues that traditional analogue mapping techniques that have been used successfully in many countries are still relevant. One main reference on the topic, the United States Census Bureau's 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 programme, enumeration area delineation and statistical areas continue to be relevant, and the *Handbook* makes extensive use of material from its predecessor. As technology has progressed, however, there are now better ways of doing many of census- mapping tasks. The *Handbook* therefore aims to complement the earlier guidelines by providing information on recent technologies.

1.13. The present *Handbook* targets two main audiences, a managerial one concerned with the costs and benefits of investment in geospatial technology and a technical one tasked with implementing the geographic aspects of a census plan. The five chapters of the *Handbook* assume a basic knowledge of GIS and cartographic concepts. For readers less familiar with these subjects, annexes I and II 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.

C. Chapter-by-chapter summaries

1.14. Chapters II to VI of the present *Handbook* cover areas to which geospatial technology can make contributions before, during and after the census enumeration. The revised organization of the *Handbook* reflects the operational decision in many member countries to bring about a transition to digital census operations.

1.15. Chapter II covers managerial considerations for statistical agency heads and presents some issues to consider when reorganizing the NSO to permit the full use of geospatial infrastructure. Planning to answer the new demand for disaggregated statistical data, including data for disaster management, calls for working backward from the delivery of new products to the requirements needed at every step along the way. Efficiency arguments, including cost-benefit analysis, are covered to underscore the role of geospatial technology as a strategic investment. Critical success factors for geospatial implementation are covered next, followed by a discussion of the planning process in detail. This includes conducting a needs assessment; the determination of products and options; staffing and other human resources considerations; and institutional cooperation with other data producers and users through NSDIs.

1.16. Chapter III provides technical content for the data-processing manager or cartography/GIS chief to use for the practical establishment of a digital geographic database. It covers the step-by-step process of constructing an EA-level database, beginning with the importance of geographic coding, administrative hierarchies and census database components. EA delineation issues are covered next, followed by discussion of existing geographic data sources for EA delineation, including both digital and analogue sources. Digital input through scanning and digitizing is covered next, followed by a discussion of map integration, topology construction and georeferencing. The construction of a geographic database may raise new data quality issues associated with accuracy and precision, and quality assurance procedures. The chapter ends with a detailed section on the mechanics of metadata, stressing the importance of documentation and suggesting strategies for both internal and external data users.

1.17. Chapter IV continues the technical focus, spelling out the process of constructing an EA-level geodatabase and using technical advances, such as GPS and remote sensing, to make corrections, integrating them where necessary with ground-based work. Employed here is a triage approach to direct emphasis on areas that have changed since the prior census. The basics of GPS are explained, along with new technical improvements of differential GPS, training requirements and the use of handheld computers. Covered next is a section of satellite remote sensing (R/S), which lists available R/S sources along with visualization tools, such as Google Earth, followed by a section on the uses of aerial photography for census work.

1.18. Chapter V covers the process of creating the maps needed for enumeration, with an operational focus that picks up where the geographic database discussion left off. It includes sections on quality assurance, the production of maps for field use, map elements and design, map printing and distribution, and the use of maps for census logistics basics. It stresses that a project management approach must include plans for contingencies in the event of setbacks.

1.19. Chapter VI covers the use of geospatial infrastructure for the dissemination of census results. It employs a similar approach as in earlier chapters to the post-census phase, but builds on the notion of the geographic database (geocoded to specific locations) as the central data interface for the publication of census products. Topics covered include the aggregation of collection units to dissemination units, geographic database maintenance and archiving, and planning census products. A section on spatial analysis provides examples and options for interpreting census data visually. Additional issues covered include 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. Also covered are geographic data products, such as map viewers, attributed spatial files for use in commercial GIS packages or online products.

1.20. Annexes provide a handy reference to those planning and implementing geospatial solutions to census projects. Annex I provides basic information on GIS data models and levels of accuracy and precision. Annex II introduces coordinate systems and map projections, paying particular attention to issues involved with data input to a GIS. Annex III covers data-modelling and includes an example template. Annex IV provides an example of a data dictionary. Annex V suggests some pointers in thematic map design for census data analysis and display, including conventional choropleth maps. Annex VI contains a glossary of common GIS terms. Annex VII contains a listing of contact information and useful URLs for gathering more information.

Chapter II
Managerial considerations for heads of national statistical offices and other decision makers

A. Introduction

2.1. This chapter is intended for managers of national statistical organizations (NSOs), census directors and geographic section heads. It covers mainly institutional issues (i.e., the non-GIS technical content), focusing on various considerations associated with the use of geospatial technologies. NSOs can produce geo-referenced data with a higher degree of accuracy in less time, but only if activities are carefully planned. Included are examples of country experiences illustrating the utility of geospatial technologies for census work.

2.2. Planning for a census that makes use of geospatial functionality may require some reorganization of the NSO, putting geography — including cartography and GIS but also such topics as geographic coding and administrative boundaries — at the centre. Since there is no one technology solution that will allow every NSO to modernize its census the same way, the *Handbook* presents a range of choices that allow the NSO to adopt new technologies at the appropriate level of skill and experience. 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.

2.3. Mandated by national Governments for producing accurate population counts, censuses are also 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. It promises measurable productivity gains in almost any industry, using the central organizing principle of geography that place matters. Organizing information geographically through ordered display of data centred around location can effectively tie social observations to their location. As a data-collection effort, a geographically referenced census provides a matchless opportunity for a country to learn detailed information about its population.

2.4. The role of GIS and geospatial technologies in general, including remotely sensed satellite imagery and GPS, in governmental operations is changing. Increasingly, countries see the merit in organizing government information around a spatial model that implicitly incorporates geography. The use of geographic databases promises and has in many cases delivered gains in such important areas as efficiency and customer service, when government agencies communicate results and inquiries among themselves using geospatial tools.

2.5. Many studies of "e-government" have illustrated the benefits of technology adoption. O'Looney (2002) highlights the increased permeability of boundaries between government and business, civic groups and individual citizens. Using information technology to deliver services brings about a higher level of interaction and can be done with a relatively small staff. Garson (2003) stresses the higher level of accountability that accompanies this, especially when information users can check the arithmetic and challenge results. Khosrow-Pour (2005) addresses the new systemic administrative challenges requiring greater engagement at senior levels; "24/7 government" has clear implications for product and service delivery, inviting more citizen/client self-service capabilities. White (2007) highlights the need to establish transparent in-house policies governing Internet use, and also urges senior decision makers to make better use of contractors to achieve near-term human resource goals with lessened long-term impact.

2.6. Because of the large initial investment in terms of both hardware and software and reorganization, the use of geospatial technologies in census operations needs to be carefully planned. GIS industry-leader David Rhind of the 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 must be weighed carefully. Roger Tomlinson of Canada maintains that good planning leads to success, whether for GIS or any other technology implementation.

2.7. A census is an elaborate national undertaking that will influence policy for years to come. NSO heads and other managers are urged to 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 planning and data needs, they should decide which conventional census products — including tabulations, totals, age and sex break-downs — will need to be released. They also ought to plan new offerings, such as atlases, DVDs with disaggregated data or small-area data or micro-data, as well as electronic maps that could meet the needs of a host of new data users, thus increasing the overall level of customer satisfaction. Using the power of geography, such new capabilities are within reach. As Tomlinson also points out, GIS is a particularly horizontal technology, with wide-ranging applications across the industrial and intellectual landscape. This technology should be adopted to address specific needs. The NSO's strategic purpose should be considered. Most likely the main goal is doing an accurate census, on time and under budget. The specific ministerial objectives or mandates placed on the organization should be clarified, as well as how they will affect census plans. Tomlinson asks organizations to consider how their strategic purposes can be improved by using new technologies, such as GPS and geodatabases. He recommends doing extensive cost-benefit analyses of technology adoption over the specified planning horizon of the census. Managers are urged to plan their agency's products after talking with users and developing an information product description that can then be used to meet requirements. They should then create a data design, using the requirements identified; choose a logical data model for the planned data and determine system requirements; then finally plan implementation.

2.8. In the realm of the census, 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 centring operations around a data hub or information core that stores and serves data within the organization.

2.9. A "geocentric census" means organizing the census process around the geography. For many NSOs on the cusp of fully embracing digital capabilities, often the investment is in analogue-to-digital conversion of paper enumeration area maps, which involves careful scanning and correction procedures so that EAs 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 then field-corrected using GPS. Incorporating such technology 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.

2.10. During the enumeration, geospatial technologies can help plan logistics and give decision makers updates on progress. After the count is completed, technologies can allow greater access to society through information dissemination, especially using the Internet. A "geospatial one-stop" approach can help to coordinate the delivery of census products to thousands of new data users and empower existing users to request new products and services.

2.11. For any new technical undertaking, a key issue is building capacity. Many NSO managers do not think they have the budget or the institutional capacity to reorganize their agencies. This should be considered less a question of budget than of the scale of the planning horizon. Managers should think ahead five or even ten years; the NSO will need a far finer level of detail in its information products than can be currently provided.

Box II.1

Four country case studies

1. Namibia

Namibia began its digital mapping and GIS programme in preparation for its 2001 census, with its main aim being the efficient production of base maps needed for fieldwork. GIS infrastructure was put in place with the help 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 kilometres (km^2), 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 at that time 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. For its 2011 census, Namibia will establish a web-based GIS using Postgress open-source software 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 (for further information, contact Ottilie Mwazi at: omwazi@npc.gov.na).

2. Bhutan

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 km^2 and a total population of approximately 2.3 million. All structures, 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 subnational administrative units. For data dissemination, NBS set up a web server and made a set of administrative boundaries available. 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 (for further information, contact Thinley Jyamtshow Wangdi at: thinly_j@yahoo.com)

3. Saint Lucia

Saint 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 and the Ministry of Agriculture, and on the private sector side with Cable and Wireless and St. Lucia Electricity Services, the CSO digitized enumeration district maps and used high-end GPS to capture spatial data from the field. 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 Saint Lucia to be defined. Settlements were geocoded using a nine-digit number. During the build-up to the conversion to digital-mapping operations, Saint 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 (for further information, contact Sherma Lawrence at: sherma_l@slucia.com).

4. Brazil

In 2007, the Brazilian Institute for Geography and Statistics (IBGE) led a census operation encompassing three work fronts: the Census of Agriculture, the Population Count and the National Register of Addresses for Statistical Purposes. These work fronts were performed concurrently and in an integrated way. The Territorial Database that supported this operation is composed of a set of maps and record files (cadasters) that depict the division of the Brazilian territory in small geographic areas or "enumeration areas".

The Territorial Database itself holds 249,068 enumeration areas in total. Of these, 162,770 were visited during the 2007 census operation, 70,085 of which corresponded to rural lands and 92,685 to urban areas.

IBGE has 27 state offices, whose headquarters are located in the capital cities of each state. The state offices coordinate the activities of 530 local offices. Each local office covers a group of municipalities. In this way, all of the 5,564 Brazilian municipalities can be covered in the census operations. Every state office headquarters has a territorial database sector, responsible for the census-mapping of the corresponding state. Moreover, in six of the state offices there exist geodesy and cartography managerial units that operate regionally throughout the country, providing support to the territorial database sectors for the production of digital (statistical) maps of the municipalities.

That was the infrastructure which supported the creation of the enumeration area maps in 2007. Existing paper maps were updated in the 530 IBGE local offices and digitally processed in the territorial database sectors with the support of the aforementioned GGCs. All of this was centrally coordinated by two managerial units under the IBGE Directorate of Geosciences, in charge of urban and rural census-mapping activities, respectively. After this stage, the maps were sent for revision and approval to the census commissions of the municipalities, whose members are representatives of the corresponding local government and society.

In order to ensure the uniformity of the Territorial Database updating process across the country, computer network automated systems were used, along with detailed operational manuals. Regional and local training programmes were organized. Supervision activities were developed by the technical coordinators. The project development was monitored following the established time schedule through a production accompanying and control system .

The Territorial Database construction for the 2007 census operation was quite an endeavour. This labour-intensive job demanded a thorough inventory of available maps in both cadastral and topographical scales and other ancillary documents; the establishment of partnerships with third parties for documents exchange and the consolidation of territorial information; the updating and digitalization of cartographic documents; the generation of localities, municipalities and cadastral maps; the creation of digital boundaries files of the enumeration areas; and maps for use in personal digital assistants (PDAs) and to support the publishing of census results.

The enumeration area map production for the municipalities/rural lands relies on systematic topographic maps produced by and available at IBGE itself and also at the Geographic Service Directorate of the Brazilian Army. From these topographic maps taken as input data, the production is carried out through the Semi-Automatic Municipal Maps Elaboration System (SisCart), specially developed for IBGE in Visual Basic 6.0 on top of the Bentley/Intergraph´s MicroStation/MGE graphic platform and the alphanumerical data management product Access97, from Microsoft.

SisCart significantly facilitated the construction of the municipal maps. Operating in a decentralized way, SisCart provides functionality in support of various key tasks, including the homogenization of map projection and scale; geocoding of the topographic map sheets comprising the municipal map; edge-matching between joining sheets; cropping of sheets following the municipalities perimeter; and editing of framework and footnote data.

The enumeration area map production for the urban areas relies on cadastral maps, whose scales range from 1:2.000 to 1:10.000. These maps are produced by state and local governmental agencies, utilities companies and other sources. The system used for the urban maps production is based on the MicroStation platform (fully customized in order to meet IBGE requirements). The system can handle all the transformations necessary to ingest map data from the most varied sources and coordinate systems. It can also provide support to the map-updating tasks that get input data from both fieldwork and office activities.

One of the main innovations introduced in Brazil's 2007 census operation was the National Register of Addresses for Statistical Purposes, originally developed from the records of the 2000 census enumeration areas and updated through fieldwork in the 2007 census. The technological innovations introduced in 2007, in particular the use of PDAs integrated with GPS, enabled the creation of new Territorial Database products by IBGE, such as:

- Maps of 70,085 rural enumeration areas and 92,685 urban enumeration areas in PDF format.
- Rural and urban enumeration areas description in PDF format.
- Maps of rural enumeration areas in JPG format.
- 70,085 Google Earth geocoded images.
- Municipal/enumeration areas digital boundaries in the shapefile vector format, encompassing urban perimeters and the isolated urban areas in all 27 federation units, with about 77,000 polygons.

All the material related to the Territorial Database reached the enumerators at the local offices and collection stations, points of convergence in the beginning and end of field data-collection activities. The maps of the 2007 census enumeration areas were delivered digitally through the PDAs and rendered on the enumerator's notebooks, along with the perimeters´ descriptions of the enumeration areas.

Historically, the handling of such a great volume of information originated from the census collection has been a point of concern for the Territorial Database sectors. With the adoption of PDAs, the field data to be displayed on the maps could be transmitted and stored on the IBGE servers for further treatment, analysis and inclusion in new map releases, thus offloading the Territorial Database sectors. During the collection, modifications in the printed maps were pointed out by the enumerators, who also contributed to the process by analysing the pertinence of the updates to be added by SisCart.

The operation costs of the Territorial Database updating in the 2007 census, including its participation in the total budget of the census operation, is detailed in the following table.

The recruitment of temporary personnel was carried out by public procurement, organized by outsourcing service bureaus and following the legislation for hiring personnel under the Federal Government. In 2004, 500 survey and mapping assistants, who worked in IBGE local offices in the Territorial Database sectors and in the DGC's headquarters were hired to execute several stages of the Territorial Database updating process.

Approximately 68,000 enumerators and 18,000 census supervisors were hired in 2007 for data collection. All enumerators received field training and classes on concepts of the Territorial Database, on the relevant characteristics of each enumeration area and on the procedures for Territorial Database revision (for further information, contact: Rafael Castaneda at: rafael.march@ibge.gov.br).

Costs of Territorial Database updating project					
	Database updating	2007 census	Database updating	2007 census	% for database updating
Year	(R$)		US$a		
2004	10 774 885.19	—	3 468 273.47	—	
2006	4 041 240.44	179 200 904.00	1 867 917.93	82 829 167.55	
2007	834 380.58	428 919 454.00	433 354.41	222 769 011.11	
2008	—	1 457 000.00	—	835 292.09	
Total	15 650 506.21	609 577 358.00	5 769 545.81	306 433 470.76	1.88%

[a] Reference values: Brazilian real (R$) to US dollar exchange rate 30 June of each year: 2008, US$ 1.7443; 2007, US$ 1.9254; 2006, US$ 2.1635; and 2004, US$ 3.10670.

B. The role of maps in the census

2.12. 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.

2.13. Mapping has been an integral part of census-taking for a long time. Traditionally, the role of maps in the census process has been to support enumeration and to present aggregate census results in cartographic form. Very few enumerations during the last several census rounds were executed without the help of detailed maps.

2.14. In general terms, mapping serves several purposes in the census process, as follows:

(*a*) *Maps ensure coverage 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 data-collection units. Maps showing enumeration areas thus provide an essential control device that guarantees coverage of the census;

(*b*) *Maps support data collection and can help monitor census activities (during enumeration)*. During the census, maps ensure that enumerators can easily identify their assigned geographic areas, in which they will enumerate households. Maps are also issued to the census supervisors assigned to enumerators to support planning and control tasks. Maps can thus also play a role in monitoring the progress of census operations. This allows supervisors to strategically plan, make assignments, identify problem areas and implement remedial action quickly;

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

C. From maps to geographic databases: the mapping revolution continues

2.15. Today, maps are but one form of information display subsumed under the broader term 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 did not therefore find wide application beyond a few government and academic projects. It was not until the 1980s that commercial geographic information systems reached a level of capability which led to their rapid adoption in local and regional government, urban planning, environmental agencies, mineral exploration, the utility sectors and commercial marketing and real-estate firms. This had an impact of the initial use of mapping technologies by NSOs.

2.16. New information sources also shorten the time from project planning to the creation of an operational database. The most important recent developments have been in navigation, remote sensing, image analysis, data manipulation and Internet mapping. GPS has 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.

2.17. 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. Census users develop data models for the 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 modelling 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. This enabling technology has permitted data producers to move from a paper map base to a digital base, updating existing information. Intelligence is still needed, however, to add value to the subsequent products and services made possible by the technology adoption.

2.18. Similar advances are occurring in the areas of geographic data dissemination. All major GIS vendors now provide the tools to make geospatial or 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. Correspondingly, maps themselves are no longer static objects but dynamic snapshots of a changing geographic database.

2.19. Internet mapping applications 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.

2.20. Some statistical offices were 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 in many Latin American countries the national statistical and mapping agencies 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.

2.21. Cartographic automation, GIS and other geospatial tools have enabled more efficient production of both enumerator maps and thematic maps of census results. In addition, advances in technology and new tasks for GIS using new data sources, such as remote sensing and GPS-enabled locational recording, have expanded the power of geographic representation within the NSO. Nevertheless, new technology should be used judiciously based on strategic goals that are determined by the leadership of the census organization. It is not a magic bullet that can solve all organizational problems but must be planned carefully.

D. Increasing demand for disaggregated statistical data

2.22. 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. If carefully planned — that is, if the NSO can collect the information in small units and then aggregate it appropriately — then it should be able to satisfy the needs of many new data customers. Some examples of applications for such data include:

(*a*) **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 the following. 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 centres, schools, water supply systems and government buildings, etc.? If digital maps of population distribution and housing characteristics could be 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 above all their location. Standard "P-Codes" for populated places eases the difficulty of locating affected areas in need of assistance;

(*b*) **Flood plain modelling.** 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;

(*c*) **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 allow 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 centres and more rational decisions concerning the location of new facilities;

(*d*) **Poverty analysis.** In countries where income or consumption data are not collected during a census, household characteristics are an important indicator of 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 channelling resources to areas of greatest need while avoiding leakage of subsidies to non-poor communities;

(*e*) **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.

(*f*) **Labour force analysis.** Whether it is a private company looking for a suitable site to locate a factory or a government agency attempting to match labour 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 for transportation planning;

(*g*) **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;

(*h*) **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;

(*i*) **Epidemiological analysis.** Small area census data, in combination with health incidence and biophysical 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;

(*j*) **Agriculture.** Geographic information on agro-ecological conditions, together with production data and 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.

E. Investing in geospatial technology: costs and benefits

2.23. This section discusses the costs involved and the potential benefits realized in adopting geospatial technology for census operations. There are a variety of options, ranging from a fully integrated in-house geospatial infrastructure 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 geospatial technology in the census process. However, advanced innovation and lowered costs for entry make conversion at this point certainly within reach as GPS technologies become affordable and 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.

2.24. As we have stressed, the adoption of any new technology presents challenges to existing organizations 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, contract support, geodatabase design, transitional costs, data acquisition, data capture, quality control, system maintenance and product development. The transition to geospatial data incurs costs associated with digitization, namely the conversion of analogue enumeration area maps to geospatial formats.

2.25. The 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.

2.26. *Effectiveness benefits* pertain to the impacts of policies or programmes 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 non-governmental organizations and improved outreach to the public at large. Some societal benefits, such as humanitarian response and health infrastructure planning, can be inferred from the examples presented in the previous section. Indeed, the effective use of geospatial technology can mean lives are saved and human well-being overall is improved.

2.27. For various reasons, it is also 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 labour costs are low, may still be the traditional manual approach. However, from a societal standpoint, it may be more cost effective to invest initially more money in a digital approach. This is because the digital output products will realize much larger long-term benefits inside and outside the census or statistical agency, creating a truly national initiative. There is also no guarantee that labour costs will remain low.

Box II.2
Technological and cost hurdles

1. Cambodia

Cambodia undertook EA-mapping 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 online, using a CamInfo system based on DevInfo. Cambodia estimates that the cost of cartography and EA delineation, combined with that of data processing, represents approximately 15 per cent 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.

Source: Cambodia presentation, Bangkok workshop, 2007.

2. Lesotho

Lesotho undertook its most recent population and housing census in 2006. The pre-enumeration census-mapping was done using the latest technology in remote sensing and GIS. This includes total coverage SPOT 2.5m natural colour satellite imagery, GPS and GIS. Preparatory office work was done by a staff of 5. This involved preliminary EA delineation through image interpretation (and local knowledge), as well as the preparation of satellite image maps for EA demarcation fieldwork. EA demarcation fieldwork was done by 8 teams of 4 fieldworkers each. Each team was equipped with a vehicle and a GPS. EA boundaries were verified and amended, where necessary. Additional attribute information, such as place names and names of local leaders, were collected. The information collected during fieldwork was captured in digital format and EA and supervisory areas maps were ultimately created and printed using the GIS. A total of 4,500 EAs were demarcated. The pre-enumeration census-mapping was done in a period of 9 months. The pre-enumeration census-mapping represented approximately 5 per cent of the total census budget. Around 40 per cent of the pre-enumeration census-mapping budget was spent on satellite imagery. Although this is a significant figure, it resulted in a saving of more than 60 per cent in fieldwork because there was no longer a need to prepare individual sketch maps for each EA. In the end, this modern approach resulted in a better quality product, as well as a saving of about 20 per cent when compared with the original budget for the sketch map approach.

Source: Geospace Inc. presentation, Lusaka workshop, 2007.

2.28. 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 obvious difference is that in the traditional mapping case, maps are recreated manually for each census. This involved much needless duplication of effort. 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 were only used for the enumeration. Several years later, the process started again for the next census.

2.29. Having a long-term strategy for geospatial technology adoption 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 analogue case, the costs tend to be higher than the benefits, since the hard-copy maps are useful for census purposes only. In the digital case, an initially large investment will likely result in lower maintenance and updating costs and sustainable benefits in the long term. The long-term benefits will probably be significantly higher because the process results in a digital database with multiple uses by the NSO and other government agencies.

F. Critical success factors for geospatial implementation in the national statistical office

2.30. Besides the obvious costs that can be quantified for a given geospatial application in the census process, 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 implementations. The following prioritized list of critical success factors is adapted specifically for NSO work:

1. Detailed strategic, operational and managerial planning, based on a realistic assessment of costs and effort involved, with clear goals and objectives defined for geospatial operations. Well-established quality control/quality assurance procedures.
2. The decision to invest in geospatial technology is need-based and problem-driven rather than technology-driven. Technology is treated not as an independent add-on but as an integral part of the overall management strategy.
3. A champion or key person(s) to promote geospatial development within the organization and high-level management support.
4. In terms of human resources, the ability to provide training and support for relevant employees and management. The ability to meet staffing needs, including the ability to retain skilled staff and hire contractors, where appropriate. Well-defined written contracts, with vendors, consultants, partners and clients within and outside the government.
5. Completion of a user needs assessment and a priori definition of output products, with a clear implementation schedule. Defined long-term funding plan, including cost recovery and data-pricing strategies, including accurate estimates for maintenance and associated costs. Frequent milestones and delivery of output products to encourage adherence to pre-set time frames.
6. The development of cooperative agreements with other interested parties, including spatial data-infrastructure arrangements. Facilitation of data-

sharing through detailed documentation and metadata across user networks.

7. The usability of systems and products.
8. Adherence to geographic standards.
9. The use of data-integration methodologies.
10. The use of clear protocols, for example with GPS for the collection and processing of data.

G. Planning the census process using geospatial tools

2.31. 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 the organizational structure for geographic support (including staffing and cooperation with other agencies), the explicit delineation of census geography and the design of the geospatial database. As illustrated in figure II.1, these stages can be carried out more or less simultaneously using organizationally approved methodologies, and many of the choices depend also on the chosen data-input strategy.

H. Needs assessment and determination of geographic options

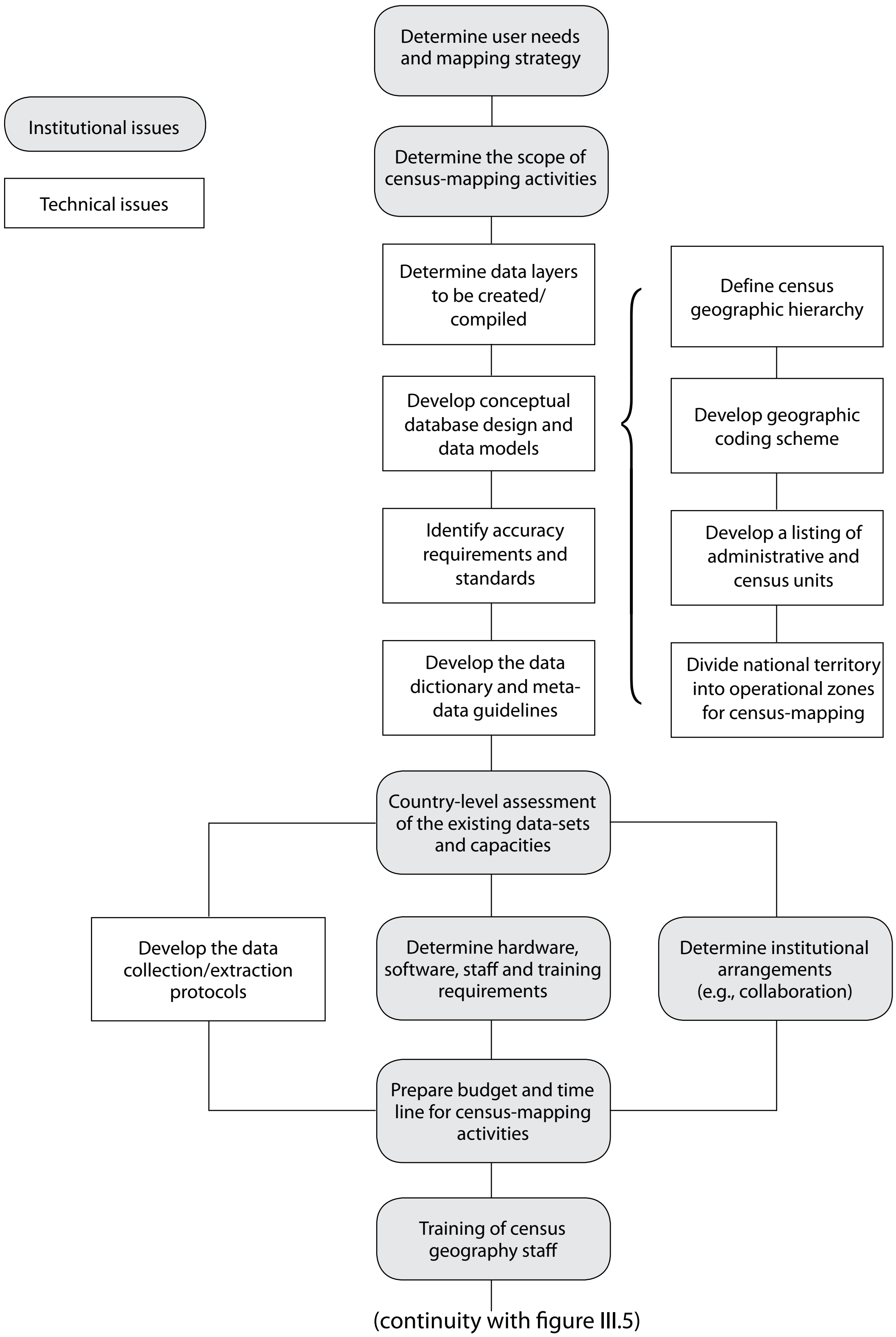
1. User needs assessment

2.32. One of the first steps in a census-mapping project is a detailed needs assessment, followed by an investigation of feasible geographic options. The census-mapping agency must then reconcile user expectations with what is feasible given available resources, working backward from final products and services to requirements.

2.33. 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 include consultation on *geographic content*, that is, geographic structures, including administrative hierarchies or summary levels, and also on *geographic base products* that support the analysis of census data. This planning component should be embedded in the general consultation programme for the census. As the demand for spatially referenced census data increases, consultations concerning geographic products will receive a more prominent role in this process. Institutions that use statistical maps, for instance, 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 census operations.** In order to obtain full information about resources and potential bottlenecks, the NSO must carry out an intensive survey of available human resources in the country, available equipment that can be used, existing digital and analogue map products, and ongoing or planned relevant activities by other

Figure II.1
Stages in planning census geographic work



public and private entities. Avoiding duplication of efforts is key to reducing the cost of census geographic operations and to timely delivery of census 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 information display options, private users are also an important user group. Citizens may, for example, want to obtain statistical information about their own neighbourhood or a neighbourhood they intend to move to. With the current rapid changes in technology, the census office must plan carefully to anticipate demand for its data.

2. Determination of products

2.34. 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 VI, should always include proper documentation, including coding and metadata, to make them most useful to users. Examples include:

- A set of digital enumeration area maps or derived dissemination units, which 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 analysing 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.

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

- Available financial and human resources.
- Existing digital and analogue geographic products.
- The degree of integration between the mapping and statistical agencies or other relevant agencies in the country.
- 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 the increased use of low-technology labour which may provide a beneficial boost to local economies.
- The size of the country, both population and areal extent, and also accessibility as affected by terrain and water bodies.
- The time frame available to plan and carry out the census-mapping process.

3. Geographic data options

2.36. 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 geographic database for census-collection and dissemination purposes. The following partial list of available options, in increasing order of complexity, is dependent on budgets and time:

- Production of digital maps created on the basis of existing sketch maps.
- Geo-referenced enumeration area layers or spatial files, with proper coding and metadata 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.
- A digital database of precisely located dwelling units, created with the aid of geographic positioning systems.

2.37. This list is incomplete and is provided here for illustrative purposes. All of these issues are discussed in detail elsewhere in the present *Handbook*. The most appropriate census geographic strategy for a country will consist of a tailor-made approach that considers the country's needs and resources. The *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.

4. Human resources and capacity-building

2.38. 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 geographic staff since similar products are created using different techniques. Furthermore, a digital geographic database is useful for many more purposes. A census office is thus likely to fulfil additional demands for products and services which were not available before. Every member of the census geographic staff should therefore have some degree of computer literacy.

2.39. 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, including databases, spreadsheets, file management and basic network operations. 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 map-making, 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.

2.40. The profile of tasks for which staff are required in a digital census-mapping project is set out in detail below. 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.

2.41. **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, geospatial 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 converted to digital census geography or from international organizations should be involved in the planning process as they can provide useful input into the planning process and help to cost alternatives.

2.42. **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 geographic information science 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 geospatial technology developments and trends, and must be prepared to adapt the census geographic strategy if conditions change or better solutions become available. It is important not to formulate strategy once the process begins and to be aware of "mission creep." Change-control mechanisms provide an essential means of responding to unexpected challenges to the plan.

2.43. **Geospatial 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 techniques, such as scanning, digitizing and editing of geospatial 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.

2.44. **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 a 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 task and, while it requires careful concentration, attention to detail and a good understanding of the structure of digital geographic databases, it can usually be assigned to clerical staff. The best-performing staff should also receive training in quality control/quality assurance approaches. This training requirement pertains as well to heads-up digitizing.

2.45. **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.

2.46. **Field work.** The requirements for census geographic fieldwork have changed with the techniques used for digital map production. As global positioning systems (GPS) 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 hand-held 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.

2.47. **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 down time, 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 geographic team since almost every aspect of the work depends on a well-functioning computer system. The 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.

2.48. **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 the optimization of software systems.

2.49. **Levels of training.** In many countries, there may be a shortage of trained geospatial 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. The following different types of training will be required for various purposes:

(*a*) Short seminars to raise the awareness of the digital census-mapping programme 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 geographic staff;

(*b*) 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 reallocation to other duties;

(*c*) The core geographic staff involved in census-mapping should receive additional training in GIS software 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 who 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-the-trainers" approach, which is particularly appropriate for a decentralized approach to census-mapping;

(*d*) 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, sending a staff member to a university for training. Many universities and training centres around the world now specialize in professional one-or two-year degree courses in GIS, remote sensing and related techniques (for listings of institutions offering training in geospatial information and techniques, see annex VII).

I. Institutional cooperation: national spatial data infrastructure: ensuring compatibility with other government departments

2.50. 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 transportation, health, environment and water resources units, also use geospatial technology 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.

2.51. 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 software applications. In this framework, NSOs are seen as information nodes in networks of data and content providers that span the country.

2.52. National spatial data infrastructures (NSDIs) now exist in over 100 countries in the world. Some possible sources of information and expertise that could benefit the NSO in building its spatial data infrastructure include national mapping

Box II.3
Three examples of data-sharing collaborations

1. UNFPA and Fiji hardware-sharing

The development of new technologies for census geography provides powerful tools for increasing the efficiency of census-taking. In Fiji, 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 GPS devices. This was one of the examples of how to add value to the census, as this procedure enabled a much more accurate future mapping application and will facilitate future censuses and statistical exercises. Funding for the GPS exercise was facilitated by the United Nations Population Fund (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 (for further information, contact Scott Pontifex, Secretariat of the Pacific Community).

2. Secretariat of the Pacific Island Applied Geoscience Commission (SOPAC) and the Pacific island countries data-sharing and processing agreement

The costs of acquiring satellite imagery that would enable using 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 borne by the majority of national statistical offices. International and national agencies can greatly contribute their existing data-sets to aid in census-taking activities. In that context, the Secretariat of the Pacific Island Applied Geoscience Commission (SOPAC) had access to satellite imagery, which it can provide 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 (for further information, contact Scott Pontifex, Secretariat of the Pacific Community).

3. Data-sharing experience of the United States

The United States Census Bureau works with tribal, state, county and local officials, as well as coordinating agencies, such as regional planning commissions, to implement several programmes designed to share geographic information. The knowledge provided by local partners allows the United States Census Bureau to meet the statistical and spatial data needs that are part of its mission: to be the statistical source for a better understanding of the nation.

The United States decennial census and the annual American Community Survey are conducted principally by mailing questionnaires to households throughout the country. A close partnership with the United States Postal Service offers regular address updates to the Geography Division's Master Address File, which serves as the source for mailing questionnaires. Under agreement to protect confidentiality, the United States Census Bureau shares its address list with tribal, state, county and local government officials to assure an accurate mailing, thereby assuring an accurate count.

Correct locations of boundaries are important for collecting, tabulating and disseminating statistical data. The United States Census Bureau conducts an annual boundary and annexation survey, in which local governments review their boundaries and report any changes. In preparing for the decennial census, regional planning organizations and governments have an opportunity to review statistical areas and recommend changes to better meet their local data needs.

The growth of GIS systems in local governments in the United States offer an opportunity for a federal agency like the United States Census Bureau to acquire, through partnerships, highly accurate and current geospatial data which helps the agency maintain its nation-wide Topologically Integrated Geographic Encoding and Referencing System (MAF/TIGER). MAF/TIGER is the source of all geographic activities at the statistical agency, such as address support, spatial data use, geocoding and mapping.

Coordination and partnership with other federal agencies in efforts to build a national spatial data infrastructure occur through the United States Census Bureau's active participation with such groups as the Federal Geographic Data Committee. Agreements with such agencies as the National Geodetic Service help the agency to acquire more accurate positional information on locations of housing units. Lastly, close collaboration with such national organizations as the National States Geographic Information Council and the National Association of Counties help to build the partnerships necessary with local governments.

Current trends in partnerships are critical for maintaining the geospatial framework needed to conduct censuses and surveys. In the United States, it is no longer cost effective to build and maintain the amount and quality of address and spatial data that in previous decades were dependent on a federal, centralized effort. Close collaboration with partners at local levels offer new opportunities for better data to support the geospatial infrastructure of a statistical agency (for further information, contact Tim Trainor at: timothy.f.trainor@census.gov).

* * *

It is crucial to ensure cooperation at the national level when it comes to acquiring, using and developing GIS applications. The establishment of GIS focal points and user groups is an optimal way to ensure synergy and avoid duplication and waste of otherwise scant human resources. The importance of 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.

authorities, 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.

2.53. Establishing contact with the existing NSDI requires identifying the contact person in the lead agency. The main contact is typically the national mapping agency. 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, which 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.

2.54. Numerous users inside and outside government agencies require access to 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, which provide the basis for many mapping and data-collection activities, are termed "framework data". Core data layers that form the elements of national spatial data frameworks include (see Economic Commission for Africa (2007)):

(*a*) **Geodetic control.** A system of precisely determined geographic control points that serve as the reference for all mapping activities in a country, sometimes termed benchmarks;

(*b*) **Base geography.** Includes imagery (air photos or high resolution satellite images), hypsography (terrain) and hydrography (surface water features; can be natural, such as rivers and lakes, or artificial, such as canals);

(*c*) **Administrative and spatial organization.** Includes subnational governmental units; such as provinces and districts, that have been delineated on the ground, geographic names and land management units/areas;

(*d*) **Infrastructure.** Includes roads, inland waterways, railroads and any infrastructure used to move people or goods, utilities and services;

(*e*) **The natural environment.** Includes soil types, vegetation zones, geographical names of population settlements.

2.55. Most relevant for the census office are the governmental administrative units. Clearly, 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.

2.56. 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 data-set containing the centre 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.

2.57. The concept of a national spatial data infrastructure, consisting of basic geographically referenced geographic databases, has three implications for census-mapping activities:

(*a*) The census office has a responsibility to contribute to the national spatial data infrastructure a consistent set of data-reporting units that are consistent with the administrative hierarchy and to which socio-economic 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;

(*b*) 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;

(*c*) NSOs should pay particular attention to issues concerning geographic referencing. In order to overlay different data-sets that cover areas in the same geographic extent, one needs to know how the position of features has been defined. In other words, one needs to know the projection and datum and details of the coordinate system to ensure the correct spatial relationships between features in different data-sets.

2.58. A statistical agency's contribution to NSDI efforts nationwide 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.

Box II.4
International agency participation and coordination

1. The Global Map Initiative

The Global Map Initiative was proposed in 1992 at the United Nations Conference on Environment and Development, 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 developing data for, 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 km. Data will be updated at approximately five-year intervals to monitor change over time. Global Map is a platform supporting four base maps in raster format (land use, land cover, vegetation and elevation) and four data maps in vector format (population centres, drainage, transportation and boundaries). 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).

2. DevInfo

The United Nations Children's Fund (UNICEF) DevInfo is a software tool to assist countries in monitoring the Millennium Development Goals and advocate their achievement through policy measures, multisectoral 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 and maps) for inclusion in reports and presentations. DevInfo is expected to be a powerful advocacy tool that will contribute to greater awareness and knowledge of the Goals at the country level and to evidence-based policymaking (for more information, see http://www.devinfo.org/).

3. United Nations Spatial Data Infrastructure

The United Nations Spatial Data Infrastructure (UNSDI) 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 the access, retrieval and dissemination of geospatial data and information in a rapid and secure way. UNSDI enables inter-operability between spatial data infrastructures developed for specific purposes within United Nations agencies, among groups of United Nations agencies sharing common interests and among United Nations Member States and their regional and thematic groupings and partners. The access, retrieval and dissemination of geospatial data and services will be enabled by UNSDI, avoiding duplication within the United Nations (for more information, see http://www.ungiwg.org/unsdi.htm).

J. Standards

2.59. To facilitate data exchange between data users it is clearly necessary to coordinate the 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, supranational organizations, such as the European Umbrella Organization for Geographic Information (EUROGI), the Global Spatial Data Infrastructure (GSDI) Association, the International Organization for Standardization Geographic Information/Geomatics Committee (ISO-TC/211) and the OpenGIS Consortium (OGC) are active in defining geographic data standards.

K. Collaboration

2.60. In the process of creating a digital geographic database, the census organization is encouraged to collaborate with other government agencies or with the private sector. Both options have been used successfully in different countries. As noted above, the national mapping agency is the natural first point of contact among government agencies, especially for NSOs that are just getting started. But other agencies (such as cadastral, environment and local government agencies) 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. Other collaborators can include data providers, academic departments and road authorities. It should be noted, however, that collaboration with other entities is desirable but not mandatory. Since the census-mapping agency has as its main priority the production of a geographic base for the census, it needs to avoid dependence on an outside supplier of data.

2.61. 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:

(*a*) **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. Important issues to consider include the need to define standards for both internal and external use, the value of clearly stated goals and the need for management support;

(*b*) **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. Protocols should be spelled out for making use of data from municipalities and other lower-order geographic or administrative entities;

(*c*) **Responsibilities.** Who will perform which tasks and functions? Issues that need to be addressed include data development, maintenance, data access, project supervision and use of resources;

(*d*) **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;

(*e*) **Resource requirements.** Resources include staff, computing environment, materials and communication. Resources required for management and project supervision must also be considered;

(*f*) **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 may take the form of cash, data, labour, equipment use or some other form.

(*g*) **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.

(*h*) **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.

L. Summary and conclusions

2.62. Planning for a geospatial census will require consideration of funding, staffing, and project management issues. As demonstrated above, national statistical organizations have a great opportunity to use new geospatial technology, including the widespread availability of personal computers, hand-held devices, GPS and low-cost aerial and satellite imagery, to obtain information about their populations. For many NSOs, the task is to enlist new talent and reorganize the enterprise to accommodate it. This means recognizing the challenge of attracting and retaining trained personnel and many countries are expressing their concerns about how to tackle these issues.

2.63. NSOs require information about 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 thereby be able to obtain the answers they need to develop a geocentric digitization plan, scaled appropriately for the country, that will help them to realize efficiency and effectiveness benefits. It should be remembered that planning is paramount. Agency leadership should state goals clearly and support the strategic plan in order to ensure success.

Chapter III
Constructing an EA-level database for the census

A. Introduction

3.1. As noted in chapter II, using geospatial technology to create better data is in some ways an organizational issue, involving goal-setting and harnessing the right human resource skills. Reorganizing that the NSO around a geographic information core means embracing the relationship between a country's geography and the various sets of information that the NSO uses and produces. The relationship between geography and databases occurs through the mechanism of coding. The first step is to link the management material described in chapter II to technical content, showing how the geographic census database becomes the focus of activities where forms of census information are stored and accessed.

3.2. A key goal of the present *Handbook* is to construct an operational plan for building an 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 advancing their development goals. Whereas it was formerly a one-off exercise usually done every 10 years, maintaining an accurate geographic base for the census is now a continuous process.

3.3. Supporting census operations by 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 towards 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 redesign of census-mapping operations, the NSO can build off the work of its predecessors and use maps developed in prior years.

3.4. The present chapter will introduce a concept of geographic coding that differs slightly from that conventionally understood by the GIS industry to mean address-matching. The United Nations definition of "geocoding" is broader. It represents the connection between statistical observations and real-world locations expressed in terms of latitude and longitude or other locational attributes. Simply put, geographic coding is a way to ensure that the data know where they are. While the present chapter advances this broader definition of geocoding, it also emphasizes the importance of traditional coding (i.e. attribute coding) for a census.

3.5. Geocoding is designed to cover a continuum of spatial scales, from individual housing units through enumeration area-level up to higher administrative or national levels. Its successful use depends on a country establishing 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.

3.6. Administrative hierarchies are based on the idea that inside the territory of the country there are boundaries that serve to demarcate the actual land extent at the 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 number of units. For example, units at administrative level two (ADM 2s) are provinces, while units at administrative level three (ADM 3s) are districts. Ideally, any geospatial operation would have access to these units in GIS format for use in its various projects.

3.7. The present 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 and defensible plan. One must highlight, then, the need for regional coordination, especially in big countries.

3.8. The NSO role in administrative boundary delineations will vary by country. It is possible that the NSO will have legal authority to define subnational units on the ground, or perhaps it is assumed to have de facto authority. If the NSO is not the custodian of subnational boundaries, then it will have to work with other agencies within the Government to use the existing geocoding plan, possibly adapting it for particular census needs. Some regional coordination may be necessary as well.

3.9. Among the topics covered in the present chapter are the definition of the national census geography, including administrative hierarchies and criteria and the process for EA delineation; the coding of geographic areas, including EAs, and ensuring their compatibility with previous censuses; geographic data sources for EA delineation and importation procedures; geographic data conversion through scanning and digitizing; constructing and maintaining topology; implementation of an EA database; data-quality issues; and metadata development.

3.10. As always, plans should be made carefully. In addition, it is important to emphasize that, with the growing demand for small-area data, there is a need to assess country needs realistically.

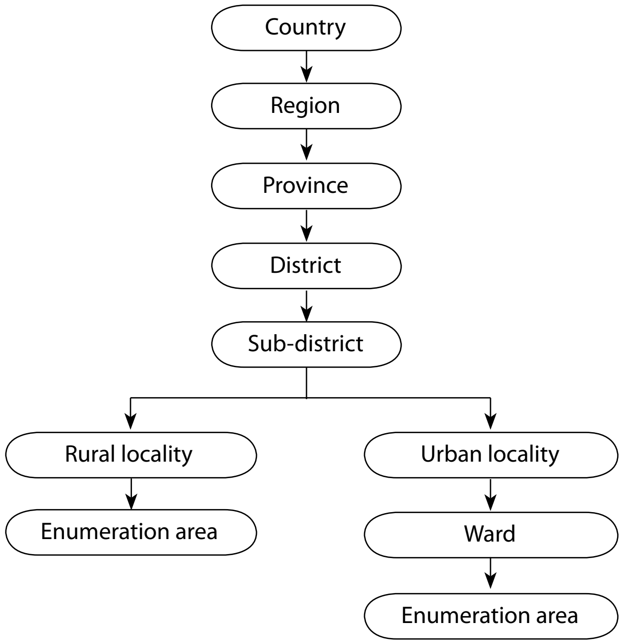
B. Definition of the national census geography

1. Administrative hierarchy

3.11. 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. The relationships among all types of administrative and reporting unit boundaries should be 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 (see fig. III.1).

3.12. 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. Other units may have statistical roles alone; that is, they are designed for the display of data and not for administering territory. Figure III.2 illustrates the nesting of administrative and census units,

Figure III.1
A generic census geographic hierarchy



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.

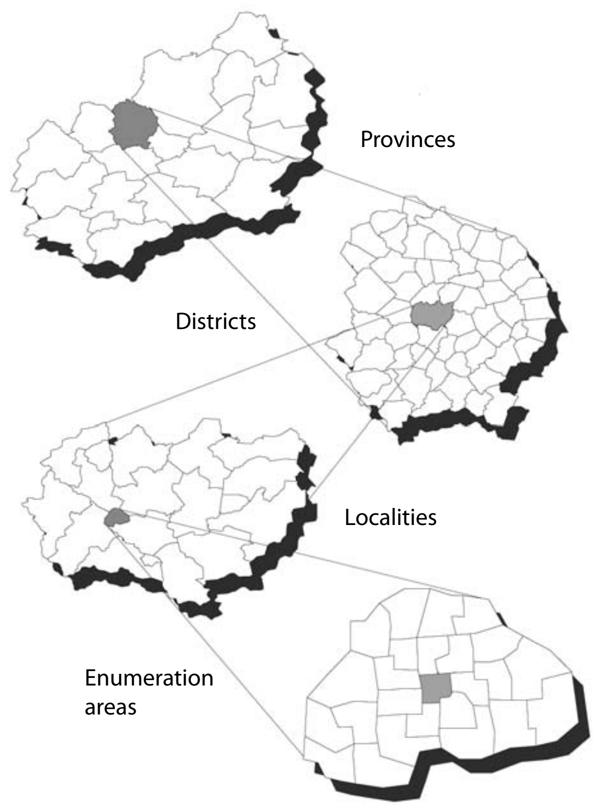
3.13. All levels do not have to be equally important. For example, many countries divide the territory into major regions, which are often geographically defined, such as "North/South/South-West/East", or "Mountain-Plains-Coastal". These regions often do not serve any administrative function but may still be used to report statistical information.

2. Relationship between administrative and other statistical reporting or management units

3.14. 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.
- Labour 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.
- Utility zones (water or electricity supply districts).

Figure III.2
Illustration of a nested administrative hierarchy



3.15. More can be learned about these special geographies through interaction with the responsible agencies in the 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.

3.16. 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 programmes 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.

3.17. 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 of population census information and 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.

3.18. Coordination should also be maintained between the statistical office and other agencies or levels of government as to who has custody of administrative boundaries and when changes can be made. Prior to a census, the boundaries should be frozen (this is recommended at least six months prior to the census) so that boundary changes do not create discrepancies in the actual coding of areas. NSOs can also choose to "time-stamp" specific versions of administrative boundaries so that they reflect the internal divisions of the country at the time of enumeration.

3. Criteria and process for ground delineation of EAs

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

- Be mutually exclusive (non-overlapping) and exhaustive (cover the entire country).
- Have boundaries that are easily identifiable on the ground.
- Be consistent with the administrative hierarchy.
- Be compact and have no pockets or disjoined sections.
- Have populations of approximately equally size.
- Be small and accessible enough 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.
- Address the needs of government departments and other data users.
- Be useful for other types of censuses and data-collection activities as well.
- Be large enough to guarantee data privacy.

3.20. 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. It should be kept 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 should thus take precedence over convenience of census enumeration. However, EA delineation must also make sense logistically for field operations.

3.21. 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 generally determined on the basis of pre-test 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.

3.22. 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 fieldwork, cooperation with government officials, extrapolation from previous census results or by means of aerial photographs or satellite imagery as discussed in the next chapter.

3.23. 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, including brush, forests and elevation contours, such as ridges, are less ideal. Administrative boundaries are often not visible. 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.

3.24. 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. Again, naval personnel are often assigned to their home ports. When planning for locating hard-to-enumerate populations, it should be recalled that operational costs are sometimes 10 to 20 times higher than for residential populations in urbanized areas.

3.25. 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 pre-test can determine the number of housing units (HUs) that an enumerator can cover per day. As an illustrative example, if 16 HUs can be enumerated per day in urban areas but only 10 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; the presence of collective living quarters, such as barracks, hotels and school dormitories; and the mode and availability of transportation.

3.26. 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.

4. Delineation of supervisory (crew leader) areas

3.27. Supervisory areas provide the means for crews of enumerators to be effectively managed. After delineation of EAs, the design of supervisory maps is usually straightforward. Supervisory areas consist of groups of usually 8 to 12 contiguous EAs that share some of the same characteristics as enumeration areas. The EAs assigned to the same supervisory area must be compact, in order 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.

3.28. 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.

5. Geographic coding or "geocoding" of EAs

3.29. 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. A more meaningful identifier is needed 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.

3.30. 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.

3.31. 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 in the design of a coding scheme are flexibility, expandability, 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.

3.32. A hierarchical coding scheme will usually facilitate the 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. In other words, the gaps that are left will depend on how many units are added. 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 the 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.

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

Figure III.3
A generic enumeration area coding scheme

3.34. An EA code of 1203501750023 means that enumeration area number 23 is located in province 12, district 12035 and locality 120350175. The unique code is stored in the database as a long integer or as a 13-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.

3.35. 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 American Standard Code for Information Interchange (ASCII) characters rather than a sequential number.

3.36. 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 should 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 and the latitudes and longitudes of centroids (central points), possibly with populations enumerated as well.

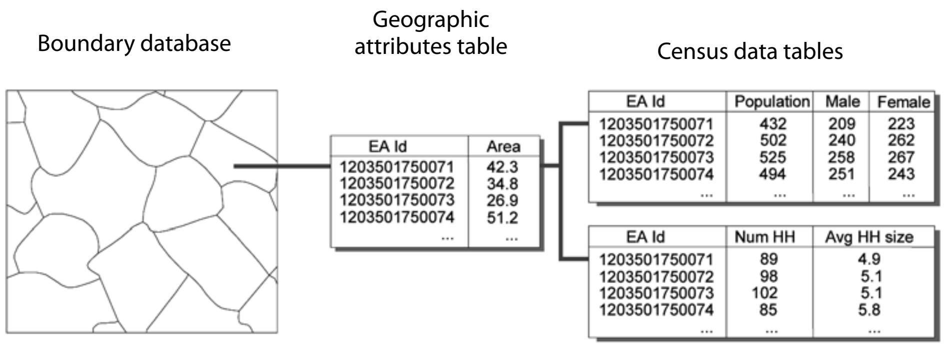
6. Components of a census database

3.37. A comprehensive census geographic 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 of all types, rivers, buildings or settlements, which are stored as separate entities in a geographic database. Roads and rivers, for instance, while both represented as lines, would not be stored in the same digital file. It is necessary to provide written specifications that specify outputs, sizes and other requirements to ensure consistency.

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

3.39. The first step is to think about what the maps to be used by enumerators will look like. The complete digital enumeration area database, for example, will probably consist of such elements as features and attributes, which can be represented as follows (see also figure III.4):

Figure III.4
Components of a digital spatial census database



- The spatial *boundary database,* consisting of area features (polygons) that represent the census units.
- The *geographic attributes table*, a database file linked internally to the spatial database that contains one record for each polygon. This table contains the unique identifier for each census unit and possibly some additional static or unchanging variables, such as the unit's area in km^2.
- 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 that 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, health facilities or other buildings may be useful for orienting fieldworkers during the enumeration. Such features, which are recorded during preliminary field-canvassing or house-listing, may prove useful later to other government agencies or non-governmental organizations, thus saving time and money. Collaboration with other data-users offers many efficiency and effectiveness benefits and should be pursued whenever feasible.

7. Consistency of EA design with past censuses

3.40. 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 analyse 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. The problem of changing the geographic base between censuses is no less serious than changes of definitions of items on the census questionnaire.

3.41. In designing the census geography, the census office should therefore attempt, as far as possible, to preserve boundaries from the previous census. Owing 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.

3.42. One component of EA delineation that can facilitate change analysis is the compilation of compatibility or equivalency files, sometimes also known as relationship 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.

3.43. 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. To ensure that the census databases that are the product of the data entry and tabulation programme will match the geographic boundary files, close cooperation between the census geographic and data-processing sections is required.

3.44. 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, also have to be made 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.

3.45. 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 that case, it might not be necessary to store the descriptive attribute information in a separate table. For simplicity, all attributes could be contained in the geographic attribute table itself.

3.46. At certain stages 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 could be produced for each statistical reporting unit for which data are required. Those 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.

3.47. Existing digital databases, such as products created by another government agency and coordinates collected in the field using GPS, can be imported into a census geographic database. 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. Using that information, census staff can delineate enumeration areas interactively, with the geographic reference information as a backdrop.

3.48. GPS as a data source for mapping and evaluating EAs is covered in more detail in chapter IV and annex II. Additional material is also provided on using GPS and remotely sensed data to correct the geographic database.

3.49. 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 name, geographic level and locational reference. This computerized list is the geographic attributes table and will be linked to the completed GIS database.

3.50. The flow chart in figure III.5 shows only one of many possible sequences in 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 hard-copy 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.

3.51. Whatever 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 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 by allowing a better evaluation of staffing and equipment requirements and the time required to perform all activities.

3.52. The pilot area should be representative for as many different regional types 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.

3.53. 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, GPS 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.

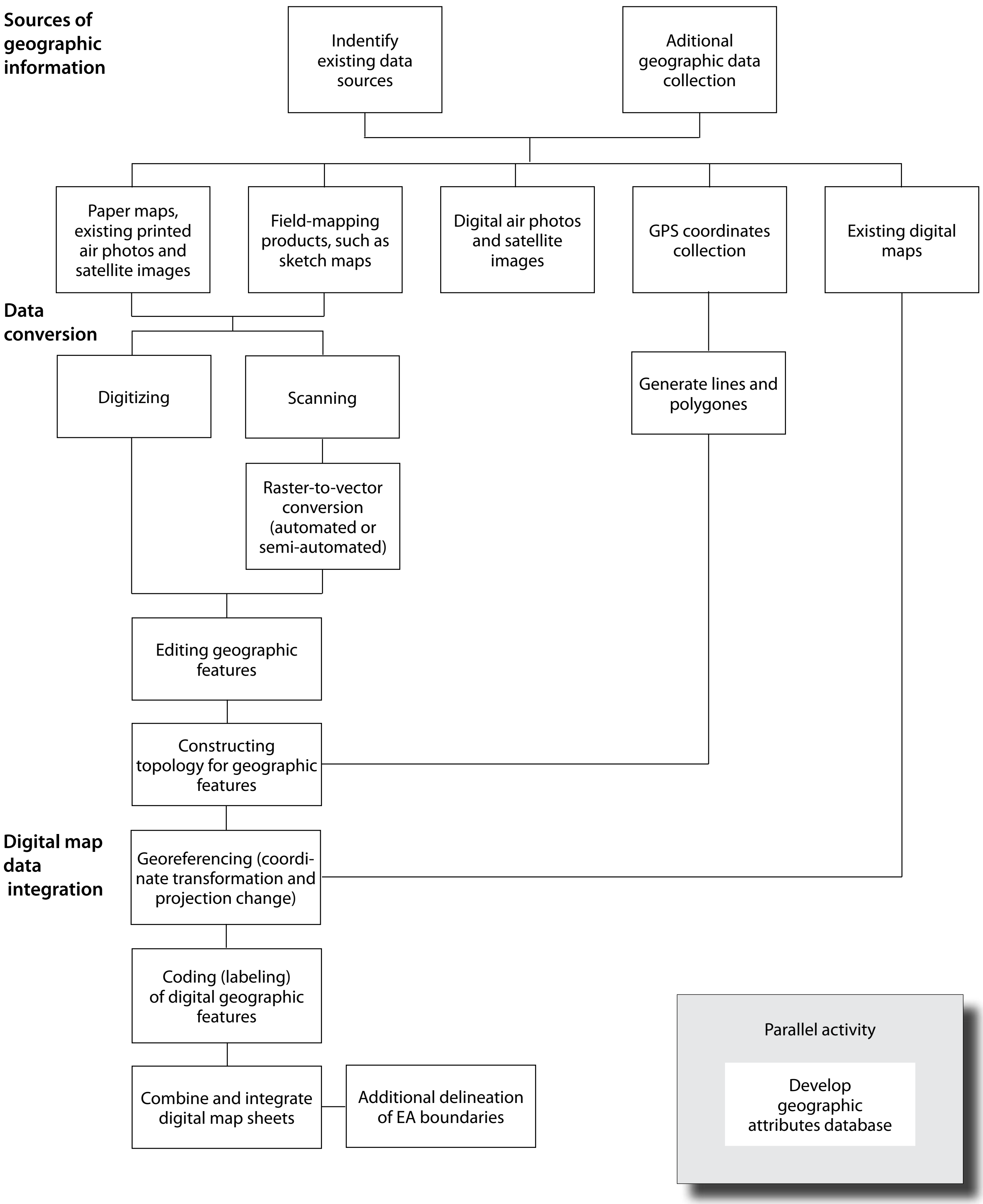
C. Geographic data sources for EA delineation

1. Types of maps required

3.54. In nearly all cases, a census cartographic programme will have to consult existing hard-copy 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 country's territory, including:

(*a*) 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;

Figure III.5
Stages in the census geographic database development



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

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

(*d*) Maps of administrative units at all levels of civil division;

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

3.55. For incorporation in a GIS database, ideally these maps should all have comprehensive documentation. This includes geographic referencing information, including map scale; projection and geographic datum; 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 onscreen 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.

2. Inventory of existing sources

3.56. All maps that have been obtained should be well documented and organized according to the organization of the census-mapping programme — i.e., by census region or district. In addition to hard-copy 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 or metadata is absent, it is often not possible to determine the correct projection information and data quality is difficult to evaluate (for more information on metadata development, see section F below).

3.57. Through activities sponsored by the NSDI or through person-to-person contacts, the following agencies and institutions can be contacted to see whether they can contribute useful hard-copy or digital maps:

(*a*) National geographic institute/mapping agency. This is often the lead agency in the country concerned with mapping and may have already begun a digitization programme for topographic maps. However, in some countries the mapping agency may lack the resources or legal mandate to collaborate extensively with the NSO;

(*b*) 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;

(*c*) Province, district and municipal governments. These can also include urban or municipal planning offices. Local government organizations increasingly use geospatial technology to manage information about transportation, social services, utility services and planning relevant information;

(*d*) Various government or private organizations dealing with spatial data, including:

(i) Geological or hydrological survey;

(ii) Environmental protection authority;

(iii) Transport authority;

(iv) Rural electrification authority;

(v) Utility and communication sector companies;

(vi) Land titling agencies;

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

3. Importing existing digital data

3.58. The direct import of digital data is in most cases the easiest form of digital spatial data conversion. The GIS industry has transitioned from stand-alone 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.

Box III.1

Geospatial software selection criteria: COTS, image analysis and FOSS options

1. Selection options for 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. Different agencies and staff should be consulted about the preferred software platform. An informed choice is the best choice.

Commercial off-the-shelf (COTS) software can be divided into those offering raster/vector integration and primarily image analysis software.

2. Raster/vector integration

Software includes Environmental Systems Research Institute (ESRI) 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 data sets.

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 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 United States 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.

3. Image analysis

Raster-based analysis is becoming commonplace in census-mapping operations. Much 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 data sets, 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 the 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 4.5 (ITT Solutions) integrates raster imagery with geographic information systems. Vector layers can be superimposed on 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, allowing for simultaneous updating of different locations of one large data set. In addition, one can update portions of large vector data sets. An extension for ArcGIS allows ArcCatalog to define a collection of maps as a Definiens' workspace, enabling users to review and edit entries.

Google Earth's strongest selling point is its free imagery. However, there are costs associated with feature upgrades, such as Google Earth Pro. Google Earth offers global coverage, democratizing access for many users. It is user friendly, has a large knowledge base and is useful for quick display of raster and vector data.

Issues to consider include the initial cost of the software, 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.

4. Free and open source software 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 software is freely downloadable from the Internet and provides similar functionality to commercial software. FOSS implies that users can access the source code of the application, meaning that NSOs with programming expertise can tailor software 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 software has traditionally been used by programmers or others with IT experience. This has fortunately changed due to a larger user-base and subsequent product development. Software has become more user-friendly, with online training and product support. FOSS software offers "interoperability", defined by the Global Spatial Data Infrastructure Association as "The capability to communicate, execute programmes, 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 software includes the following:

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 vector features. Multiple annual software updates keep users ahead of product developments. A strong user community provides 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 an application for viewing and analysing 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.

3.59. 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. As a result, difficulties may arise in data conversion. 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.

3.60. 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.

3.61. Apart from problems in converting the data files from one format into another, the most often difficulty encountered 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 and metadata be provided.

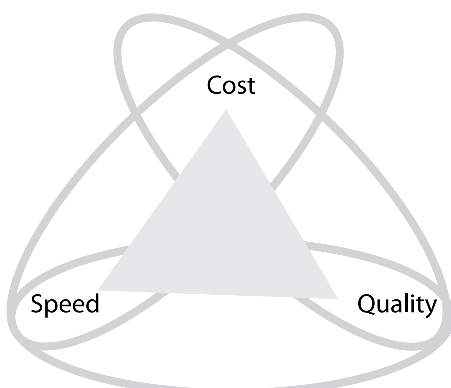
3.62. Other possible problems that may need to be addressed include differences in definitions and coding schemes, the use of different cartographic reference systems, incompatible spatial scales and varying accuracy standards, which may result in features that should match across two databases being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

4. Geographic data conversion: analogue to digital

3.63. 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 hard-copy or digital form and the collection of additional data, using fieldwork, air photos or satellite images. Collectively, we will use the term "data conversion" to refer to these steps. The guiding wisdom for data conversion is to plan carefully. If more care is taken early on in the process, fewer problems will arise when the process is under way.

3.64. 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 (see figure III.6). 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 III.6
Trade-offs in the data conversion process
(after Hohl, 1998)



3.65. Figure III.7 outlines the basic steps in the data-conversion process that lead to a complete digital census database. A survey of existing digital and hard-copy 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.

3.66. 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 III.7
Photo of feed scanner

Source: Ideal.com.

3.67. The process of converting features that are visible on a hard-copy 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.

3.68. The conversion of hard-copy 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 that 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.

3.69. 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 the 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.

3.70. The two main approaches for converting information on hard-copy 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 discussed in more detail below.

(*a*) Scanning

3.71. 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 colour 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 colour signals, when combined, determine the pixel colour. The result of the scanning process is a raster image of the original map that 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.

3.72. 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 III.7).
- **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 advantage of flat-bed scanners is their low cost and easy set-up and maintenance. They are useful for scanning text documents — for example data tables — which are later interpreted by means of 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, which require many large format topographic and thematic maps 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 that 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 colour 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.

3.73. 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 any further processing steps anticipated. The most important parameters are:

- **Scanning mode.** Binary or "line art" is appropriate for monochrome drawings or sketches, as well as for colour separations, where all features are basically of the same type. Gray-scale mode preserves variation on a map and subsequent image manipulation can be used to extract only the features that have a certain reflectance value in a graphics or image-processing system. This is even easier when the maps are scanned in colour 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 colour 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.
- **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.

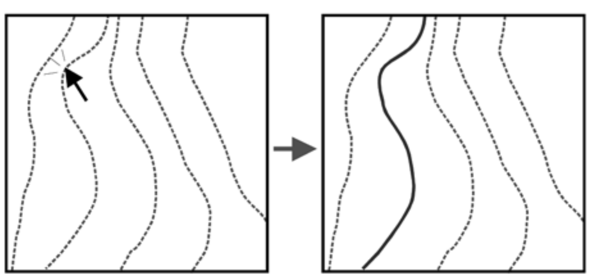
3.74. Scanning the source document is only the first and fairly straightforward step. Since the end-result of the conversion process is a digital geographic database of points and lines, the scanned information contained on the raster images needs to be converted into coordinate information. This process is "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 method of data entry. Recent advances in software development, pattern-recognition techniques and processing speeds have led many to adopt scanning as their preferred method for data input.

3.75. Raster-to-vector conversion can be performed in automatic, semi-automatic or manual mode. In automatic mode, specialized software automatically converts all lines on the raster image into sequences of coordinates. 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 that 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 can be deleted automatically. Also, if the image has been scanned using a colour scanner, raster-to-vector software often allows the user to specify line codes to be assigned to colours. 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.

3.76. In semi-automatic mode, the operator clicks on each line that needs to be converted (see figure III.8). 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 the heads-up digitizing mentioned above.

Figure III.8
Semi-automatic vectorization

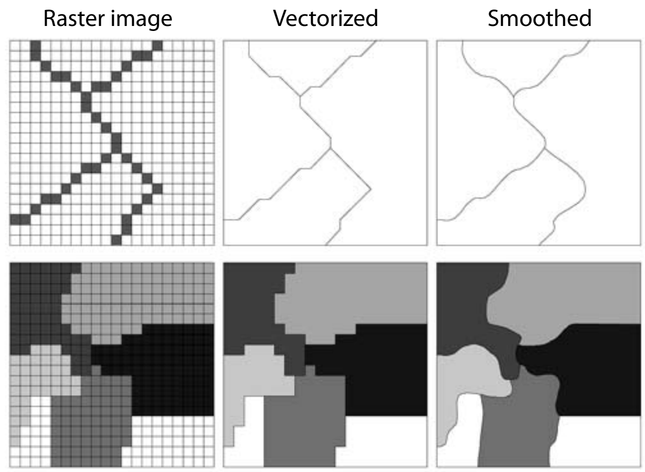


3.77. If linear or area features are converted automatically from relatively low resolution raster images into vector format, the resulting line work may show unnatural sharp edges. It is common practice to smooth the vector data, using spline or generalization functions available in GIS packages. Figure III.9 shows examples for a line and a polygon data set.

(i) *Additional considerations*

3.78. 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

Figure III.9
Vectorization and smoothing of scanned image data



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.

3.79. An additional step is often introduced for the conversion of relatively complex maps that show many different features (e.g., topographic maps) or 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 colour 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.

3.80. 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.

3.81. 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 stand-alone commercial and non-commercial raster-to-vector packages available (for example, Vextractor, AbleVector and PTracer), as well as software extensions, such as ESRI 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.

(ii) *Advantages and disadvantages of scanning*

3.82. The advantages of scanning are:

- 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, using heads-up digitizing.
- Clear base maps or original colour separations can be vectorized relatively easily, using raster-to-vector conversion software.
- Small-format scanners are relatively inexpensive and provide quick data capture.

The disadvantages of scanning are:

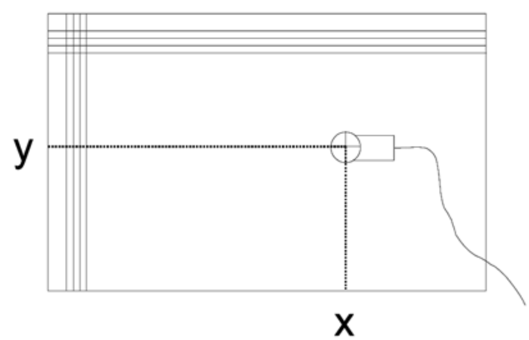
- Converting large maps with small format scanners requires tedious reassembly of the individual parts.
- While large format, high-throughput scanners are expensive, their cost can be justified if they are put to use on large-scale map scanning and vectorization resulting in a digital geodatabase.
- Despite recent advances in vectorization software, considerable manual editing and attribute labelling may still be required.
- Scanning large volumes of hard-copy maps will present challenges for file storage on many desktop computer systems. NSOs considering scanning all their EA maps should consider investing in a separate hard drives with a backup system to accommodate the large volumes of files produced.

(*b*) Digitizing

3.83. Manual digitizing has traditionally been the most common approach for spatial data automation. Manual digitizing requires a digitizing board that may range in size from small tablets of 30 x 30 cm to large digitizing tables of 120 x 180 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 such shrinkage introduces distortions that will be carried into the digital map database.

3.84. 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. First, 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. Second, control points are chosen for which the real-world coordinates in the base map's projection system are known. The ideal control points are therefore the intersections of the graticule of latitude and longitudes that are shown on many topographic maps. In the georeferencing step that precedes or follows the digitizing of point and line features, this information is used to convert the coordinates measured in centimetres or inches on the digitizing tablet into the real world coordinates — usually in metres 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 III.10). 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 III.10
Digitizing table



3.85. 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.

3.86. 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.

3.87. 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 that contains data about the polygon (see annex I).

3.88. 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 (see figure III.11). 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.

Advantages and disadvantages of digitizing

3.89. The advantages of digitizing are:

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

Figure III.11
Heads-up digitizing

3.90. The disadvantages of digitizing are:

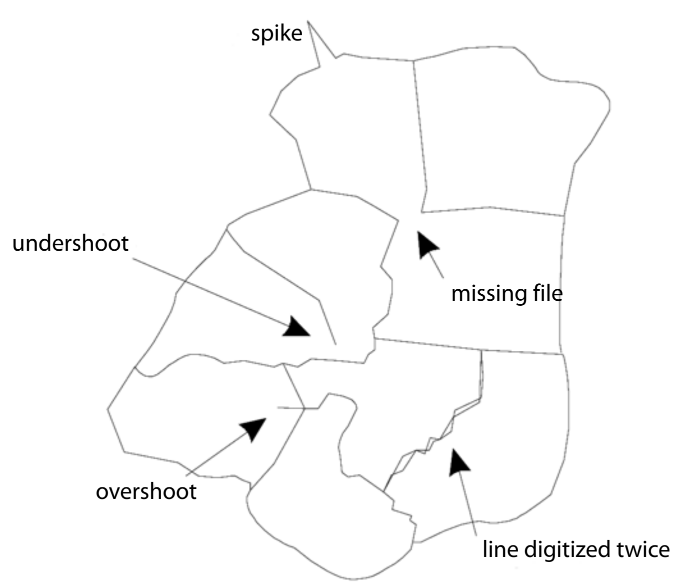
- Digitizing is tedious, possibly leading to operator fatigue and resulting quality problems that could require considerable post-processing.
- Manual digitizing is quite slow. Large-scale data conversion projects may thus require a large number of operators and digitizing tables.
- In contrast to primary data collection using GPS or aerial photography, the accuracy of digitized maps is limited by the quality of the source material.

(*c*) Editing

3.91. The objective in converting geographic information from analogue 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 III.12. When undershoots and missing lines occur, the net effect is to create one polygon when two areas actually exist. When lines are digitized twice, the net effect is to create one or more additional polygons when none exist. 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.

3.92. Some of the common digitizing errors shown in figure III.12 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 end-points 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.

Figure III.12
Some common digitizing errors



(*d*) Constructing and maintaining topology

3.93. 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 the 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 neighbours. 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 I 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 neighbours 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, para. A1.5).

3.94. 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.

(*e*) Digital map integration

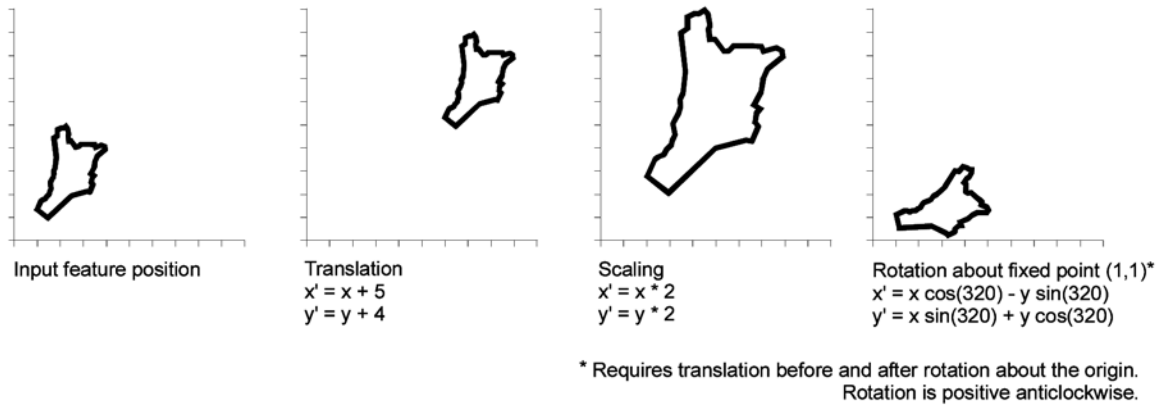
3.95. A census-mapping project should take advantage of all suitable geospatial data sources. These are likely to be stored in different formats, using varying map scales and cartographic projections. Integrating these heterogeneous data sources requires considerable knowledge of GIS data-integration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

(*f*) Georeferencing

3.96. Georeferencing is the process of assigning coordinates from a known reference system, such as latitude-longitude, to the page coordinates of a raster (image) map. 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 III.13):

- **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.
- **Rotation.** The geographic feature is rotated about the coordinate systems origin by a given angle. Rotation will make sure that the resulting digital

Figure III.13
Translation, scaling, rotation

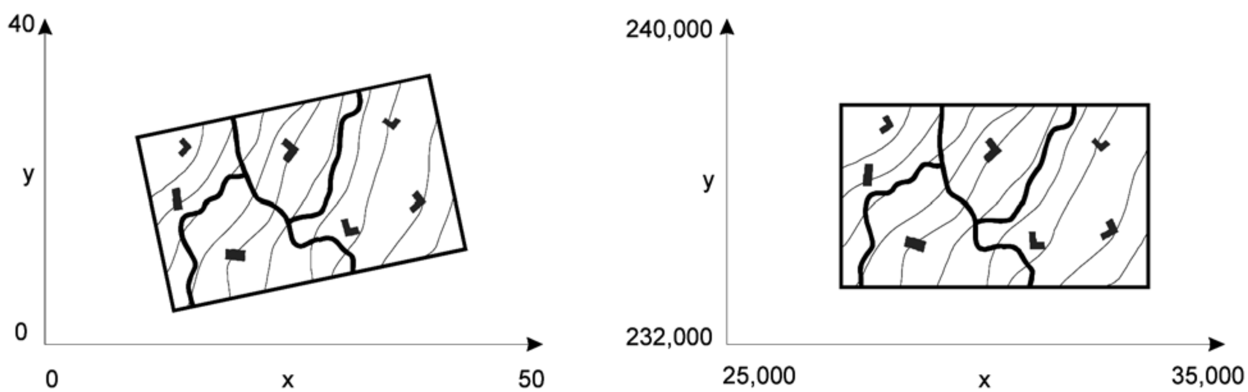


map has the proper orientation even if the paper map had not been correctly aligned on the digitizing board.

3.97. 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 III.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 provided in annex II.

3.98. 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

Figure III.14
Map in digitizing units; map in real-world coordinates



Input map in digitizing units
(centimetres)

Output map in real-world coordinates
(metres)

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 programmes is a good candidate) or to use what is known as rubber-sheeting.

3.99. Rubber-sheeting 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 rubber-sheeting is quite large and this operation should therefore be avoided if at all possible. In some instances, however, where the input maps clearly do not conform to any well-defined projection, rubber-sheeting is a viable way to make use of the available geographic information. A good example in the context of census-mapping is the georeferencing of hand-drawn sketch maps. Paragraph A2.6 in annex II provides a practical example of georeferencing that illustrates the process of converting, for instance, how to convert a digitized map into a properly referenced digital database.

3.100. Georeferenced layers for the same area or coordinates collected using a global positioning system will not be compatible with 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 measured in metres or feet (see also annex II). As noted above, this step can be done in most systems either at the start of digitizing or after spatial data automation has been completed.

(*g*) Projection and datum change

3.101. Projection change is necessary when maps that were digitized from different map sheets need to be assembled into a seamless database. It is related to the transformation process that converts the coordinates of digital map features without changing their shape. When converting from one projection to another, the shape and distortion of map features do change, although the changes may be all but imperceptible at large cartographic scales.

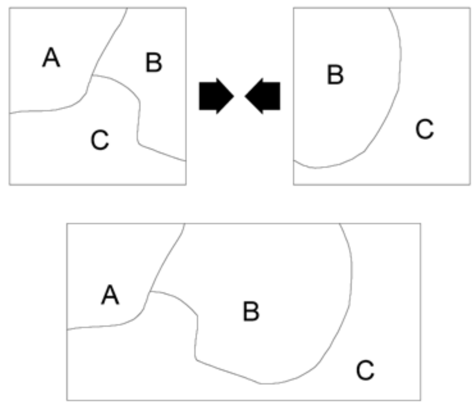
3.102. Maps issued at different map scales sometimes use different projections. In other instances, a mapping agency may have changed the standard projection used for mapping in the country, so that older map sheets may use a different projection from those map sheets that were revised more recently. 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.

3.103. Projections and geographic datums are discussed in more detail in annex II. It will be useful for the census-mapping agency to arrange for trained staff or experts from an outside agency to provide advice 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 programs provide the required projection change functions.

(*h*) Integration of separate map segments

3.104. 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 III.15).

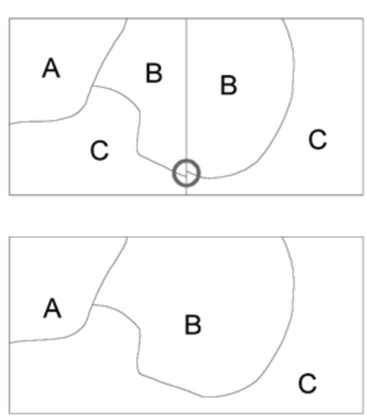
Figure III.15
Joining adjacent digital map sheets



3.105. This process is usually 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 (see figure III.16). Errors may 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 are represented by different symbols.

3.106. 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. It is advisable to be careful when mixing scales.

Figure III.16
Edge-matching after joining adjacent map sheets



3.107. The process of fixing these errors is called edge-matching. 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 edge-matching functions provided by some GIS packages.

D. Implementation of an EA database

3.108. 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 above as playing a key role, particularly in boundary and administrative area delineation.

3.109. 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.

3.110. 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, which when imported in a GIS allow them to "behave" predictably in simulated conditions and permit sophisticated modelling, such as travel-route analysis.

3.111. Geographic databases (hereafter referred to as geodatabases) are more than spreadsheets. Entity types can be defined as having specific properties that govern behaviour 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, nests within administrative units and is composed of population-based units.

1. Relational databases

3.112. 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 database management system (DBMS) is where databases are stored.

3.113. 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 such instances as "gas lamp," "sodium vapour lamp," and "mercury vapour lamp."

3.114. 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.

3.115. In a geographic context, an "entity" can be an administrative or census unit or any other spatial feature for which characteristics will be compiled. For example, an entity might represent the feature "enumeration area" (figure III.17). 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 III.17
Example of an entity table-enumeration area

Entity: Enumeration area
Types (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

3.116. Database management systems (DBMSs) can be divided into various types, including relational, object and object-relational. Relational database management systems (RDBMSs) 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 programmes include Microsoft Access and Oracle.

3.117. Object-database management systems (ODBMSs) are designed to address a central weakness in RDBMSs, namely their inability to store complete objects directly in the database. ODBMSs can store objects persistently and provide object-oriented query tools. Object-relational database management systems (ORDBMSs) are a hybrid object-relational database, consisting of a relational database 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.

3.118. Software companies have answered the need for spatial capability in their relational databases through the use of geographic DBMSs extensions. These are major DBMSs, with spatial database extensions. Examples include the IBM 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.

3.119. 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 programmes 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 behaviour.

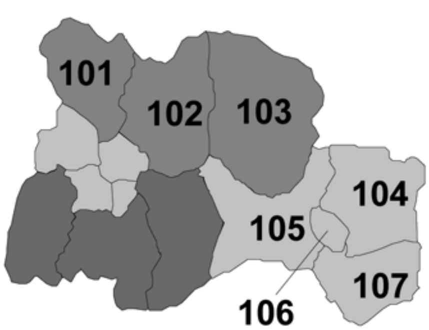
3.120. Relations define the association between entities. For instance, a table describing enumeration areas can be linked to a table for the entity crew leader area. This table has attributes, such as the name of the crew leader, the regional office responsible and contact information. The primary key 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.

3.121. 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 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.

3.122. Figure III.18 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

Figure III.18
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
...	...	...

only wastes storage space but 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.

3.123. Defining a clean database structure is not a trivial task. Some database management programmes 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.

2. Definition of database content (data-modelling)

3.124. 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 known as data-modelling 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.

3.125. It should be noted that many GIS databases are created without detailed data modelling. 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-modelling imposes a level of rigour 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 the process.

3.126. As discussed earlier, many national and international agencies have already been active in developing generic data models for spatial information as part of an 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.

3.127. Annex III provides an example of what a data model description in a data dictionary might look like. Related to the data model are both metadata standards, which are discussed below, and simpler database dictionaries that accompany databases distributed to the general public (see annex IV).

E. Data-quality issues

1. Accuracy requirements

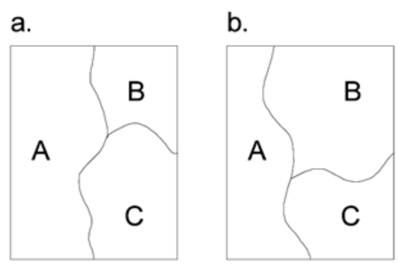
3.128. The development of acceptable data accuracy standards is perhaps one of the most important tasks in planning a digital database development project. In many fields, such as utilities and facilities management, terrain or hydrological mapping, accuracy database standards exist that can be adopted for any new project. Census-mapping, in contrast, has traditionally been carried out 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.

3.129. Accuracy in GIS refers to both attribute data — the geographic attribute table and the census data that can be attached to it — and 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.

3.130. 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. Or 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 look very different. For instance, in figure III.19, the two maps correctly represent the neighbourhood relationships between three administrative units.

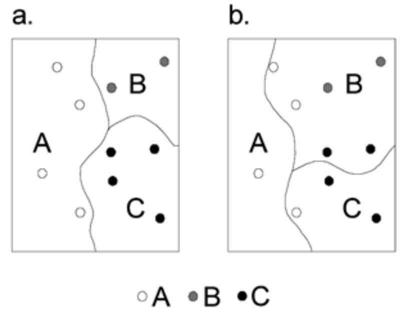
Figure III.19
Logical accuracy



3.131. 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. This 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.

3.132. 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 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 III.20
Problems if positional accuracy is not maintained



3.133. Figure III.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 III.20b, by 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.

3.134. 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 per cent 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 per cent requires a greater than proportional input in time and other resources. In fact, some estimates claim that increasing accuracy from 95 to 100 per cent would require 95 per cent of the total budget of a project (Hohl 1998).

3.135. 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 metres from their true position in y per cent 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.

3.136. Although a high degree of positional accuracy is desirable, accuracy standards that are too high 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 that 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.

3.137. 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.

2. Quality control

3.138. 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 for Population and Housing Censuses* (United Nations, 2008) stress the importance of quality control and contain an overview of these issues in the census process. These general concepts also apply to the construction of geographic databases.

3.139. Quality control must permeate the entire census process and it is no less important for geographic programmes. As a general strategy, the best programme designs include protocols to ensure quality control at every stage of the census process.

3.140. 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 geographic staff to limit errors at every step of the data-conversion process. Census staff should be encouraged to report problems in 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 not be afraid to report problems with their own work and clearly understand the overall objective of quality-control procedures.

3.141. 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.

3.142. 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. Statistical tests can be done to identify outliers. 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 geographic 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.

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

3.144. Quality-control steps should be documented thoroughly. A hard-copy log is generally the most appropriate means of documenting data quality, although automated digital logs can also be used. The log 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 but can also point out which staff members may require additional training.

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

3. Partitioning of national territory into processing units

3.146. A complete digital EA 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.

3.147. 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 work simultaneously on separate parts of the database.

4. The digital administrative base map

3.148. The choice of operational design depends on the processing as well as the organizational environment. 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 subdistricts as well. These boundaries should be of high accuracy and should show an amount of detail that 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.

3.149. 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 above about NSDIs). Codes used in the administrative base should obviously correspond to the codes used in the census database.

3.150. 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 neighbouring 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 neighbouring regional office or operator.

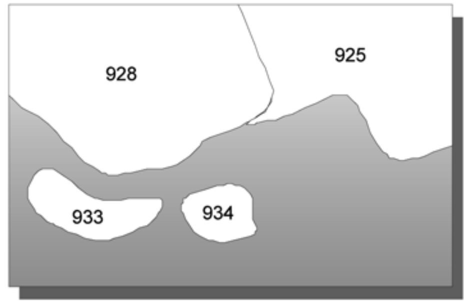
5. Dealing with disjoint area units

3.151. 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 to dealing with this problem.

3.152. 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 of 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.

3.153. 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 (see figure III.21). 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 III.21
Dealing with administrative units consisting of several polygons



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

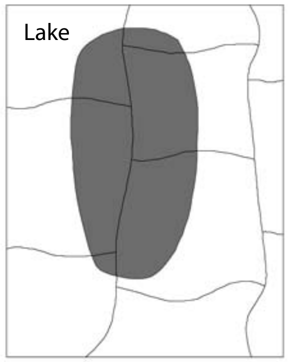
6. Computing areas

3.154. 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 owing 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.

3.155. 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 neighboured a large lake. The reported areas included the portion of the districts that extended from the lake shore to the centre line of the lake (see figure III.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 programmes, the definition of population density can have severe consequences.

3.156. In countries where this is a problem, the census office may decide to report two area fields: one that is the "total area" of an administrative unit and one that 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 avail-

Figure III.22
A lake covering a large area in several administrative units



able 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 for the definitions of the net areas to be well documented.

3.157. 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.

F. Metadata development

3.158. The present *Handbook* argues that the construction of a geographic database should be considered as a long-term process and 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 institutional memory needs to be based on more than the recollection of the geospatial analysts involved in initial data development. The detailed documentation of all steps involved in developing the digital spatial census database is therefore mandatory.

3.159. 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.
- To support the integration of externally produced data sets into an organization's data holdings.

3.160. What different data producers consider essential metadata can differ widely. Many countries have therefore started the development of general geographic metadata standards. These aim to unify 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/TC211) 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).

3.161. 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 geospatial 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 interdepartmental 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.

3.162. One example of a metadata standard is the Content Standards for Digital Geospatial Metadata (CSDGM) developed by the United States Federal Geographic Data Committee (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 United States Census Bureau (http://www.census.gov/geo/www/standards/scdd/ (see also United States Federal Geographic Data Committee 1997b). Here, we will only discuss the main components of the metadata definition.

3.163. 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 and must 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 labelled "mandatory if applicable" or "optional".

3.164. The main components of the standard are the following:

- **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 in paragraph 3.127 above.

- Distribution information include the data distributor, file format of data, off-line media types, online link to data, fees and order process.
- Metadata reference information provides information about the metadata itself, most importantly who created the metadata and when.

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

- Citation information ensures consistent referencing of the originator, title, publication date, publisher.
- Time period information includes single date, multiple dates or range of dates.
- Contact information, such as contact person and/or organization, address, phone and e-mail.

3.166. 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.

3.167. 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.

G. Summary and conclusions

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

3.169. 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.

3.170. The final task is to shore up the decision to produce an EA-level database for census-canvassing operations. This will make the census more accurate and provide far better inputs for subsequent analysis and dissemination. It will also support the post-census dissemination of results, especially for humanitarian applications (e.g., disaster preparedness and management), which is covered in chapter VI.

Chapter IV
Integrating fieldwork using GPS and remotely-sensed data

4.1. This chapter continues the step-by-step process of constructing the EA-level geographic database that was introduced in chapter III. Here we recognize the value of new tools and data sources made possible by satellite technology — namely global positioning systems and remote sensing (including aerial photography) — by addressing the new tools and data sources directly.

4.2. The main topic for chapter IV is using GPS and R/S for EA delineation. The main reason to do this is to field-validate the EA boundaries that were created in the NSO's GIS lab from the prior census's maps. Or, when accurate maps are not available, it is done as a basis for EA delineation in the census main office, before conducting fieldwork for completion and validation. With the help of remotely sensed data, geographic analysts can identify the 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 collective living quarters) and extraction of other features. Mobile computing and hand-held computers will be covered. Remote sensing will be addressed in terms of both satellite imagery and aerial photography, presenting some basics and guidelines for NSO use.

4.3. Ideally, at this stage in the process, the statistical agency has scanned the EA maps from the previous census and transformed them into a digital geographic 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. The present *Handbook* assumes that headquarters and field operations will be integrated through field-based operations and the sharing of data.

4.4. Globally, the overriding goal of geographic digitization is to harness new technologies to make better maps faster and improve census data quality overall. Once again, we retain our focus on the country's particular conditions, paying attention to how territory is divided into administrative regions and the EAs that the statistical agency uses for the census. Through the 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.

4.5. Remote sensing 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 has answered the question of the specific use of the data, only then can impacts be assessed in terms of resources, especially human resources.

A. Global positioning systems (GPS)

4.6. Once an exotic gadget, 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 has gained widespread personal use, with big gains in the consumer market in cars, boats, construction and farm equipment, and built into hand-held 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 such areas as biology, forestry and geology, and also finds increasing applications in epidemiology and population studies. GPS is also becoming a major tool in census cartographic applications.

4.7. Most of the discussion refers to the United States 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 Galileo system and the Chinese Beidou system are also briefly introduced and discussed below.

1. How GPS works

4.8. 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 3 spares — and their ground stations. The system, which is called NAVSTAR, is maintained by the United States 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, 5 to 8 GPS satellites are within the “field of view” of a GPS receiver on the Earth’s surface.

4.9. 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.

4.10. 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 where the two circles intersect. To confirm our exact position, we need to 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.

2. GPS accuracy

4.11. Inexpensive GPS receivers can provide reasonably accurate information about the latitude, longitude and altitude of the user's 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 metres for civilian applications. Altitude information is somewhat less reliable than latitude and longitude.

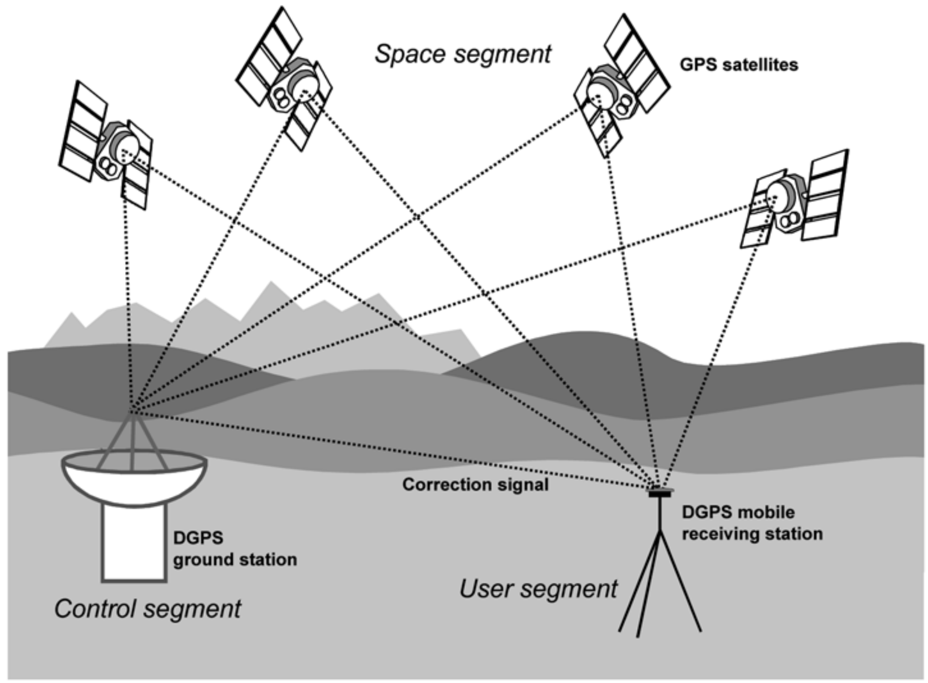
4.12. Accuracy is influenced by several factors. One of these is the number and position of the satellites. Ideally, these are spread out over the sky to allow optimal geometric computation. Positional dilution of precision refers to the spread of satellites in the sky and can be quantified with a number. 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 the 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, instituted by the United States Department of Defense to reduce accuracy in the signal. Though selective availability has been discontinued, it could be reinstated during a time of war.

4.13. 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.

3. Differential GPS

4.14. For applications requiring higher accuracy, differential GPS systems (DGPS) systems use correction information transmitted from a base station with precisely known coordinates to correct the satellite signals (see figure IV.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

Figure IV.1
Differential GPS



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-metre accuracy.

4.15. 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.

4.16. 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 receiver's 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.

4.17. Some new national systems worthy of mention make use of new satellite technology and communications over the Internet. The Wide Area Augmentation System or WAAS is a continental DGPS system instituted by the United States Federal Aviation Administration. 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 United States. Continuously operating reference stations or CORSs, 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 centimetres. This service is at the time of publication available only in North America, although at the time of this edition a similar system may be soon available for Africa.

4.18. Governments in other regions are developing similar satellite-based differential systems. In Asia, the Japanese are developing a Multi-Functional Satellite Augmentation System or MSAS, 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.

4. Other global satellite navigation systems

4.19. Several alternatives to the United States NAVSTAR GPS system exist. A more generic term for such systems is global navigation systems or GNS. The Russian counterpart of GPS is the GLONASS system, which is operated by the Ministry of Defence 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 Government of India will see two satellites launched from Indian territory in return for access to high-precision signals.

4.20. The European Union 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 United States-operated 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 inter-operable with the United States GPS system at the user level.

4.21. 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.

4.22. 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.

4.23. Most vendors offer integrated products that combine a GPS receiver with a hand-held 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 below, along with integrated field-mapping systems.

5. GPS in census-mapping applications

4.24. GPS technology offers many applications in mapping activities, including the preparation and correction of enumerator maps for census activities. As noted above, we must emphasize the importance of fitting the use of new technology into a robust, detailed overall plan. With DGPS, geographical positions of EA boundaries can be corrected and the location of point features, such as service facilities or village centres, 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. Guidelines for demarcating EAs in the field are set out below, followed by examples of specific census operations that can make use of GPS (for more information on EA delineation, see chap. III).

4.25. As we know, EAs 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 such features as roads and water bodies for 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.

4.26. When delineating EAs, 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 pre-test to determine the number of days needed for the enumeration.

4.27. Population estimates represent the most critical element of EA delineation. 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. The failure to obtain good estimates for EA populations in advance will hamper enumeration and threaten the quality of the results.

4.28. 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.

4.29. 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 and 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 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.

4.30. 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 travelling 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.

6. Some specific GPS-related mapping tasks

4.31. Some specific GPS-related mapping tasks include:

(*a*) **EA boundary delineation.** 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 probably 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;

(*b*) **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 carefully weigh the potential benefit of conducting detailed administrative boundaries at the time of the census against the cost in time and labour. 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. Enquiries should be made into accompanying metadata, including datum and projection information, before trying to use these data in a GIS project;

(*c*) **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 duplication of work;

(*d*) **Collective living quarter locations.** Collective living quarters 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 humanitarian planners are increasingly requesting the geographic locations of collective living quarters in order to effectively plan disaster response. Moreover, locating collective living quarters with GPS units may be less taxing — since they are fewer in number — than measuring all housing units in the country;

(*e*) **Other relevant features (including roads).** Such features 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.

4.32. For census applications, the 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 fieldworkers would probably be beyond the resources of a census project.

4.33. 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. In this way, line features can be recorded automatically by walking along a road or travelling 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. GPS units cannot operate without battery power, so adequate supply, including multiple backups, should be ensured. Other functionalities of the GPS units should be locked down so that field personnel cannot reprogram them or change their settings. Finally, units should be marked clearly so that if stolen they are not easily resold.

7. Training requirements for GPS use

4.34. For a GPS project to be successful, the NSO must manage equipment purchases carefully, arrange for training and personnel needs and 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 fieldworkers to make sure they work consistently. A training programme for fieldworkers can include understanding the operation of the GPS units, as well as how a receiver calculates a position and troubleshooting problems with the units.

4.35. 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.

4.36. 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.

4.37. 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 in the format of a correctly formatted table or as "event data."

8. Summary: advantages and disadvantages of GPS

4.38. The advantages of GPS include the following:

- Fairly inexpensive, easy-to-use component of 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.
- New GPS systems will be coming online in the next five years.

4.39. The disadvantages are as follows:

- As a physical component, GPS is cheap, but field operations using GPS can be quite time-consuming (and therefore expensive) if not properly planned.
- Comprehensive planning includes determining what products will result from extensive GPS use.
- Signal may be obstructed in dense urban or wooded areas (multi-path error).
- 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.
- The more complex the GPS unit measured, the more training is required.

4.40. In the application of GPS, problems can 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 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.

B. Integrated field-mapping systems using hand-held computers

4.41. New technology that combines personal computer functionality with GPS has made inroads in census operations in some countries. One advantage of using hand-held computers or personal digital assistants (PDAs) is that of "direct capture" — the ability to record information directly without transcription, removing several intermediate steps. Maps can be updated immediately. 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 geographic database. This information can then be incorporated into the geographic database at the home office. Given that notebook computers and other portable computing devices are becoming less expensive, integrated field-mapping systems are becoming a viable option for field data collection for census purposes.

4.42. Advances in technology, including GPS, wireless communication and computer miniaturization, have made possible numerous new applications for hand-held GIS, particularly the development of specialized 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, and 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 hand-held computers includes "thin" versions of popular office applications. Examples of GIS software used for hand-held computers include Autodesk OnSite, ESRI ArcPad and Intergraph Intelliwhere.

4.43. NSOs contemplating the use of hand-held computers for either pre-census-mapping or the actual enumeration must consider the cost. GPS-equipped hand-held computers 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 onscreen map readability, power requirements, especially in areas without dependable electricity, and other adverse environmental conditions for computers, even ruggedized ones.

C. Satellite remote sensing

1. Using imagery to field-verify EA maps produced at census headquarters

4.44. Since the publication of the 2000 *Handbook*, 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. The time is ripe 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. R/S data can be used as an independent check on the field-verification process. The present *Handbook* advocates an approach that performs triage on the surface area of a country, partitioning it into areas needing more and less attention.

Box IV.1

Case study of GPS experience: Fiji

Fiji's 2007 census was the first in the Pacific region to use GPS technology to link census questionnaires with georeferenced locations for all households in the country. In Fiji, as 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), thus 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.

About 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; and 200 GPS operators were trained over a three-week period. GPS operators and supervisors were given a "cheat" sheet, 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 so 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 first put a sticker with a unique 6-digit code on the house where the interview was conducted and also one on the gate for fenced houses. A small sticker with the same number was placed on the form for that household. A reserve sticker was placed on the gate for households where the front door and gate were far apart. That facilitated waypoint taking. Second, the GPS operator visited the household and keyed in the same number as waypoint identification so as to later link it to the questionnaire.

Once fieldwork was completed, two databases were created for the waypoints, one 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 be learned for future censuses:

- In the early stages of the GPS point collection, operators were taking waypoints before acceptable precision levels were obtained. To solve this problem, waiting time was increased.
- Loss of stickers due to two religious festivals (where houses were 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 was the fact that the GPS waypoint- taking lagged behind the enumeration, sometimes by months. If the enumeration team is also taking the waypoints then those errors would be eliminated.

One of the expected 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.

Source: Fiji presentation at a workshop held in Nouméa, 2008.

Figure IV.2
EA boundaries delineated over a panchromatic satellite image

This is what is meant by a "change-detection" approach, which is especially useful for establishing the perimeters of populated areas.

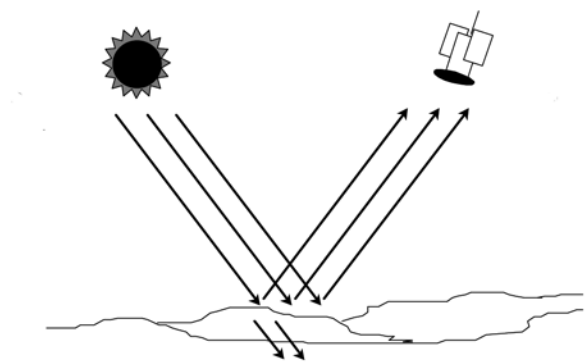
4.45. Following the schema introduced in chapter III, at this stage the NSO has digitized old census EA maps and overlaid other geographic data and has entered all of it into a geodatabase, but it has not yet corrected the results in the field. With the provisional EA boundaries superimposed on R/S imagery (see figure IV.2), population settlements can be quickly located and priority areas identified. Planning for such activities must, however, be detailed and realistic.

2. Principles of satellite remote sensing

4.46. 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's 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.

4.47. 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 (see figures IV.3 and

Figure IV.3
The remote-sensing process

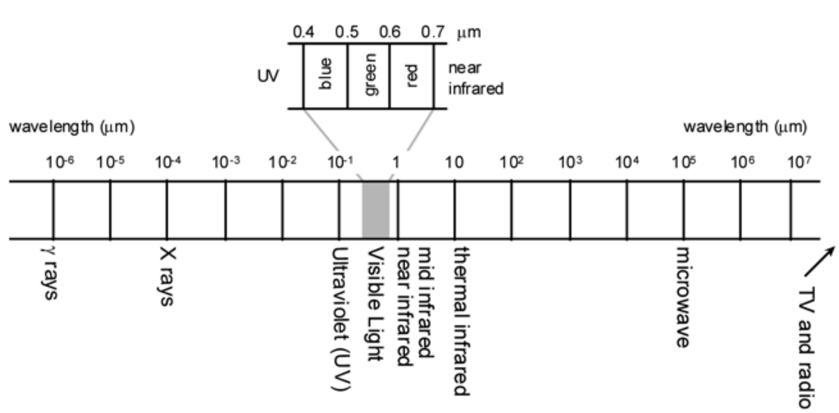


IV.4). Most satellite data collection is considered passive, in that it receives emitted energy from the Earth, contrasting with active sensors, such as radar, which incidentally can also penetrate cloud cover. Satellite systems do not use photographic film to record the reflected energy. Instead, an electro-optical detector array — similar to a charge-coupled device camera — measures the intensity of electromagnetic radiation and records it digitally as a regular raster or image of rows and columns.

4.48. Satellite sensors operate in multi-spectral and 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.

4.49. 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 receiving stations

Figure IV.4
The electromagnetic spectrum



on Earth, where they are radiometrically and geometrically corrected and georeferenced. The resulting digital or printed images can be interpreted visually, in a similar way to air-photo interpretation, or they can be analysed 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.

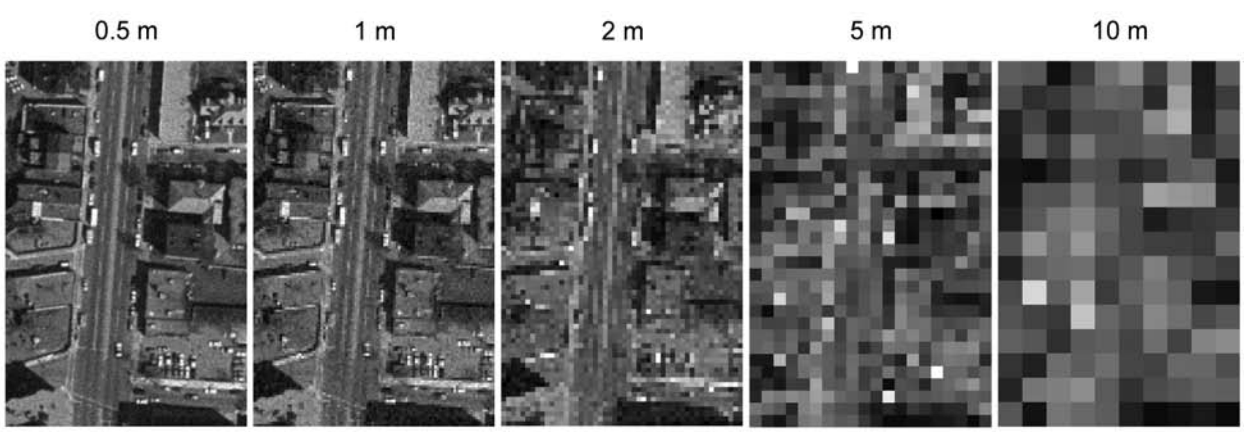
3. Resolution of remote-sensing data

4.50. 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-1-metre of the most popular high-resolution systems, such as Quickbird and Ikonos. Indian Remote Sensing, 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. As of March 2008, the remote sensing company GeoEye plans a launch of a sensor with a pixel size of 0.41 metres.

4.51. Figure IV.5 compares pixel sizes that were simulated from a 0.5 metre resolution digital air photo by aggregation. The image covers an area on the ground of 100 x 150 metres. 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.

4.52. Among the varieties of imagery available, from low spatial resolution/high temporal resolution METEOSAT and 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.

Figure IV.5
Illustration of pixel size in aerial photographs and satellite images



4.53. 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, making the prospect of complete coverage expensive.

4.54. With a spatial resolution of 30m or 15m for the Enhanced Thematic Mapper (ETM), Landsat can identify linear features, such as roads and rivers, and other ancillary layers, such as lakes and other water bodies. Landsat has an important additional advantage, in that it is freely available. In some cases, Landsat 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.

4.55. Most 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 georeferencing with or without ground control points. For example, archived imagery available from Digital Globe (as of March 2008) costs US $16 per square kilometre (km^2), with a minimum order of 25 square kilometres. Ikonos imagery tends to cost less, about $8 per km^2. Prices will depend on the amount of area being covered in the imagery purchase, with larger areas costing less per km^2. Raw image data will be considerably less expensive than a digital orthophoto map produced from satellite images. Normally, however, images are purchased fully processed. A listing of very high spatial resolution civilian satellite remote-sensing products is provided in table IV.1. The Office of Outer Space Affairs of the United Nations Secretariat maintains a more comprehensive list (see www.oosa.unvienna.org). Other sensors include ALOS, the Japanese system used in the Americas, the Alaska Satellite Facility (ASF) and CBS26, and IMPE, the China-Brazil system which is free for Africa.

Table IV.1
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

4.56. 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.

4. Online sources of satellite remote-sensing data

4.57. 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 such applications as Google Earth, ArcGIS Explorer, 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.

4.58. Google Earth is a virtual globe program that maps the Earth by cataloguing and displaying satellite imagery. Google Earth has made an impact on the geospatial community 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. Users cannot access the actual source of the data, they can only view it. However, they can add their own data.

4.59. 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 as of March 2008; check for most recent pricing). 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 e-mail and a data importer that can read address points from a spreadsheet, using comma-separated values, but this is limited to 100 points/addresses.

4.60. The functionality of the Pro version includes add-on software, such as 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.

4.61. 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, such as 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.

4.62. 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 stand-alone imagery. The 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 but not enough to count housing units. The weaknesses of Google Earth include the lack of necessary resolution for use with EA delineation, the difficulty in transferring imagery into a GIS program, the need for a high-speed Internet connection to download imagery, and metadata and authenticity. Other online satellite data sources, such as those obtained through the Environmental Systems Research Institute (ESRI) free ArcGIS Explorer application, may circumvent some issues with Google Earth by allowing the direct import of imagery into GIS projects. But the spatial resolution of the available imagery may still not be sufficient for some census applications.

4.63. 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 that 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 (see paras. 4.73-4.89 below).

5. Applications of remote-sensing data for population analysis

4.64. R/S techniques have the potential to identify areas of rapid growth or change, allowing NSOs to concentrate resources where they are most needed. Population analysis using remote sensing is in its infancy, but gains are rapidly being made. A United States 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 discussed, Landsat is problematic for estimating population size and also has long-term problems with its sensor, in addition to funding variability. Applications to replace Landsat with other high-resolution data, such as SPOT, are likely to be available to researchers in the near future.

4.65. 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 substitutes for residential spread and population density (Jensen and Cowen, 1999). In some disaster-prone areas, aerial photos have an advantage over satellite imagery in their ability to capture scenes below cloud cover. Radar has yet to play a significant role in population analysis although it has an advantage in that it can penetrate clouds.

4.66. A "change-detection" approach can be used to measure population change spatially, in particular to identify areas of rapid growth, using two or more images of the same place over a time period of five or more years. For quantifying the spread of urbanization, analysts categorize each image using a hard classification technique, so that 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 populated areas.

4.67. 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 km^2 per year. This increased to 25 km^2 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 build-up 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 argues for a continuous approach in using GIS for census-mapping operations.

4.68. Health studies have used remotely sensed imagery to focus on intra-urban disparities in such phenomena as disease prevalence. Castro (2004) used aerial photography and neighbourhood polygons to identify potentially harmful malaria breeding sites. Detection of informal settlements and neighbourhoods 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 neighbourhoods tend to have minimal texture, meaning low variability in brightness, and high concentrations of impervious surfaces (Weeks, 2007).

4.69. 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 colour, then labelled each area as a particular surface type, such as a "rooftop". Such a technique would then allow all rooftops of a particular type to be grouped and labelled as part of an informal settlement. Applying the technique to several study areas, Pellika was able to achieve 97 per cent 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.

4.70. As with air photos, the 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.

6. Advantages and disadvantages of satellite remote-sensing data

4.71. The advantages of satellite remote-sensing data include the following:

- 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 areas at a level of detail sufficient for EA delineation provided that population estimates exist for the areas delineated.
- Imagery can permit the mapping of inaccessible areas.
- Imagery can serve as an independent check on field verification.
- 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.
- Update of topographic maps in rural areas is possible, e.g., identification of new settlements or villages that are missing on maps.

4.72. The disadvantages are as follows:

- The spatial resolution of many systems, especially low-cost ones, is not sufficient for census applications.
- In the case of optical sensors, cloud and vegetation cover restrict 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.
- Image-processing requires a large amount of expertise that 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 their mapping tasks. Using such an institutional collaboration approach, expenses and expertise can be shared.

D. Aerial photography

1. Aerial photography overview

4.73. Even with the rise of high-resolution satellite imagery, aerial photography continues to be useful for mapping applications that require high accuracy and rapid completion of tasks. Aerial photographs are similar to maps and satellite images because 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, aerial photos also do not provide the geometrical accuracy of a map. Camera angle and terrain variation distort the view of an aerial 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 IV.2).

4.74. 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 support many aspects of urban and regional planning. Census projects have also frequently made use of aerial 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 reasonably short time frame.

4.75. Aerial photos have been used for mapping since shortly after the invention of airplanes. Early applications made use of standard cameras. Very soon, however, customized camera systems that minimize geometric distortion were mounted on specially adapted airplanes that allow 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, the interpretation of stereo pairs of images became the dominant method for producing maps of elevation contours.

4.76. Aerial photography is obtained using specialized cameras on board low-flying planes. The camera captures the image either on photographic film or digitally. In comparison with digital sensor systems, film has traditionally 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 1½ inches per image pixel, with 4,096 gray-scale shades versus 256 from film. South African company Rob Wooding and Associates compare digital aerial imagery favourably with 1m satellite imagery, finding the aerial imagery to be both less expensive and more precise. This will depend on the location of the area flown, so needs should be evaluated on a case-by-case basis.

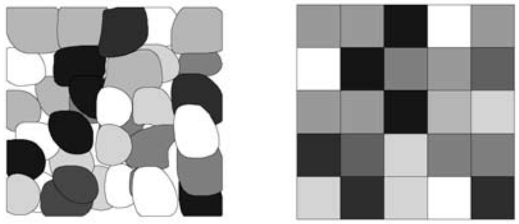
4.77. Traditionally, the end-products of an aerial photography project are printed photos of an area on the ground. The air photo survey is designed so that the resulting photos overlap by between 30 and 60 per cent. The photogrammetrist can combine these photos to produce a seamless mosaic, covering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

4.78. Recent advances in digital image-processing have changed the format in which aerial images are transformed into useful products. In analogue 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 IV.6. 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).

2. Application of aerial photos for census-mapping

4.79. 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-storey 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

Figure IV.6
Photographic film versus the scanned image



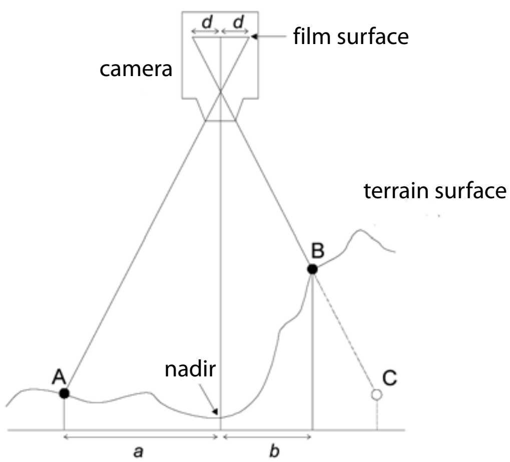
Box IV.2

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 IV.7 (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 centre 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 IV.7

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 soft-copy mapping systems support a high degree of automation for the 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, wire-frame 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 after distortion removal, the initial air photos will 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. Neighbouring orthophotos can be digitally combined to create seamless image databases for an entire city or region or even a whole country. Mapping technicians can extract or delineate features on these orthophoto maps through onscreen digitizing. Or they can simply be used as a backdrop to provide a context for existing GIS data layers.

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 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.

4.80. 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 a mouse or a similar pointing device. 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).

4.81. It must be stressed that an EA boundary file that is based on either satellite remote sensing or aerial photos is meaningless without adequate reference data — such as landmarks and street names — because enumerators may not be able to locate themselves.

3. Implementation and institutional issues with aerial photography

4.82. 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 agency, 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 international mapping companies which can provide the airplane, camera and processing equipment.

4.83. These services are not cheap, however. Fortunately, air photos are useful for many different applications, including planning service provisions, updating 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. One example is the use of aerial photography to estimate the number of people living on boats by the Statistical Office of Hong Kong, China (see Netherlands Interdisciplinary Demographic Institute, 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.

4.84. As described above, the development of orthophoto maps requires considerable technical expertise and specialized equipment. By contrast, the use of orthophoto maps does not require significant additional training, although it should be compatible with the overall census plan. 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 (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.

4.85. 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 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, remotely sensed imagery is more likely to be built into a GIS project than included as a separate map.

4.86. 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 centimetre range (usually 5-30 cm). Resampled digital orthophoto images, with pixel sizes between 0.5 and 2 metres, are sufficient for delineating EAs in urban areas.

4.87. 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 cm. 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.

4. Advantages and disadvantages of aerial photos

4.88. The advantages of aerial photos include the following:

- Aerial photos provide a large amount of detail and can be interpreted visually. Information about many types of features — roads, rivers and buildings — is shown concurrently.
- Data collection is faster and map data can therefore be produced much more quickly than using cartographic ground surveys. Recent aerial photos are therefore a more reliable basis for census-mapping compared with maps that are updated infrequently.
- Aerial photos can be used to produce maps for hard-to-reach areas or areas in which fieldwork is difficult or dangerous.
- Topographic mapping using aerial photography can be less expensive 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.
- Printed aerial photos are useful in fieldwork 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.

4.89. The disadvantages are as follows:

- Aerial photo-processing requires expensive equipment and specialized expertise. Census offices therefore need to collaborate with other agencies to gain access to orthophotos or otherwise rely on outside support.
- Aerial photos may be copyrights, with limited distribution rights.
- Aerial photos still require information on the names of features which need to be extracted from possibly outdated maps. Aerial photography does not necessarily make fieldwork redundant. It will probably not be adequate for remote areas, although it may provide backup for hard-to-enumerate areas.
- Aerial 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 terrain).
- Digital aerial photos consist of very large amounts of digital data and therefore require fairly powerful computers for display and further processing.

E. Summary and conclusions

4.90. Chapter IV has reviewed 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 V
Use of geographic databases (maps) during the census

A. Introduction: use of geospatial tools during census enumeration

5.1. Maps are used for all kinds of planning purposes immediately before and during a census. Activities that use maps include the allocation of enumerators to territory; the identification of rugged or inaccessible areas; managing logistics for the transportation of field staff and supplies; locating hard-to-enumerate populations and collective living quarters; delineating administrative boundaries at multiple levels; monitoring census progress; and creating locator maps.

5.2. The present chapter will cover the process of creating such census maps, focusing on how to use the versatility of the geographic database to get the right information into the hands of enumerators in a timely and well organized manner. More broadly, it spells out how geospatial technology can support census operations during the enumeration phase. Topics covered here include map compilation, determining the relevant layers for enumerators and supervisors and the basic elements of printing and distribution.

5.3. Overall, the present chapter will adopt a project management approach to enumeration, which is plan-centric. With the logistics planned in detail, error is minimized and slowdowns averted. The census is a territorial exercise, in which the country is divided into operational units that can be canvassed. A digital enumerator map has the virtue of being modifiable for the specific context of census-taking in the country. Building EAs digitally allows the NSO to have a living document, building on the work of previous censuses with additional value added from remote sensing and GPS.

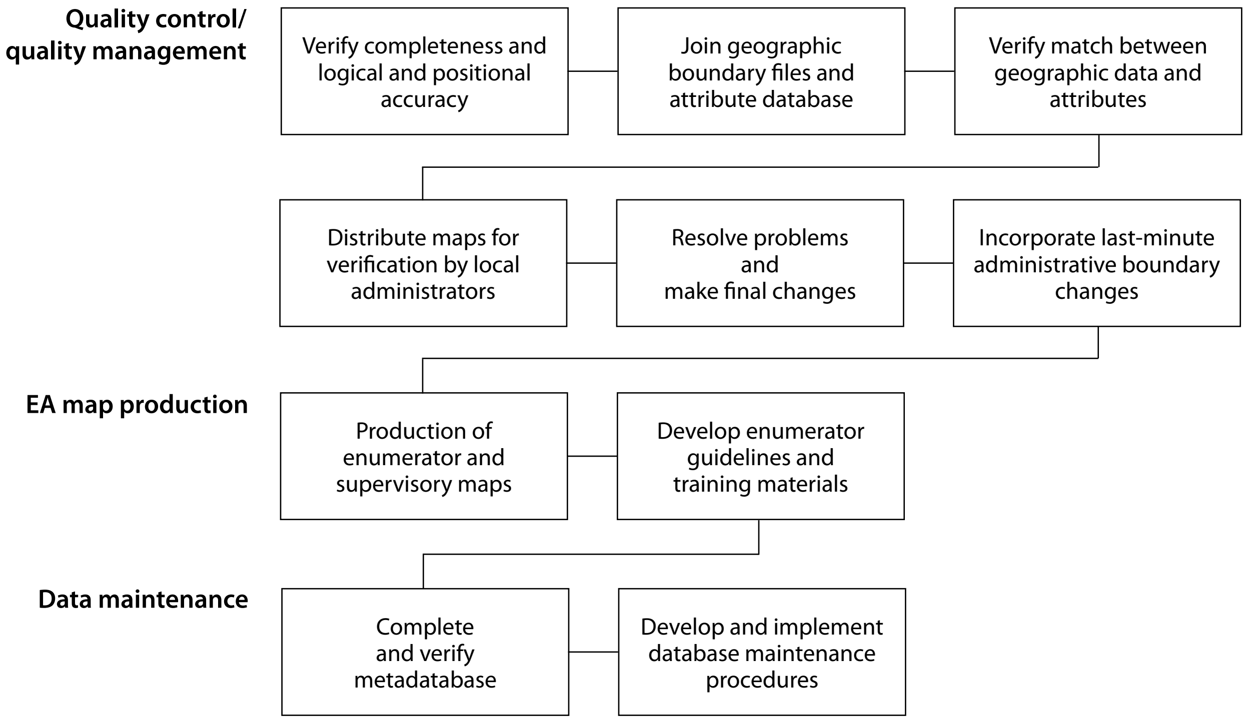
5.4. By this stage of the process, map compilation from previous censuses and all fieldwork has found a place in georeferenced form. Census planners have identified the areas in need of the most attention, following the change detection approach introduced in chapter IV. If it has done its work correctly, the NSO will have everything it needs at this point except the count. The philosophy here is that a thoroughly edited and updated database will become the basis for the maps that enumerators take into the field. For the purpose of the enumeration, then, the task is to get the maps out of the EA database and into the hands of enumerators.

B. Quality assurance, EA map production and database maintenance

1. Overview

5.5. The accuracy and completeness of census data depend substantially on the quality of the cartographic base maps used by enumerators. In addition to a con-

Figure V.1
Stages in quality assurance, output production and database maintenance



tinuous 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.

5.6. 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 the estimation of workloads and travel costs; making arrangements for the distribution and receipt of materials; the identification of trouble spots; and the arrangement of visits by office personnel to field locations.

5.7. Production of EA maps is conceptually straightforward provided that the quality of the digital geographic 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.

2. Draft map production and quality assurance procedures

5.8. The following sections walk through the process, creating census maps for the enumeration. Steps include the integration of geographic databases, the compilation of map features, verification by local authorities, and map printing and distribution.

3. Match boundaries and attribute files and print overview maps

5.9. 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 geographic 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.

5.10. 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 to ensure complete coverage. Errors are minimized when EA boundaries use the centre-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 may be found in chapters III and IV.

5.11. Once the geographic data and attribute information are correctly matched, labels need to be added to the map and map symbols chosen to identify features on the base maps (for information dissemination through maps, see also chap. VI). Labelling can be done interactively (manually), semi-automatically or automatically, using a GIS package or a more specialized cartographic design software. In a very large census map-production project, the labelling of features will be a time-consuming and tedious task. When EA map design is quite complex—for example, when 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 especially large.

5.12. Most GIS and desktop mapping systems provide functions for automated label placement. The user simply specifies the field in the GIS database attribute table that should be used for labelling, 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.

5.13. For very large EA mapping programmes, the census office might consider purchasing a specialized name placement software package, such as Maplex. Such software has 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.

5.14. The packages base label placement on a number of heuristic rules that 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.

4. Quality assurance

5.15. Although much consistency checking can be done interactively on computer screens, final quality assurance is best performed using printed hard-copy 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.

5.16 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 is discussed in chapter III. 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.

5.17. 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 neighbouring 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, as described below. 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.

5.18. 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 which feature a text label refers to. In some cases, arrows may be necessary to clarify the assignment.
- **Data layers.** 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 may require an inset or a separate map to ensure that all details can be identified.
- **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. This includes the date the map was produced so that enumerators can determine whether data is up to date or further field corrections need to be made.

5. Verification by local authorities and final administrative unit check

5.19. As a critical consistency check, the printed EA maps should be sent to local authorities for verification. Local administrators—inside and outside the census administration—should 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 reduce the risk of errors of map interpretation by locally recruited enumerators.

5.20. 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 can sometimes pose problems for the NSO, which needs to produce summary statistics for these units. The procedures for managing administrative boundaries are covered in chapter III. Ideally, administrative boundaries should be 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. Options for dealing with this problem closer to the time of the enumeration include:

(*a*) If administrative boundary changes continue to occur, then continuous tracking before the census is an 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 will require additional resources. The NSO should consider the use of a spatial database engine to record the date of establishment of the various administrative boundaries;

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

(*c*) 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 or other boundaries, the household questionnaires for these units must be reassigned to the correct units. This introduces an additional step after enumeration and may therefore delay the dissemination of census results.

6. EA map production (including map printing)

5.21. After the 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 may 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:

(*a*) The entire enumeration area, defined by a clearly indicated boundary line;

(*b*) Some parts of the neighbouring areas (i.e., the surrounding areas) to facilitate orientation;

(*c*) Any geographic and text information contained in the census cartographic database that will facilitate orientation within the EA, using standard cartographic conventions for symbolization (for example, used dashed lines to indicate paths, blue colouring for water etc.). Features for EA maps should include:

(i) Streets and roads;

(ii) Buildings;

(iii) Landmarks;

(iv) Hydrological features;

(v) Other notable or relevant features, possibly including topography, water bodies etc.;

(vi) A consistent map legend or map key, 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.

5.22. Figure V.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 the EA boundaries layer. In addition, annotation and 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 neat-lines (a box marking the outward extent of the map) and a legend that is used consistently for all EAs. Figure V.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.

5.23. In many countries that are not fully digital for the upcoming census round, EA map design may be simpler than that illustrated in figure V.3. 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 centrelines for the streets and polygons for entire city blocks rather than for individual houses.

5.24. Decisions must be made concerning format and colour of the printed EA maps. (for printer selection criteria, see chapter VI.) Given the high resolution available on laser printers, EA maps should produced on A3 (420 x 297 mm, the size of two A4 sheets) or 11 × 14 inch paper, if possible. Compared to larger-format printers or plotters, standard-sized printers have the advantages of lower cost and higher output speed. 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.

5.25. 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 simply 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 use software that creates a barcode for each map produced (for more information, see www.barcodehq.com/primer).

Figure V.2
Sample components of a digital EA map

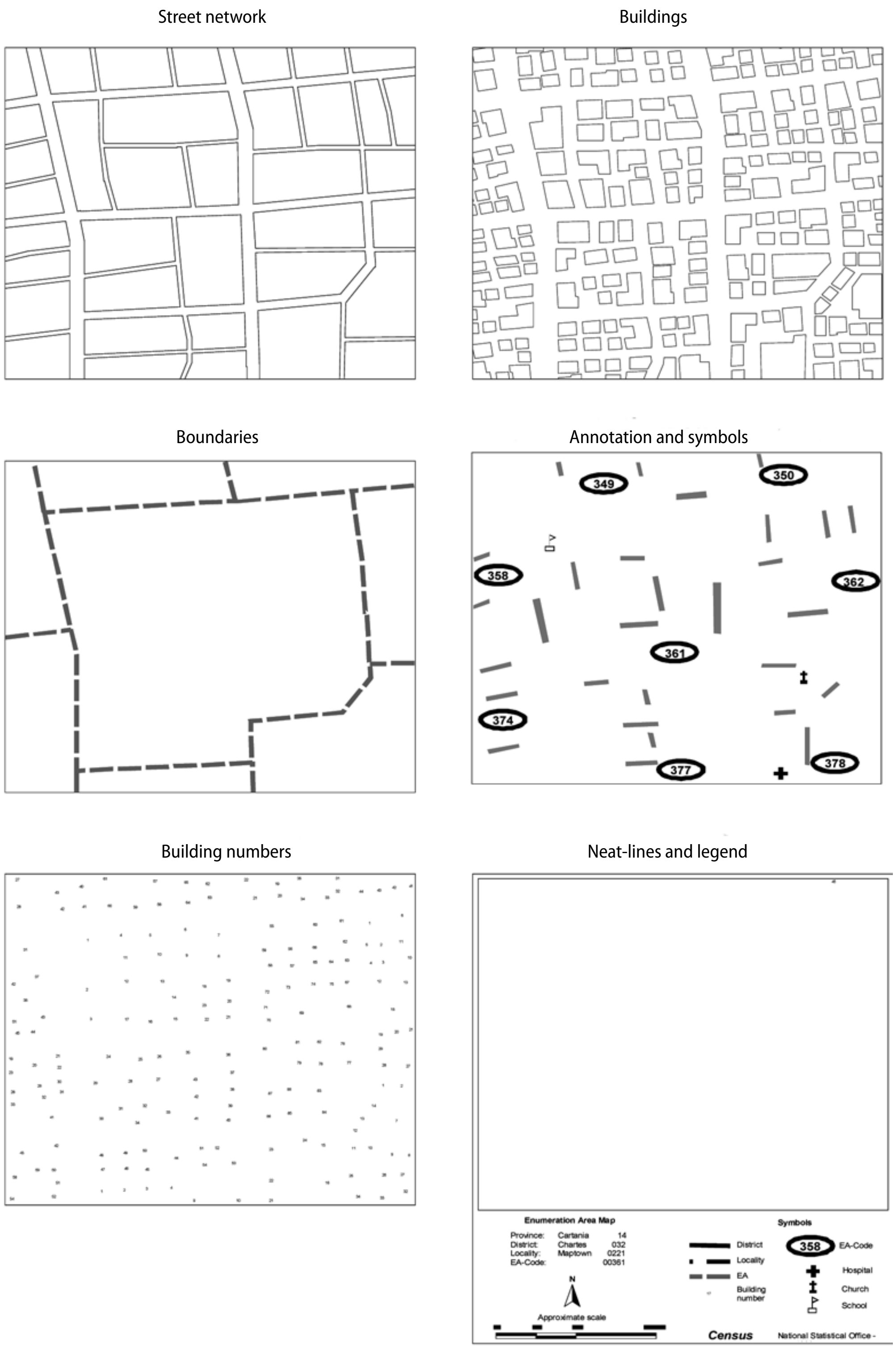
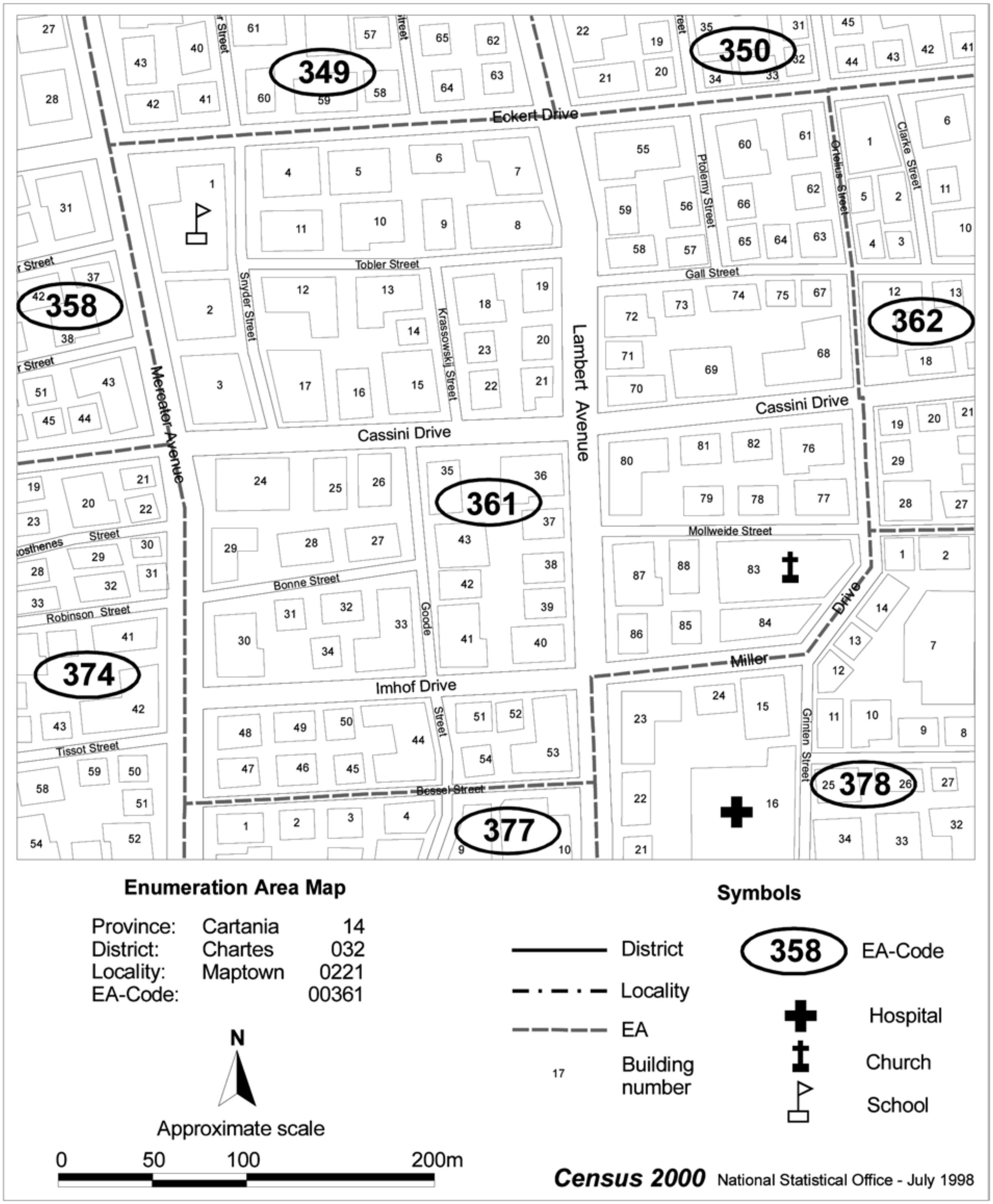


Figure V.3
Example of an urban enumeration area map



5.26. A well designed EA map will usually work well in black and white. Although colour printers have come down in price, their throughput is more limited than black-and-white printers and supplies are often 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, colour may contribute to the clarity of map design. For instance, the EA boundary can be indicated by a brightly coloured 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-coloured highlighter (for more discussion on the topic of colour use, see annex 5).

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

5.28. 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 or 5 copies. The decision whether to distribute maps digitally or in hard copy rests on how centralized the census operation is. If mapping activities are housed in one or a few census offices, a preferred approach is to distribute digital map files rather than hard-copy maps. These files can be transmitted to local census offices on CD/ROM or DVD or via the Internet on a dedicated ftp site. 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 such problems as lost hard-copy maps. Census officials should opt for the most reliable and cost-effective approach given the available options.

5.29. If the census geographic database is consistent and well organized, EA map printing should be fast, and there is usually no need to outsource the work. 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 (also known as the 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.

C. Use of geospatial infrastructure during census enumeration

5.30. The main contribution of geographic databases to a successful census takes place before and after the actual enumeration. Geospatial infrastructure, 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 questions are discussed below.

1. Use of digital maps for census logistics

5.31. Maps are needed for many purposes in the census process. Among these, geospatial technology can also play an active role in planning preliminary work and the logistics of the enumeration. The assignment of administrative units to operational areas, the location of field offices, and planning the travel of fieldworkers and enumerators are some of the tasks for which the use of a geographic database may be helpful. If digital maps are to be used for these purposes and the NSO has opted not to create a digital geographic database, then the census cartographic unit can make use of a coarse-resolution geospatial database, consisting of small-scale digital maps of settlements, roads, rivers and administrative divisions. These maps can be at scales ranging from 1:50,000 to 1:500,000; 1:1,000,000 may be too generalized to be of much help. These data sets can in most cases be obtained from existing sources, including the national mapping agency or survey department. The NSO can further develop its own data sets or use "canned" data that came with the GIS software.

5.32. 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 software program can be used to determine and display possible regional assignments.

5.33. The use of geographic databases for logistical purposes is not quite as critical as their use for EA delineation. 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 geographic database for these purposes is that distance and travel-time estimates will be more accurate and census staff can quickly produce maps showing various aspects of the census-planning process. Furthermore, the development of a small, coarse-resolution geographic database for the country is a good precursor for the much more challenging task of producing a detailed georeferenced census database.

2. Monitoring progress of census operations

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

5.35. 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.

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

5.37. 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 via the Internet. If local and regional supervisors have Internet access, information can even be submitted through a password-protected database interface.

3. Guidelines for map use by enumerators during the census

5.38. The training programme for enumerators should include basic map-reading and navigation techniques. Once in the field, enumerators should locate 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 that 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.

5.39. 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 that it points in the direction of travel. As each corner is turned, the map should be reoriented so that 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.

4. Updating and correction of EA maps during enumeration

5.40. Even if a thorough quality-control programme 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 fieldwork, buildings or streets may have been overlooked or registered incorrectly on maps. Since fieldwork to construct the geographic base for the census needs to be conducted several months or even years ahead of the enumeration, new construction and infrastructure developments may not be updated in the enumerator maps.

5.41. In addition to providing 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 any suggested revisions through an established procedure of incorporating edits into the master geographic census database (for further discussion of this topic, see chap. VI). This may require making the corresponding corrections in the digital census database, or it may require some additional field-checking or accessing satellite or aerial photo data to verify changes. This may be a very complex process but should ensure that the NSO holds the most current information on the enumeration areas, which will reduce the workload for cartographic activities before future censuses or surveys. It should not, however, delay the release of census results.

Box V.1

Production of census field maps in India

1. Census-taking in India is a gigantic operation set up to count and collect information concerning more than one billion persons in the country. Mapping plays a crucial role in ensuring full coverage of the geographical area of the country, without any omission or overlapping.

2. Preparations of maps showing the administrative boundaries of states, districts, subdistricts, villages and towns begin nearly three years in advance. Notifications indicating jurisdictional changes are collected from provincial governments and municipal authorities. The maps prepared at the time of the last census are modified, indicating the changes reported, and are certified by the designated authorities before being used in the census. Any further changes occurring in the jurisdictions are incorporated until administrative boundaries are frozen for the census,. The census organization maintains a database of boundary files for use in the updating exercise, using GIS technology.

3. At the time of houselisting operations, which are undertaken six to eight months before the population enumeration, houselisting blocks are carved out for allotment to about two million enumerators. Since the geographic database, showing detailed layout of buildings, houses etc. is not available in digital format, free-hand drawings, showing the layout of buildings, houses, roads and major landmarks, as available from the previous census, are used for rural areas. In the urban areas, the latest maps available from local authorities are used to delineate houselisting blocks. On the basis of information collected during houselisting operations, fresh enumeration blocks are carved out, covering areas with an average population of 500 to 750 persons for use at the time of enumeration. The free-hand drawing of the enumeration areas, known as "notional maps", are used by the enumerators during the population enumeration and are also shared with other agencies for field survey.

4. Whereas digital maps showing administrative boundaries are available up to the town and village level, detailed digital maps showing the layout of buildings, houses, lanes, by-lanes, road networks and major landmarks are not yet available for towns and villages. Currently, an initiative for preparing a detailed digital geographic database of major towns has been undertaken, which will make use of satellite imageries. The Survey of India, the premier government mapping agency in the country, will provide detailed ward maps for this purpose in digital format, based on satellite imagery. Special field surveys for collecting information on house numbers, types of building, purpose of use, population etc. are proposed for linking with the digital maps and subsequently for demarcating census enumeration blocks. The outcome of this exercise will be the preparation of georeferenced maps for use in the census. Maps showing the location of the enumeration block and detailed layout of buildings, lanes, by-lanes etc. will be handed out to the enumerators for enumeration. These maps would also be shared with other government agencies for implementing their respective programmes (for further information, contact Chinmoy Chakravorty at: cchakravorty.rgi@censusindia.gov.in).

D. Summary and conclusions

5.42. Chapter V 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 VI
Geographic databases for dissemination of census results, products and services

A. Introduction

6.1. Chapter V discussed the use of geospatial infrastructure to support census enumeration. Chapter VI deals with geographic tasks that the NSO will carry out after the enumeration, and with the dissemination and use of geographically referenced census information. At this point in the process, the results from the enumeration should be in. All through the present *Handbook* we have stressed that all geographic plans must align with the overall census plan. A second key to an NSO's investment in a geographic database is its ability to be used for all phases of the census process. Once the enumeration is complete, results can be used to refine the 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.

6.2. If a complete digital census geographic database has been created, then statistical 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 at this stage for producing publication-quality maps to accompany census reports, for distribution to outside users who want to analyse census data spatially or for internal applications. This database can be compiled for a suitable level of the administrative hierarchy or for other aggregated 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.

6.3. For the most part, however, chapter VI 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 NSO needs to adopt a long-term perspective. Immediate tasks after census-taking are thus only the first steps in the preparation of geographic materials for the next enumeration.

6.4. The main topics of chapter VI are management tasks relating to geographic databases after and between censuses, and the development and dissemination of output products. Other topics covered include incorporating changes from the field; the process of aggregating data to dissemination units; database maintenance; the dissemination of products and services; disclosure and data-privacy considerations; marketing; outreach; geographic products; census maps and databases for publication; some methods for analysing census data spatially; and issues relating to Internet mapping and the distribution of census databases.

B. Tasks after the enumeration and during the inter-census period

6.5. After the enumeration is where good planning pays off. If the NSO has followed an overall plan forged early in the process and stuck to it, then the determination of final products has steered the choices in the enumeration phase. If the plan calls for publication of results at the EA level, then possibly no reconciliation will be needed. If EA-level results are not to be published, then some reconciliation will be necessary. Some additional reconciliation will be necessary for geographic changes discovered during the enumeration. The NSO needs to agree up front what kinds of errors require immediate attention, and what can wait because it might slow down the release of results.

Immediate tasks

1. Incorporate updates and changes identified by enumerators

6.6. 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 of the canvassed territory. The census-mapping office should encourage enumerators to record errors, which the local supervisors can 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.

6.7. First, it ensures that tabulations and the development of digital and hardcopy map products are based on the EA delineation actually used during enumeration. Second, 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 must not, however, occur without some assessment of the impact on the census plan as a whole.

2. Aggregation of collection units and tabulation or statistical units

6.8. The most important responsibility after the enumeration 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 level of detail, not to mention the challenging file management, of an EA-level database that spans the country.

6.9. Matching of data collection (EA) and tabulation units at various levels of geography requires the development of "equivalency" or "comparability files". A table that reconciles these two different geographies is sometimes also called a "concordance". Equivalency 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 are illustrated in table VI.1, which shows that one EA (census tract) from 2000 was split into three new EAs (census tracts) for 2010, accompanied by revised coding.

Table VI.1
Comparison of old and new EA units

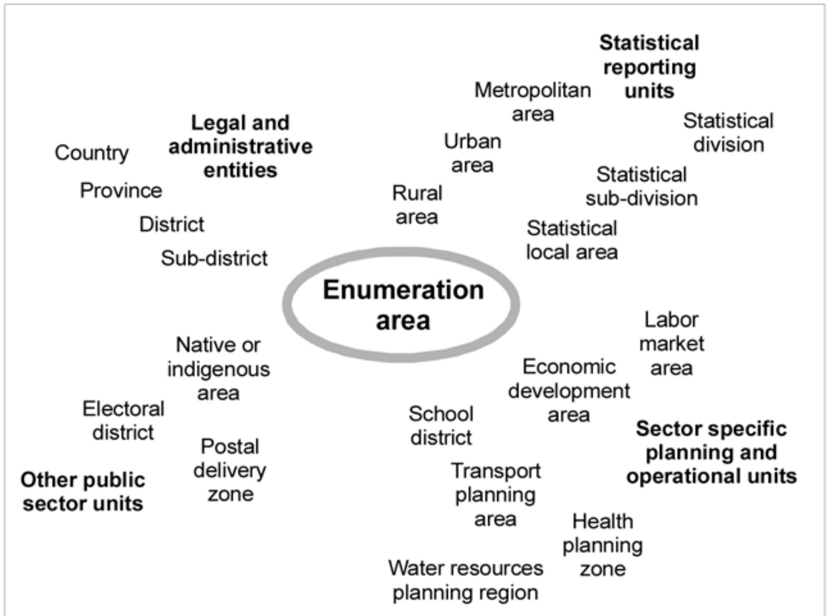
Field-length field description				
1-11 census tract code				
12 part (P) flag				
13-18 street-side mileage of 1990 census tract				
19-22: percentage of 1990 census tract street-side mileage				
23-31 2000 census tract code				
32-33 2000 census tract suffix				
34 2000 census tract part (P) flag				
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

6.10. The development of equivalency files is made easier if a consistent coding scheme has been implemented over the course of the process. Geocoding of census units and attributes are covered extensively in the present *Handbook*, particularly in chapter III. Geocoding 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.

6.11. 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 VI.1). These may in some instances coincide with administrative areas, but they will often 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 the production and maintenance of equivalency files is thus an important task of the census-mapping office. As emphasized throughout the present *Handbook*, planning effectively will have numerous seen and unseen benefits for the NSO.

6.12. Additional comparability files should be developed to reconcile past with present enumeration or statistical reporting areas. Since both data-collection and tab-

Figure VI.1
Examples of census-tabulation and reporting units



ulation 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.

6.13. 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.

3. Database archiving

6.14. After errors and inconsistencies have been addressed in the master geographic database, benchmark copies of all geographic data sets should be produced and archived. The master 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 metadata 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 the completion of database work.

6.15. For NSOs that have a continuous mapping programme, a copy of this database will serve as the basis for regular updating during inter-census activities. The advantages of a continuous mapping programme are discussed below.

4. Database maintenance: advantages of a continuous mapping programme

6.16. The present *Handbook* has argued that the benefits of a digital geographic census programme 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 programme . The continuity of census geographic activities will therefore ensure that the investment in database development is preserved.

6.17. One important consideration 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 inter-census period, a clear system of version control should be implemented. Changes implemented in the database should be documented and published. 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 different staff members making changes to different versions of the database that later have to be reconciled.

6.18. During the inter-census period, the census-mapping agency should follow industry trends and 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 inter-census period.

6.19. 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 geographic concepts and tasks. For a long-term census-mapping programme 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 inter-census period, provide geospatial 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 fieldwork. Retaining core staff will also reduce the start-up costs otherwise required for recruiting geographic experts, who would then need some time to be fully integrated in the census cartographic process (for detailed consideration of human resource issues, see chap. II).

6.20. Again, the importance of a long-term view of census cartographic activities cannot be overemphasized. The additional resources required to maintain a cartographic capability in the NSO between censuses will be well worth the benefits of pursuing a long-term strategy.

C. Dissemination of geographic census products

6.21. Dissemination is often neglected by NSOs, yet it represents the single most important way to educate the public and support other activities. It must be stressed again that the NSO should determine the array of products and services early on in the process, including structure and format. The planning of information dissemination, disclosure and data-privacy issues, marketing, outreach and education are reviewed below.

1. Planning data dissemination

6.22. The definition of geographic output products and services and the scheduling of their release needs to be closely coordinated with the timetable for the overall census project. The tabulation of census data may require information from the census geographic unit, and thematic maps and digital geographic databases can only be completed once census data processing has been completed.

6.23. 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. It is very important that these plans for dissemination products should be made very early and published widely in order to get feedback from the user community.

6.24. It is useful to establish an advisory panel of representatives from the most important census data user communities that can guide the NSO. 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 NSO and data users. The examples of the use of disaggregated census data provided in the introduction to the present *Handbook* provide some indication of the wide range of users that the census office should consider in their user needs assessment.

6.25. Past experience of what has proved popular with census data users can only be a limited guide to the definition of output products and services. Demands change, partly in response to changing technological capabilities among data users. Digital map database products and services 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 hard-copy maps may be larger in many countries than requests for digital information, this is changing as the 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.

6.26. It is advisable to look ahead several years when planning the output strategy. For example, the Internet has become 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 non-governmental data users is through participation in a national spatial data infrastructure (NSDI).

6.27. 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 and services are introduced and as new users realize their potential for their own needs. Thus, the census office needs to be prepared to serve a growing demand once products and services are made available. Effective plans are scalable with the level of demand. It is advisable to define clearly and early which census data users' needs **must** be served (i.e., are legally required to be served), which **should** (through notions of good customer service) and which **will not** be served (because of resource constraints). A clear set of priorities will also facilitate the development of a timetable for census product and service distribution.

6.28. The NSO is strongly encouraged to make results, products and services as widely available as possible through an open data-dissemination policy. 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.

6.29. Some census geographic data products will be required for internal and official use. These may include equivalency files and reference map libraries, as well as 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.

6.30. 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 to meet multiple user needs by publishing analysis as well as data and results. Some examples of such spatial analysis are provided below.

6.31. Census output products and dissemination options are discussed below, including required products, thematic maps that can be distributed in hard-copy or digital format, digital cartographic database dissemination, digital census atlases and Internet-based information dissemination strategies. A thorough background in techniques for thematic mapping is required for many of these output products. Only the more general issues concerning thematic mapping are discussed below. A more comprehensive overview of thematic map design is contained in annex IV.

2. Disclosure and data-privacy considerations: the differencing problem

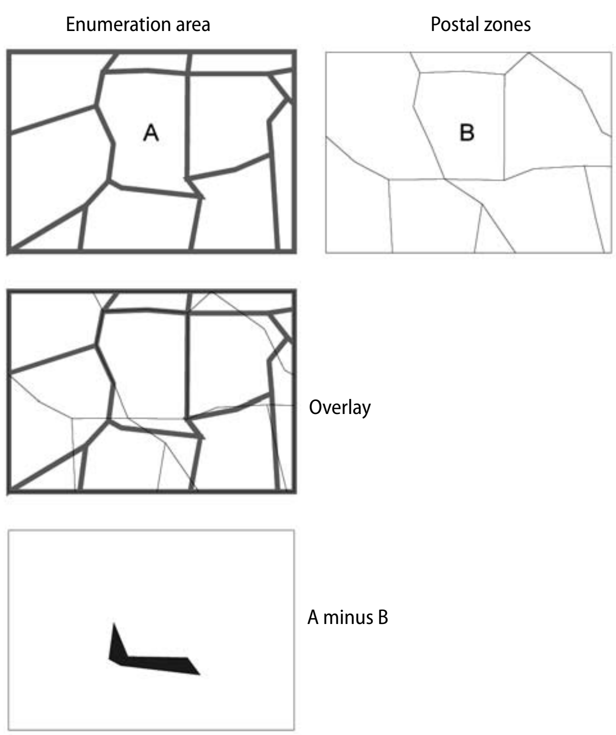
6.32. The NSO should be aware that the public's concern for maintaining the privacy of personal information on the census may affect data dissemination. Guidelines for maintaining the privacy of individuals is covered in standard United Nations guidelines (see, for example, United Nations, 2008).

6.33. 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 sophisticated user may be able to use geospatial 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).

6.34. 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.

6.35. 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 VI.2 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 VI.2
The differencing problem in statistical disclosure



6.36. To avoid data-disclosure problems, the NSO 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) analyse 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 by using 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.

3. Marketing of geographic census products and services

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

6.38. In considering whether to sell or provide free access to data, NSOs should strike a balance between revenue generation and expansion of the data-user base. Countries that aim at recovering some of the costs of developing census geographic 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 geo-demographic data. Most of the leading GIS vendors produce and sell geographic 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.

6.39. 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 geographic 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 geographic data. Demand in other countries may come from internationally operating companies or academics studying the country.

6.40. 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 noted above. NSOs should be wary of signing away marketing and distribution rights to a private company. By so doing, the NSO 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.

6.41. Other potential distribution 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 its data in relation to the costs of producing, advertising and selling the data in order to ensure that a fair and mutually beneficial agreement will be the basis of a public-private or public-public partnership.

4. Outreach and education

6.42. To ensure broad awareness of data availability and the widest possible distribution of georeferenced census data, the national statistical office should develop a publicity plan. That plan could include 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 and distribution 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.

5. List of potential products

6.43. The various products that the NSO should publish after the enumeration is completed are described below. These include equivalency and comparability files, a reference map library, gazetteer and centroid files, and thematic maps.

(*a*) Equivalency and comparability files

6.44. 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.

6.45. Equivalency files should be made available in both hard-copy 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.

(*b*) Reference map library

6.46. 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.

6.47. 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 hard-copy reference maps should therefore also be made available on demand.

6.48. 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 United States Census Bureau (available on the Internet at www.census.gov/geo/www/garm.html).

(*c*) Gazetteers and centroid files

6.49. 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 programme 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.

6.50. A gazetteer should be stored and distributed in digital form, allowing the direct use of coordinates and name information into a GIS. It will also be useful to develop a simple query system, that allows users to 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. Digital boundary files are also a possible census product. These will be discussed below.

6. Thematic maps for publication

(*a*) The power of maps

6.51. Before discussing the types of thematic maps that can be produced for census publications, it is useful to note the following reasons 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 that 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?

(*b*) Thematic mapping of census data

6.52. The adoption of geospatial technology in the census process encourages a view of maps that is quite different from traditional cartography. Using a computer, maps can be generated quickly on a computer screen. This supports a mode of work that is optimized for data validation, the 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 and others, 2005).

6.53. Maps created on a computer screen are sometimes called "virtual maps" to distinguish them from printed or drafted hard-copy maps, although increasingly the line is blurring between computer and hard-copy 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 III—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.

6.54. 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 DVD or posted on an Internet site.

6.55. Table VI.2 shows a list of possible thematic maps that can be included in a census atlas or a census office's Internet site. Many other types of maps might be considered for publications on special topics or to highlight interesting aspects of census results in the various 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 (for an overview of thematic mapping principles, see annex V).

Table VI.2
Suggested thematic maps for a census atlas

Category/title of thematic map
Population dynamics and distribution
Percentage population change during inter-census period(s)
Average annual growth rate
Population density (persons per square kilometre)
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
Percentage of population aged 0-14
Percentage of population aged 15-64
Percentage of population aged 65 and over
Percentage of female population of child-bearing age (15-49)
Total dependency ratio (percentage of population aged 0-14 and 65 and over relative to population aged 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 (aged 15 and over)
Mean years of schooling (aged 25 and over)
Illiteracy rate of population aged 15 and over
Illiterate population aged 15 and over
Educational level of population aged 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 etc.)
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

Source: United Nations (2008).

6.56. 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 EA-level data.

6.57. 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 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.

6.58. 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.

6.59. 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. The information contained in annex 5 should be of interest to core cartographic staff, as well as to people inside and outside the census agency who may only occasionally produce maps from digital spatial databases. Excellent additional references on cartography and thematic mapping are Robinson and others (1995), Kraak and Ormeling (1997) and Dent (1999). 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 (see, for example, Krygier and Wood, 2005; and Brewer, 2005).

(*c*) Digital census atlases

6.60. While a more generic geographic 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, by contrast, combines a digital geographic database and census data in a simple mapping package. The user can use the data to produce custom maps that can be printed or copied into other applications packages.

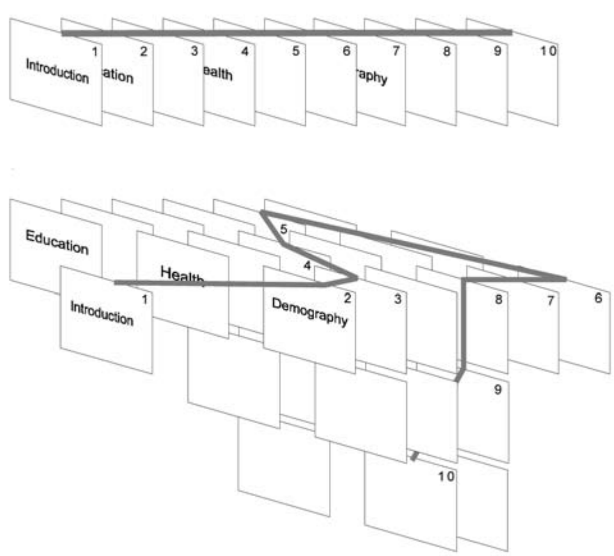
(i) *Static atlases*

6.61. A static digital census atlas can bring together maps, tables, graphs and perhaps 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 that can be distributed together with free viewer software. Most presentations or graphics can also be exported to PDF format that 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 the cut-and-paste commands in the Windows environment.

6.62. An alternative presentation platform is an Internet browser. Most computer users have an Internet browser on their computer that 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 are 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 may not be interested in.

6.63. Most presentation packages provide a better design option that is based on hyperlinks. These links allow the user to jump between different sections of the presentation. They also allow the integration of 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 VI.3
Presentation design options for a static digital census atlas



6.64. The hyperlink concept is illustrated in figure 6.3, where it is contrasted with the linear design approach. In the hyperlink design, several parallel topics are presented that 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 there, links might be provided to maps showing child health indicators (4), educational facilities (5) and so on.

6.65. 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.

6.66. Hyperlinks are familiar to anyone who has used the Internet. 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.

6.67. 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 website for viewers anywhere in the world.

(ii) *Dynamic atlases*

6.68. An alternative to a static census atlas is to publish a digital map and database, together with mapping software, which 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.

6.69. 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 geospatial database but who want more flexibility in exploring and utilizing geographical census information than is possible with a pre-packaged static census atlas.

6.70. 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", so that after installation, the user should immediately be able to produce maps.

6.71. Some census offices have developed in-house map-viewing software and distribute these with their census data products. The maintenance of such software is expensive, however, and commits 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).

6.72. As an alternative, there are now several mapping packages available that are free of charge and can be distributed with a database (for details, see chap. III).

6.73. 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 of Redlands, California, United States. ArcGIS Explorer is a mapping interface for data created by the ArcGIS package.

6.74. 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 websites. Analytical functions are limited, but the system does support different types of data query—interactive or using SQL-like commands—and address matching.

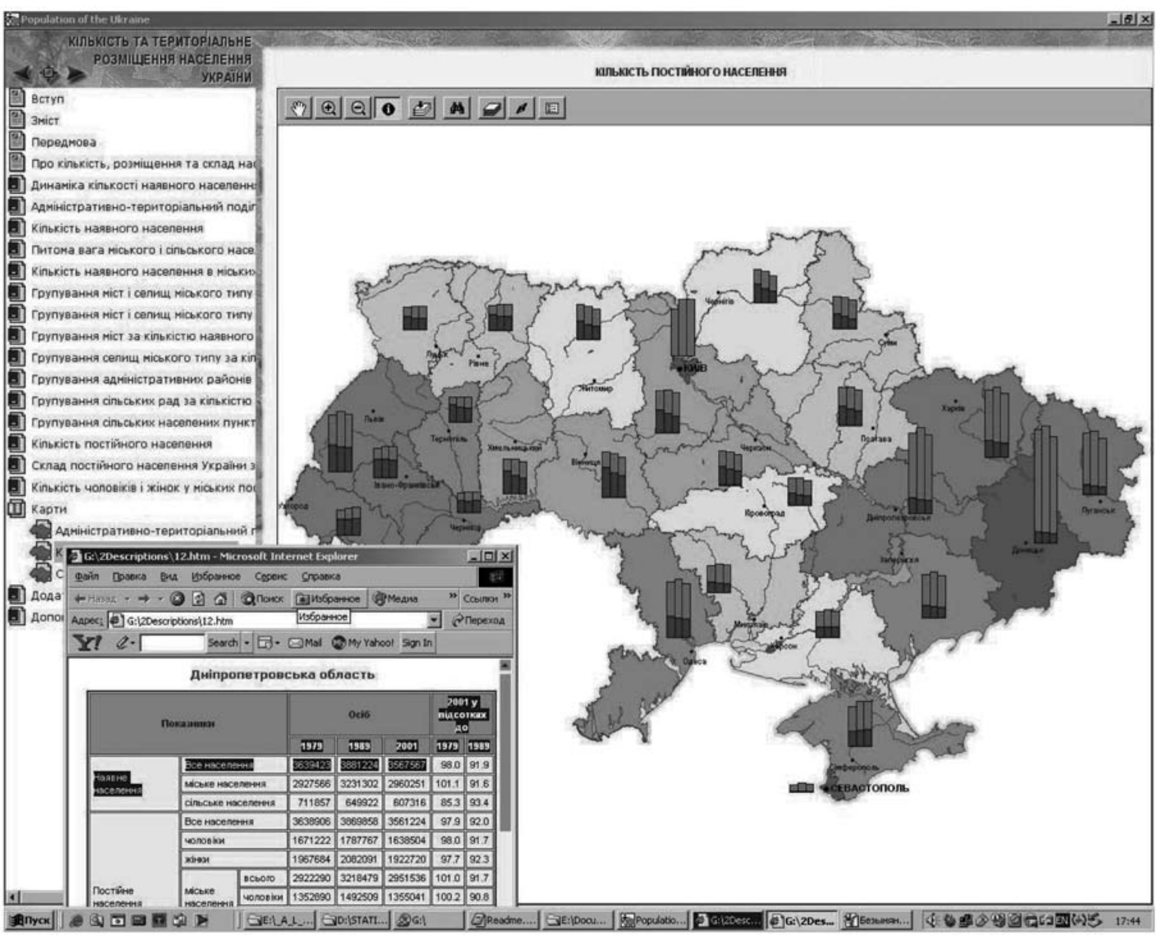
6.75. The documentation for a dynamic census atlas needs to include much of the same information that should accompany a more comprehensive census GIS database. However, the text should be designed with non-expert users in mind. Technical GIS jargon should be avoided. Since the users are unlikely to 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.

6.76. Figure VI.4, prepared by the State Statistics Committee of Ukraine, illustrates the use of statistical and graphic elements in the same map presentation.

(*d*) Spatial analysis techniques

6.77. Although spatial analysis is sometimes used during the enumeration phase (clustering, for example, can aid in identifying housing units to be canvassed),

Figure VI.4
A screenshot of Ukraine's dynamic atlas



the main use of spatial analysis is for census products and services. A variety of techniques, including buffering, linear interpolation, point pattern analysis and cartograms, offer functionality beyond standard thematic (choropleth) mapping, with many tools now available in both commercial and open-source software.

6.78. The presupposition behind use of new methods of spatial analysis is the availability of population data with a higher level of "granularity" (or spatial specificity) than previously, at the EA, population cluster or other small-area levels. If analysts or other GIS users wish to analyse 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 that range from simple queries to measurements to transformations, descriptive summaries and models.

6.79. Spatial analysis is defined by Longley and others (2005) as a set of methods whose results change when the locations of the objects being analysed 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?

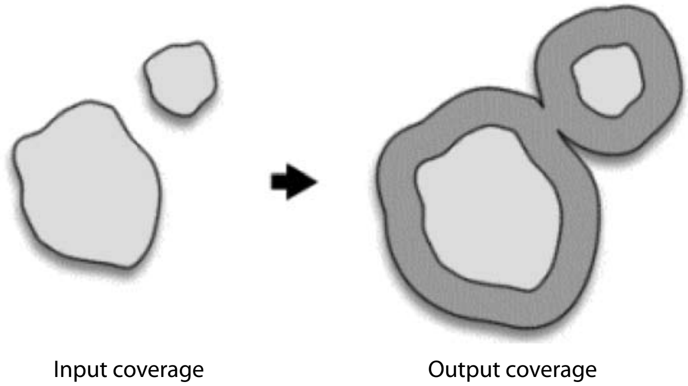
6.80. Some prevalent forms of spatial analysis especially useful for use with population data include:

- "Queries" are considered the most basic type of analysis operation. They use a GIS programme to answer simple questions posed by the user, with no changes in the database and no new data produced. Often, this is the first step in an analysis, where one seeks to create a subset of units, such as populated places, with certain characteristics, allowing the user to check how typical an observation is against other observations. An example of a query using geocoded census data would be: "Select all towns with a population greater than 1,000 persons". These towns can then have their attributes summarized, for instance, to measure their total fertility rates against smaller towns and villages, and the results would then be mapped. The term "exploratory data analysis" refers to investigations of patterns and trends in data using such techniques as querying.
- Measurement analyses use the locational properties of population data, including the simple properties of objects, such as length, area or shape, as well as the relationships between pairs of objects, such as distance or direction, to describe aspects of that data. "Distance measurements" are easily made with all GIS programs, using the centroids (or centre points) of cities, towns and villages. An analysis can be performed to select villages located more than a kilometre from a school, clinic or water source. These can then be further analysed, using the attribute information for the populated places themselves.

6.81. In the first two examples, data sets were queried or measured but no new data resulted. Transformations are methods of spatial analysis that use simple geometric, arithmetic or logical rules to create new data sets. Transformations can include operations that convert raster into vector data, or a stream of GPS coordinates into a route or a boundary.

6.82. Of all the transformational techniques, "buffering" is the most well known and important. Buffering involves building a new data layer 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 VI.5) and can be weighted by attribute values. Buffering can be used to model travel time, for instance, by creating a "catchment area" around a particular feature, such as a school or a clinic. This provides a measure of

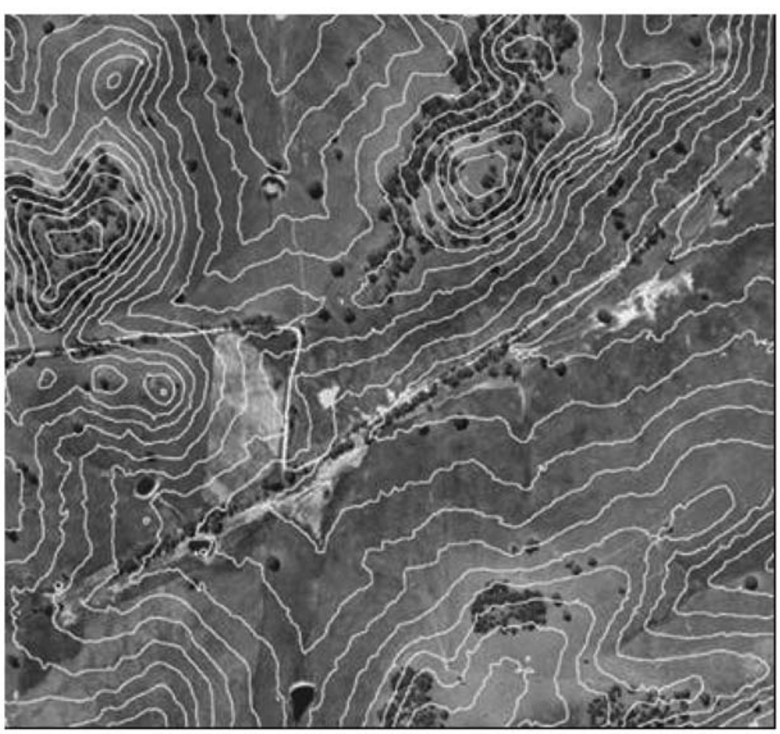
Figure VI.5
Buffering of a polygon object



accessibility that can be mapped across the extent of a country. Or population data can be used in conjunction with other data sets, such as a data layer showing territory that may be susceptible to flooding. The resulting analysis could then identify populations and settlements at risk that could be candidates for disaster-mitigation programmes.

6.83. Another example of a transformation is "point-in-polygon" analysis, which determines whether a point lies inside or outside a polygon. This can be used to compare geocoded village centroids lying inside and outside hazardous areas, such as tropical storm tracks or earthquake zones. "Polygon overlay" analysis involves comparison between the locations of two different polygonal data layers. For example, the boundaries of two administrative districts could be compared to troubleshoot errors in the field-enumeration process.

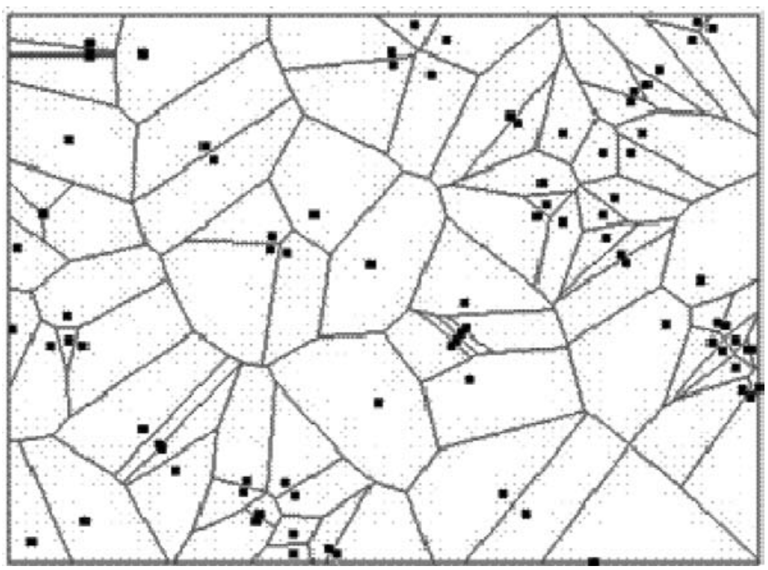
Figure VI.6
Example of linear interpolation creating contours



6.84. "Spatial interpolation" is a spatial analysis 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 modelled to estimate the missing parts of the surface (for an illustration showing how contour lines can be derived through linear interpolation from a satellite image, see figure VI.6).

6.85. 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 (for an illustration of Thiessen polygons, see figure VI.7).

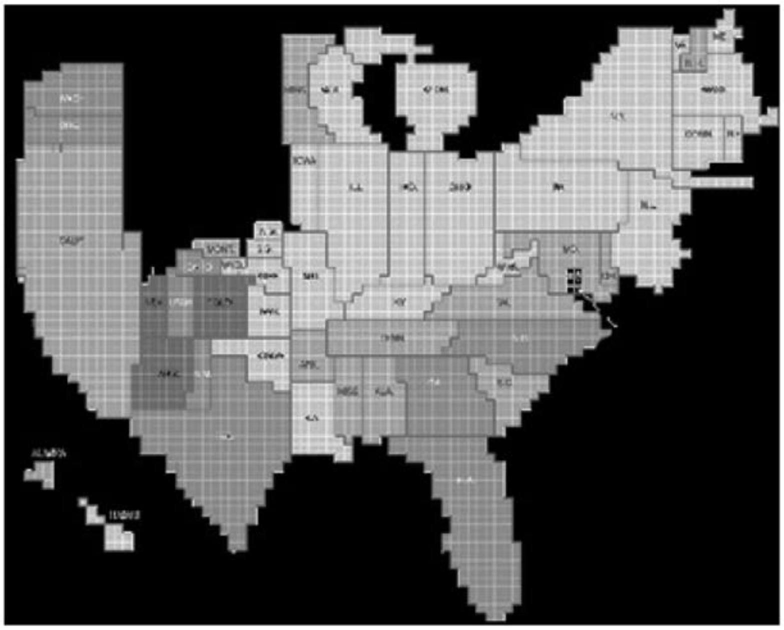
Figure VI.7
Illustration of Thiessen polygons



6.86. Descriptive summaries are a spatial equivalent of descriptive statistics (such as mean and standard deviation) that represent the essence of a data set in 1 or 2 numbers. "Centres of population" are the two-dimensional equivalent of a statistical mean and are often used to display the centre 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 are particularly useful for identifying populations at risk as well.

6.87. "Cartograms" (see figure VI.8) 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, such as population size or voting habits. Analysts can now obtain scripts and extensions from ESRI and elsewhere (for example, see MAPresso at www.mapresso.com) to create them using their polygon layers.

Figure VI.8
Example of a cartogram



(*e*) Map production and publication issues: types of output

6.88. 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-colour 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 hard-copy atlases, with dozens or even hundreds of maps.
- Digital census atlases, as described above, are a cost-effective alternative to printed versions in countries where computers are widely available. Census atlases can be based on either static pre-prepared maps or a simple thematic mapping interface, where the user can select the variables to map, the classification scheme, cartographic symbols and colours and a basic layout.
- Most maps can be published on the Internet as well as in print form. 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.

(*f*) Cartographic tools and software

6.89. New sophisticated GIS software has made the job of producing quality maps much easier. One must nevertheless continue to employ the same design principles traditionally used by cartographers to create visually effective maps.

6.90. 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 clip-art 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, in which 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.

6.91. 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 the depiction of shaded terrain, graduated fills or transparency, which give the cartographer greater flexibility in design. To move data 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. Output options are described below.

7. Output options: digital files

6.92. Various formats for the data dissemination of digital files are discussed below. Analysts should be aware of the difference between dissemination or graphic formats (which are production-quality finished formats) and data formats (which are more raw in form).

6.93. 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 the present *Handbook* were produced in this way. Graphics files can be incorporated in websites as static map images and can also be exchanged as file attachments via electronic mail.

6.94. 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 colour or graytones of tiny dots or pixels arranged as a regular grid. Continuous colour tones or gray-scales are used for photographic-type pictures. Fewer colours are needed to show the more discrete objects typically found on thematic maps.

6.95. 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 of 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.

6.96. A brief description of the most commonly used file formats is set out below. The list is by no means complete as there are dozens of different formats.

(*a*) Data formats

6.97. 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 bit-map 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. 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.
- **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 packages.

- **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 labelled 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 (for PDF (Portable Document Format), see para. 6.102 below).

(*b*) Raster image formats

6.98. Raster images can be created by GIS or graphics packages directly. In some instances, two other options for creating images are useful. One option is to use the screen-capture command in a raster-oriented graphics package. These "screen grabbers" are sometimes better at preserving the original display colours 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.

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

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

6.101. Some standard raster file formats are:

- **BMP.** The Microsoft Windows device-independent bit-map (DIB) format. It allows Windows to display the bit-map 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 colours 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 describes all geographic information associated with the image, such as the projection, real-world coordinates and map extent, 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 the 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 colours, 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 colours or gray-shades, 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. As a result, a photograph that has been exported with a high degree of compression cannot be restored to show all details in the original photograph. A new format, JPEG 2000, is now supported by GIS software programs.
- **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 colour (16 million colours), whereas the GIF file format only allows 256 colours. PNG excels when the image has large areas of uniform colour. The "lossless" PNG format is best suited for editing pictures, while 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. The new geography markup language (GML) has been recommended by the OpenGIS Consortium (www.opengis.org) and is currently being implemented (for more information, see www.opengeospatial.org/standards/gml).
- **DGN.** 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.

6.102. A document format which is arguably unique in that it does not fit into a raster or vector category but is certainly widely used is PDF.

- **PDF.** The Portable Document Format was developed by Adobe. Its initial use was for the 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 once it has been installed. The PDF reader can be downloaded free of charge from the Adobe website. Some experts predict that PDF format will replace postscript files as the main standard for high-level graphics printing. 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.

(*c*) GIS data formats

6.103. A variety of formats exist for disseminating spatial data in more raw form to dedicated GIS users and spatial analysts. Formats include those for geographic coordinate data and also tabular data, which can be joined or linked as attributes to spatial data in a GIS.

(*d*) Coordinate data

6.104. 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, those conversion functions need to be purchased separately. For the most part, many vendors' software can read other vendors' software.

6.105. Despite some efforts by commercial and public geospatial technology groups (for current discussions of the GML data format, see, for example, www.opengis.org), there is still no universally accepted and widely used generic or open source data-exchange format for GIS. 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 the following:

- **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.
- **ESRI's Shape file** (.shp) is a simpler format used by ESRI's ArcView and ArcGIS desktop 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 catalogues, network data sets, terrain data sets and address locators. Each File Geodatabase file can be as large as 1 terabyte. Personal geodatabases use the Microsoft Access file system (.mdb) and can be as large as 2 gigabytes. The file geodatabase is also used by ArcSDE spatial database engine.

- **Google's 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 (for GML, see paras. 6.101 and 6.105 above)).
- **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.
- **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.

6.106. All of these formats support boundary and attribute information. Any commercial GIS software programme will have an import function for at least one or two of these formats. Ideally, a census office should offer its public-release geographic 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.

6.107. Distribution of GIS data in its 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, path-name 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.

(*e*) Tabular data

6.108. New GIS software development has de-emphasized the import of tables in favour of relational databases, such as Oracle and Access. Most GIS packages still 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 (REtrieval of DATa for small Areas by Microcomputer) and CSPro (Census and Survey Processing System), a suite of population analysis tools available for free from the United States Census Bureau. The comma-separated values (.csv) format is also used for tables and is not vendor specific.

6.109. 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.

D. Printing

1. Overview

6.110. For small print runs or quality-control plots, a census office should have one or a number of printers available. 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 colour or gray tones that can be produced.
- Media size.
- Supported media types (plain paper, specially coated paper, transparencies etc.).

2. Printer types

6.111. The following are the most popular printer types (for most NSO needs, the best option for cost-effectiveness and reliability is the laser printer, but other options are available):

- **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. Colour laser printers have reached a price range to be considered for most graphics application environments.
- **Inkjet printers** produce output by squirting electrically charged drops of colour 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 colour output device.

- **Thermal printers** require special paper and ink-coated ribbons that are moved across a thermal head. Ink is fused to the paper where the thermal head applies heat. The colour ribbons are coated with three colours (cyan, magenta, yellow (CMY)) or four colours (cyan, magenta, yellow and black (CMYK)) so that three or four passes of the thermal head across the paper are required. In thermal wax printers, the heat causes a layer of coloured 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 colour variation compared to thermal wax printers.
- **Electrostatic printers** use toner that is transferred through electrical charges to a non-conducting surface. Toner is either attracted or repelled. Direct electrostatic printers apply the charge directly to the specially coated paper. Toner for each colour is applied in separate passes. Subsequently, the toner is fused to the paper after all colours have been applied. Another electrostatic process is colour xerography, which uses a drum or belt that is charged when exposed to light.

6.112. Many draft maps do not have to be printed in colour. 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.

6.113. Colour printers are useful for printing complex maps for which monochrome shading and symbolization would be insufficient. Inkjet printers are commonly used colour printers—from A4/letter-size desktop printers to large-format printers (e.g., 60 × 90cm or 24 × 36 inches)—although colour laser printers may supplant them. They produce high-quality maps at 600 dpi. Printing speeds are still relatively slow for inkjet printers.

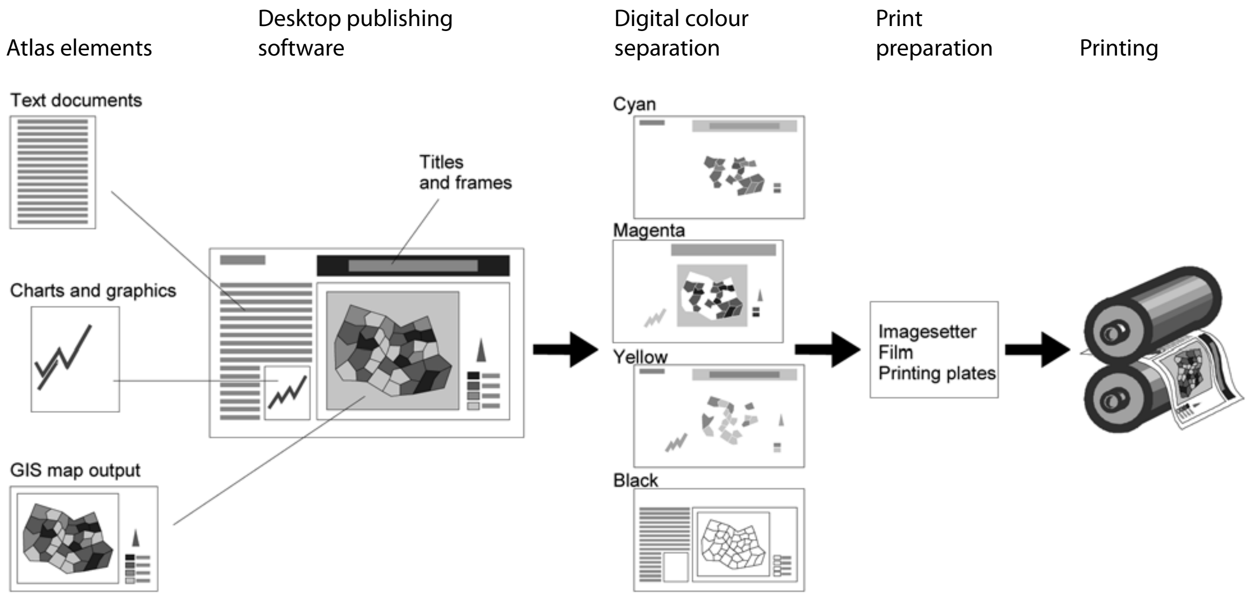
6.114. In deciding upon a suitable printer for a GIS project, cost is a major issue. Something to keep in mind is that the purchase price of a printer is only one—often relatively minor—cost component. While printer prices have dropped considerably, the cost of ink cartridges and special paper have remained 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 per cent of the page is covered by ink). Computer trade journals often publish comparisons.

3. Commercial printing

6.115. 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, analogue 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.

6.116. 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 VI.9). After an initial planning stage, in which the text,

Figure VI.9
The digital printing process



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.

6.117. 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. Considerations for deciding whether or not to contract out include quality, cost, the number of copies required and time.

6.118. 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 software can also produce colour 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 colours cyan, magenta, yellow and black (the CMYK colour model). The colours on the maps and graphics are produced as additive combinations of various percentages of these four colours. The digital files are then sent to an image-setter that 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 and 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 colour proof from the printer before final production. Many vendors of printing hardware and software also provide extensive information and other resources on their websites.

E. Digital geographic data for dissemination

6.119. Increasing software sophistication and Internet use worldwide has made the dissemination of digital geographic data feasible. The demand for digital databases that consist of extractions of the census agency's digital geographic master database will only increase. Census data are an important input in policy planning and academic analysis in many fields. Provision of health services, allocation of educational resources, design of utilities and infrastructure and electoral planning are some applications for which government agencies require spatially referenced small-area population statistics. Commercial users employ such data for marketing applications and location decisions.

6.120. Advances in data processing and GIS mean that producing a digital geographic database at the enumeration area or similar disaggregated levels is increasingly within the reach of many national statistical organizations. To meet user needs, the NSO should pursue a digital data-dissemination strategy. NSOs should consider the costs and benefits of providing data at such a detailed level.

6.121. 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.

6.122. 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. Metadata must always accompany geographic data dissemination, and the costs of metadata production should be factored into the equation as well.

6.123. One alternative to release of an EA dissemination database is the creation of a derived product at a similarly detailed scale.

1. Digital data-dissemination strategies for reaching potential users

6.124. As noted above, the NSO should consider the costs and benefits of wider data dissemination when planning its products and services. The wide range of potential users of disaggregated census data means that the NSO needs to pursue a multi-levelled digital data dissemination strategy. Broadly, we can distinguish between the following types of users:

- **Advanced GIS users** can work easily with large data sets and can use ftp (file transfer protocol) to access them. They require extensive metadata. Often advanced users 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 or "power users". They will want access to spatial and attribute information in a comprehensive digital geographic format. The census office needs to supply comprehensive documentation (see chap. III. F on metadata development) on the geographic parameters used for the geographic database, as well as on the individual census variables. The spatial information will be distributed in an open geographic format that can be easily converted into any number of commercial GIS formats.
- **Computer-literate users** in the government, commercial or private sector 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 ana-

lytical functions, such as aggregation of census units to custom-designed regions, should also be possible. This 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.

- **Novice users** mostly want to view pre-prepared maps on a computer and perhaps perform some basic queries. For this group of users, the best data-distribution strategy is to produce a self-contained digital census atlas. The 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.

(*a*) Definition of data content

6.125. The first step in preparing geographic databases for general release is to define the data content to be disseminated. The NSO should plan for the release of separate data sets, not the operational ones used for internal management. One main reason for this is that keeping the two data types separate will allow the NSO to continue maintenance and updating of its internal data on a continuous basis. For disseminating data sets, a number of questions need to be addressed, as set out below.

(i) *Up to which level will data be released?*

6.126. 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 or statistical validity. Even at the EA level, there may be special reporting zones that contain only a few households and for which census data cannot be released. If necessary, data for selected reporting zones must be deleted or recoded.

(ii) *One large GIS database or a family of census databases?*

6.127. 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. In addition, the NSO should consider the physical size of data files and the impact of file downloads on servers. For instance, downloading a 3 gigabyte data file using a 56 kilobyte server takes about 3 days.

6.128. 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 that show indicators at the subdistrict and EA level can be constructed. Individual databases can also be useful for major urban areas.

6.129. 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 pre-enumeration field activities.

6.130. Offering databases for subsections of the country will ease data use. Many users only need census information for a relatively small region, usually their own. 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.

6.131. If separate databases are distributed, care must be taken that the individual partitions are compatible. The shared boundaries between database subsets must 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.

6.132. 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, by contrast, may not be supported by the GIS software. Users would then have difficulties in employing the census database for geographical analysis applications.

(iii) *How tightly should boundaries and database be integrated?*

6.133. Census GIS databases are characterized by the large number of their 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. It is very important to re-emphasize the need for consistent coding to ensure that the various geographical entities are defined uniquely.

(iv) *What amount of metadata needs to be provided?*

6.134. While the production of metadata and documentation is absolutely essential as accompaniment to any data release, the specific amount of information needed will vary by user. One important distinction is between internal users, who normally require extensive metadata, and external users, who normally require less extensive metadata. See the later section in this chapter, as well as earlier sections (for more information on metadata, see chaps. II and III.F above and paras. 6.137-6.139 below).

(*b*) File-naming conventions

6.135. Although the Windows, Macintosh, UNIX, and LINUX operating systems all support long file names, it is good practice to use the DOS 8.3 (i.e., 8 character) file-naming conventions for all data and documentation files that are distributed, and to develop internal standards for maintaining consistency. 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. Consistent naming conventions that are explained in the documentation will make it easier for the users to find the data they need quickly.

(*c*) Compression

6.136. GIS files can be very large, so that, 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 or CD-ROM, file compression will greatly facilitate data distribution. The most widely used compression software programs for the Windows environment are PKZIP and Winzip utilities. They are available on most computers, while utilities that can extract files from the compressed archives also exist for the UNIX operating system. It should be noted, however, that some formats, including the File Geodatabase, cannot be compressed. Self-extracting files are more convenient for inexperienced users and do not require a decompression utility. However, they are operating at a system-specific level and should only be used if the target computer platform is known.

(*d*) Documentation, including data dictionaries

6.137. 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 chaps. II-V above). Simple ASCII ("read-me") text files can be read by any user. 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.
- A 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), as follows:
  - Cartographic projection, with all required parameters, such as standard parallel or meridian and false easting and northing.
  - Coordinate units (e.g., decimal degrees, metres, feet).
  - Source map scale (i.e., the scale of the hard-copy maps from which boundaries were digitized).
  - Geographic accuracy information (e.g., any numeric accuracy information available for the source maps can be reported). If a quantitative assessment of data quality is not possible, accuracy may 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.
- Disclaimers, copyright information etc.

6.138. 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. The term "data dictionary" is an older term that refers to the specific formatting and field names in the data set. The information provided should include::

- 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 centres, and so on.
  - Any available data-quality information that allows users to judge the suitability of the data for a given task.

6.139. 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 contained in annex IV.

(*e*) Quality control and assurance of deliverable data products

6.140. As discussed above, quality control is an important step before the release of final products for reproduction. After producing the final version of all databases in the form in which products will be distributed (e.g., compressed), the database should be tested on all target platforms (e.g., Windows environment, UNIX, Macintosh, LINUX).

6.141. 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 the distribution of customized data sets of 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.

6.142. 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.

6.143. 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 the distribution of very large files in many countries. Download times are often unacceptable due to shortcomings in the Internet infrastructure in many countries. The main bottleneck, however, is obtaining 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.

6.144. 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 website and incremental use of web-server resources. Census geographic databases can thus be provided to the user at very low cost or free of charge. Some organizations may, however, decide to charge for online data. One reason may be to cross-subsidize a publication programme for users without access to the Internet. Another reason is the organization may wish to recover part of the cost of data collection and compilation of the census data.

2. Legal and commercialization issues

6.145. Issues NSOs ought to consider relating to production and ownership of geographic data are discussed below. These include data copyright, trade-offs in the commercialization of geographic data and liability issues.

(*a*) Data copyright

6.146. 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 geographic 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 and cartographic census information.

6.147. 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.

6.148. 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.

(*b*) Trade-offs in the commercialization of geographic data

6.149. 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 taxpayers have already funded data collection, they should not be charged again for the use of the data. As a consequence, geographic 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.

6.150. 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 repackaged 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.

6.151. 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 since tax revenues have increased and improved access to information has 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.

6.152. 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 geographic data. As Prevost and Gilruth (1997) point out, cost-recovery efforts that put census geographic 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 the use of alternative, cheaper and lower-quality data.

6.153. 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 owing to the realization that the benefits of charging higher prices do not justify the cost of enforcement of copyrights and lost societal benefits due to the reduced use of vital information.

6.154. Access to 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).

6.155. If data from other agencies—such as the national mapping agency or local authorities—are used for producing census maps, then 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 probably require the collaboration of those agencies for future census-mapping activities.

6.156. 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 described above. For instance, special arrangements can be made between government agencies who wish to incorporate each other's 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.

(*c*) Liability issues

6.157. 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 are 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, the 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.

6.158. Another example related to liability issues that is very relevant to census data dissemination is the violation of the 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., EAs, postcode sectors, health or education districts—there is a possibility that sly or unethical GIS operations can isolate information for groups of persons smaller than the lowest disclosure level (see paras. 6.32-6.36 above on the differencing problem). In some countries, this may be grounds for legal action by the individuals concerned.

6.159. 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, by contrast, may shield the agency from such claims.

6.160. Before distribution of spatially referenced data, the agency should thus consult with legal experts and draft a disclaimer that 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 that no guarantee 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.

3. Internet mapping

6.161. Many national statistical organizations have embraced the Internet as a means to disseminating 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.

6.162. Use of the Internet for data dissemination is to be encouraged for many reasons. However, opting for Internet dissemination represents an investment on the part of the NSO, particularly for the load that Internet dissemination places on computer servers, storage, updating and backups. NSOs are encouraged to consider the total lifetime costs of investing in Internet deployment as a part of their general census planning.

6.163. The Internet is eminently 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 websites can give data users access to useful information. However, they do not allow the user to manipulate the data and produce custom maps for specific geographic areas. Approaches that allow a significant degree of user interaction with the census geographic database are outlined below.

6.164. 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.

6.165. 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 that are run from a central server, each staff member can access geographic information through his or her browser software.

6.166. Service-oriented architecture is an umbrella term for methods that include server-side client-side and hybrid approaches (each is described in more detail below):

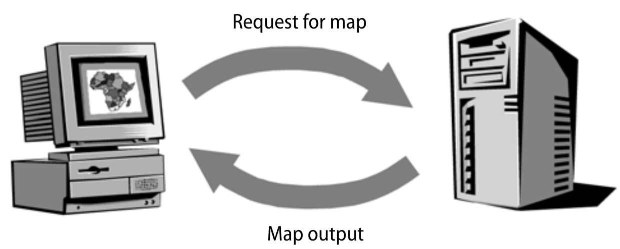
- In server-side strategies, the user sends a request for a map to the server holding the database. Mapping software on the server processes the request, produces a map—for example, in GIF format—and sends it back to the user.
- By contrast, in client-side strategies most of the processing tasks are performed locally on the user's or "client's" computer.
- Hybrid approaches, finally, combine server-side and client-side approaches.

(*a*) Server-side approaches

6.167. 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, in which a powerful central computer handles data management, storage and processing for a number of users that are connected by "dumb" terminals.

6.168. The principle of a server-side strategy is summarized in figure VI.10. The user connects to a website 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 colour scheme, and additional data layers that provide context, such as roads, rivers or administrative boundaries.

Figure VI.10
Internet mapping: the server-side approach



6.169. The user's request is sent through the Internet to the server and routed to a GIS package. The GIS software can be located either on the web server or on a separate computer connected to the server. The GIS package can 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.

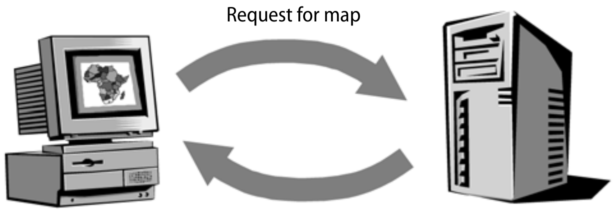
6.170. The server-side approach has the advantage that the user does not need a powerful computer to access possibly very large geographic databases. Even fairly complex geospatial procedures, such as address matching or network routing, can be carried out quickly if a powerful server is available. All 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 of 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.

(*b*) Client-side approaches

6.171. Client-side approaches—or 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.

6.172. In the first client-side approach, 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 programme or applet that enables mapping or geographic analysis (figure VI.11). An applet is a platform-independent piece of software written in the Java programming language that 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 VI.11
Internet mapping: the client-side approach



6.173. 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 programme that extends the Internet browser's 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.

6.174. 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 programme files may be very large, requiring a fast Internet connection, so that 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 computers. This is a problem if some or all of the geographic data on the statistical agency's server are copyrighted.

(*c*) Hybrid approaches

6.175. Server-side approaches are good at providing access to relatively simple maps for 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 geographic data access for census office staff.

6.176. 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.

(*d*) Opportunities for census data distribution, including MapServer

6.177. Currently available Internet mapping packages are scalable. Data providers can purchase an off-the-shelf package that works with standard data sets. Since the mapping of census data is a fairly standard application, national statistical offices should have no difficulties in finding a suitable solution. For more complex applications, a toolbox of software modules can be obtained that allows the data provider to custom-design the map server interface.

6.178. 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 (see www.opengeospatial.org) 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 features advanced cartographic output, including thematic mapping and map element automation; scale-dependent feature drawing and feature labelling; 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).

6.179. With increased network capacities, larger data sets and programme 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.

6.180. 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, including "power users" and active and passive users. These categories roughly parallel the typology of users set out above, as follows:

- **Power users** wish 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, DVDs or Internet download options of "raw" census GIS data sets.
- **Active users** have some expertise in GIS but do not have local GIS capabilities. These users want to download parts of the database, together with GIS programme modules (applets) that can perform the required tasks.
- **Passive users** simply wish to obtain a predesigned 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 data.

6.181. A flexible census data-distribution system on the Internet could look function as follows:

- 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 needs information about the demographic characteristics of potential customers could request demographic information for a circular area of five km radius surrounding a shopping centre location. A government planning agency may request data on the population living within five kilometres of a proposed highway corridor.

6.182. 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 colours. 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.

6.183. 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 EAs within a given district. For requested areas that do not match the standard census geographic hierarchy, some further processing is required. In some countries, dwelling unit geospatial databases are now available or under construction, in which 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. Where that is not possible, the server-based geospatial database should perform an areal interpolation, using techniques such as those described above.

6.184. 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 should be available as well.

6.185. Depending on the data-distribution policies of the country, these services may be free or fee-based. While requests for basic information that has been compiled already may be provided free of charge, more complex requests may be fee-based.

6.186. Data privacy is an important consideration if the custom-tabulations are based on microdata. Internet security issues are as significant in the management of census data on networks as they are in commercial Internet applications. The internal network that may provide access to census micro-data must therefore be separated by a firewall from the Internet domain that allows external users access to aggregate census data.

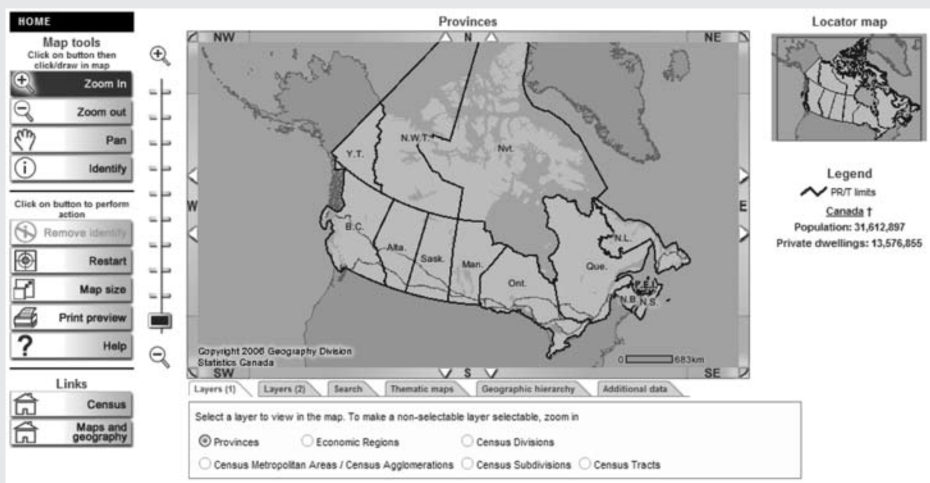
6.187. Obviously, the envisioned data-distribution interface is very ambitious. It requires fast Internet connections and can only reach a large number of users if Internet access in private households, businesses and government agencies is widespread. In many countries, these conditions are not yet 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.

Box VI.1

Case study of web mapping for the dissemination of census data: Canada

Since the late 1990s, the Internet has been the key dissemination medium at Statistics Canada. Web-based dissemination has enormous potential to provide the user with great flexibility in accessing statistical data. A key to exploiting this potential is the provision of tools that permit the user to easily explore the data that are available. Statistics Canada has developed a web-based mapping tool that permits the user to geographically explore census data through interaction with a map. GeoSearch enables the user to search for, identify, visualize and obtain basic geographic and demographic data for areas in Canada.

GeoSearch2006



To find a specific place in Canada, users can interact with a common set of mapping tools located on the left-hand side of the screen to zoom in or out, move around the map or identify their area of interest. A series of tabs under the map window serve as the gateway to a variety of different functions, such as selecting a standard geographic area type, searching for an area of interest using several types of information (place name, address, or postal code), thematically mapping census data, identifying the geographic hierarchy relationships or accessing additional data. Once an area is identified, users can view population and dwelling counts, select from over 35 census variables to create dynamic thematic maps, link to over 10,000 PDF reference maps or access a wealth of demographic and socio-economic data in the 2006 community profiles.

In addition to GeoSearch, Statistics Canada produces a full set of census geography data products that are available on its website. These products are arranged in three major product lines. The maps product line (PDF format) includes reference maps that illustrate the extent of standard geographic areas and thematic maps that display statistics related to several themes. The spatial information product line includes the boundary files for all standard geographic areas, the road network file and the hydrographic data file. These spatial products enable users to develop their own geospatial databases for analytical purposes or incorporate the census geographic areas into their existing geospatial database. The attribute information product line includes data on the characteristics of the geographic areas (e.g., population and dwelling counts, land area and population density) and the relationships between geographic areas. These products can be used independently of, or in conjunction with, spatial information products (for more information, contact Joe Kresovic at Joe.Kresovic@statcan.gc.ca).

F. Summary and conclusions

6.188. Chapter VI has examined issues relating to geographic tasks after the enumeration, including data dissemination. After changes to the geographic base have been received from the field, a process begins whereby the geographic database is continuously edited for use in surveys and future censuses. Data can be released at a modified EA level or through aggregating into new dissemination units, such as clusters. A number of static and dynamic products and services can be produced from the census, including maps, reports, atlases, CD-ROMs and interactive websites.

6.189. Releasing data at the disaggregated level raises a number of disclosure and privacy issues. An additional issue related to data release is whether to provide data free of charge or to raise revenues selling value-added products, such as data CD-ROMs or DVDs. The present chapter has also included a description of geographic data products, such as map viewers, attributed spatial files for use in commercial GIS packages and Internet mapping products.

Bibliography and references

Aronoff, S. (1991). *Geographic Information Systems: A Management Perspective*. Ottawa: WDL Publications.

ASCE (1994). *The Glossary of the Mapping Sciences*. Bethesda, Maryland: American Society for Photogrammetry and Remote Sensing, and American Society for Civil Engineers.

Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, Wisconsin: University of Wisconsin Press. First published in French as *Semiologie Graphique*. Paris, 1977.

Brewer, C. (1994). Colour use guidelines for mapping and visualization. In A. M. MacEachren and D. R. F. Taylor, eds. *Visualization in Modern Cartography*. London: Pergamon.

_________ (2005). *Designing Better Maps: A Guide for GIS Users*. Redlands, California: ESRI Press.

Bugayevskiy, L. M., and J. P. Snyder (1995). *Map Projections : A Reference Manual*. London: Taylor and Francis.

Canters, F., and H. Decleir (1989). *The World in Perspective: A Directory of World Map Projections*. New York: John Wiley and Sons.

Castro, M., et al. (2004). Integrated urban malaria control: a case study in Dar es Salaam, Tanzania. *American Journal of Medical Hygiene*, 103-117.

Dent, B. D. (1999). *Cartography: Thematic Map Design*, 5th edition. Dubuque, Iowa: Wm. C. Brown Publishers.

Duke-Williams, O., and Rees, P. H. (1998). Can census offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure. *International Journal of Geographical Information Science*, vol. 12, No. 6.

Economic Commission for Africa (2007). Determination of fundamental data sets for Africa: geoinformation in socio-economic development. Document ECA/ISTD/GEO/2007/02E. Addis Ababa.

Fothergill, S., and J. Vincent (1985). *The State of the Nation: An Atlas of Britain in the Eighties*. London: Pan Books.

Garson, G. D. (2003). *Public Information Technology: Policy and Management Issues*. Hershey, Pennsylvania: Idea Publishing Group.

Hohl, P., ed. (1998). *GIS Data Conversion: Strategies, Techniques, Management*. Santa Fe, New Mexico: Onword Press.

Hurskainen, P., and P. Pellikka. (2006). Change detection of informal settlements using multi-temporal aerial photographs: the case of Voi, SE Kenya. Proceedings of the Fifth Conference of the African Association of Remote Sensing of the Environment. Helsinki.

Jensen, J. R. (1996). *Introductory Digital Image Processing: A Remote Sensing Perspective*, 2nd edition. New York: Prentice-Hall.

__________, and D. Cowen (1999). Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing*, vol. 65, No. 5, pp. 611-622.

Johnson, J., and H. J. Onsrud (1995). Is Cost Recovery Worthwhile?. *Proceedings of the Annual Conference of the URISA*, San Antonio, Texas, July.

Khosrow-Pour, M. (2005). *Practicing e-Government: A Global Perspective*. Hershey, Pennsylvania: Idea Publishing Group.

Kraak, M. J., and F. J. Ormeling (1997). *Cartography: Visualization of Spatial Data*. Harlow, Essex, United Kingdom: Longman.

Krygier, J., and D. Wood. (2005). *Making Maps: A Visual Guide to Map Design for GIS*. New York: Guilford Press.

Longley, P., M. Goodchild, D. Maguire and D. Rhind. (2005). *Geographic Information Systems and Science*. New York: John Wiley and Sons.

Lynch, M., and K. E. Foote (1997). Legal issues relating to GIS: the geographer's craft project. Austin: University of Texas.

MacEachren, A. M. (1994). *Some Truth with Maps: A Primer on Symbolization and Design*. Washington, D.C.: Association of American Geographers.

__________ (1995). *How Maps Work: Representation, Visualization and Design*. New York: Guilford Press.

McDonnell, R. and K. Kemp (1995). *International GIS Dictionary*. Cambridge, United Kingdom: GeoInformation International.

Monmonier, M. (1993). *Mapping It Out. Expository Cartography for the Humanities and Social Sciences*. Chicago: University of Chicago Press.

Montana, L., and J. Spencer (2004). *Incorporating Geographic Information into Measure Surveys: A Field Guide to GPS Data Collection*. Chapel Hill, North Carolina: Carolina Population Center of the University of North Carolina.

Netherlands Interdisciplinary Demographic Institute (1996). Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-Scale Demographic Surveys, 22-26 April 1996. The Hague: Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund.

Niebert, D. D., ed. (2004). *Developing Spatial Data Infrastructures: The SDI Cookbook*. Global Spatial Data Infrastructure.

O'Looney, J. A. (2002). *Wiring Governments: Challenges and Possibilities for Public Managers*. Westport, Connecticut: Quorum Books.

Padmanabhan, G., J. Yoon and M. Leipnik (1992). *A Glossary of GIS Terminology*, Technical Report, No. 92-13. Santa Barbara, California: National Center for Geographic Information and Analysis.

Prévost, Y., and P. Gilruth (1997). *Environmental Information Systems in Sub-Saharan Africa*. Building Blocks for Africa 2025, Paper No. 12. The World Bank, Washington, D.C. and UNDP/UNSO, New York.

Rhind, D., ed. (1997). *Framework for the World*. Cambridge, United Kingdom: GeoInformation International.

Robinson, A. H., J. L. Morrison, P. C. Muehrcke, A. J. Kimerling and S. C. Guptill (1995). *Elements of Cartography*, 6th edition. New York: John Wiley and Sons.

Snyder, J. P. (1982). *Map Projections Used by the United States Geological Survey*, Washington, D.C.: Government Printing Office.

_________ (1993). *Flattening the Earth: Two Thousand Years of Map Projections*. Chicago: University of Chicago Press.

Tomlinson, R. (2007). *Thinking about GIS: Geographic Information System Planning for Managers*. Redlands, California: ESRI Press.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press,.

_________ (1990). *Envisioning Information*. Cheshire, Connecticut: Graphics Press,.

United Nations (2007). *Geographical Information Systems for Population Statistics*. Sales No. E.97.XVII.30.

_________ (2008). *Principles and Recommendations for Population and Housing Censuses, Revision 2*, United Nations publication, Sales No. E.07.XVII.8.

United States, Census Bureau (1978). *Mapping for censuses and surveys*, Statistical Training Document ISP-TR-3. Washington, D.C.: United States Department of Commerce, Bureau of the Census.

United States, Federal Geographic Data Committee (1997a). *Framework Introduction and Guide*. Washington, D.C.

_________ (1997b). The subcommittee on cultural and demographic data. Washington, D.C. (www.census.gov/geo/www/standards/scdd/index.html).

United States, National Center for Health Statistics (1997). *Atlas of United States Mortality*. Washington, D.C.: Center for Disease Control and Prevention.

United States, National Research Council (2007). *Tools and Methods for Estimating Populations at Risk from Natural Disasters and Complex Humanitarian Crises*. Washington, D.C.: National Academy Press.

Weeks, J. (2007). Can we spot a neighbourhood from the air? Defining neighbourhood structure in Accra, Ghana. *GeoJournal*, 69:9-22.

White, J. D. (2007). *Managing Information in the Public Sector*. Armonk, New York: M. E. Sharpe.

Yankson, P., et al. (2004). A flexible model for urban vegetation cover measurement based on remote-sensing images. China: Wuhan University.

Annex I
Geographic information systems

A. Geographic information systems overview

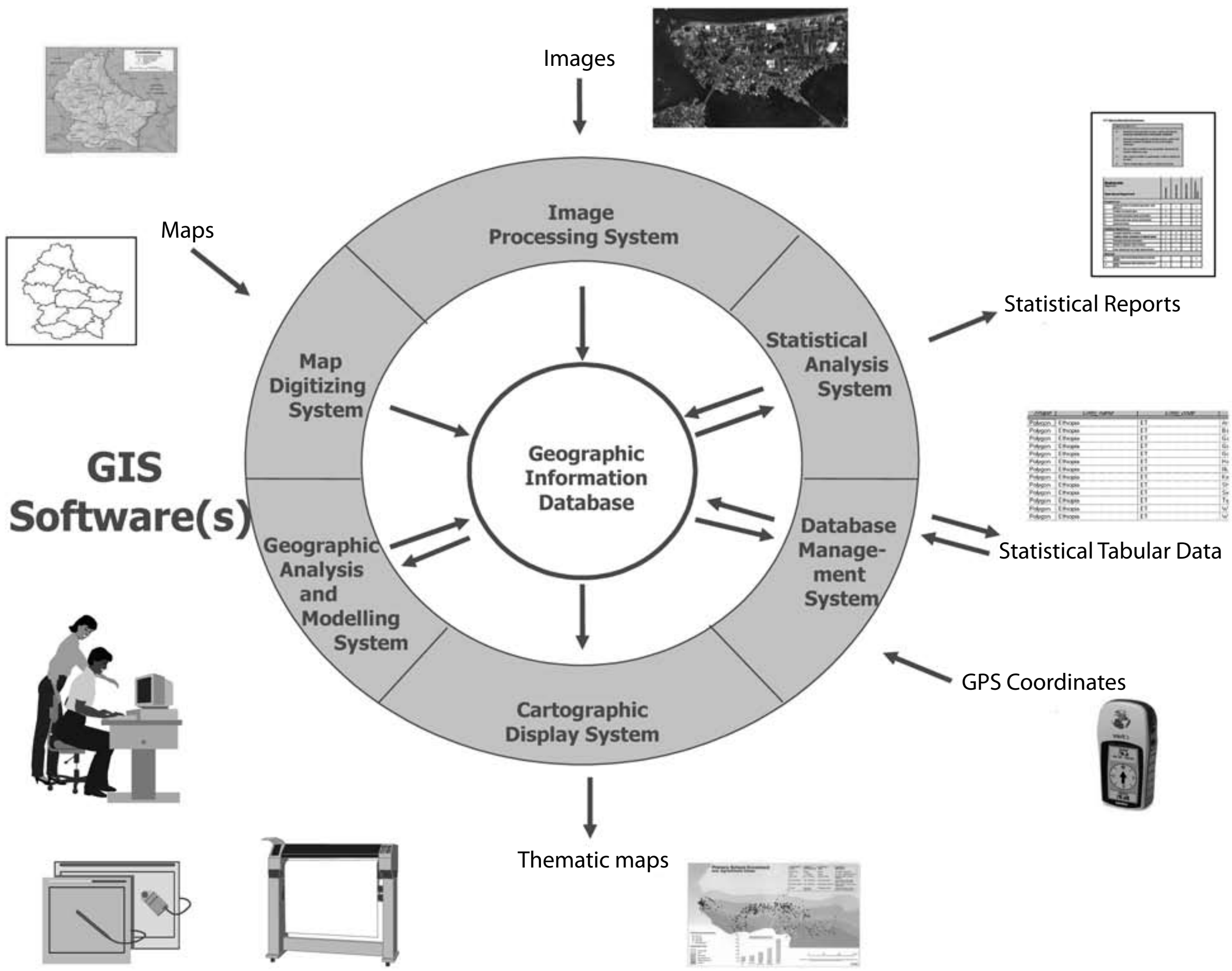
A1.1. A geographic information system (GIS) is a computer-based tool for the input, storage, management, retrieval, update, analysis and output of information. The information in a GIS relates to the characteristics of geographic locations or areas. In other words, a GIS allows us to answer questions about where things are or about what is located at a given location.

A1.2. The term "GIS" has different meanings in different contexts. It can relate to the overall system of hardware and software that is used to work with spatial information. It might refer to a particular software package that is designed to handle information about geographic features. It may relate to an application, for example a comprehensive geographic database of a country or region. Finally, it is sometimes used to describe the field of study that is concerned with methods, algorithms and procedures for working with geospatial data. For example, there are now GIS degree courses at many universities, and the term "geographic information science" is increasingly used to refer to academic research on computer-based geographic software and procedures. A related term, geospatial technology, is also a broader term that encompasses GIS, remote sensing, and global positioning system (GPS) technology.

A1.3. Several fields have contributed to the foundations of GIS, as illustrated in figure A1.1. The surveying and cartographic traditions have contributed the rules and tools for measuring and representing real-world features. Computer science provides the framework for storage and management of geographic information, and together with mathematics contributes the tools for manipulating the geometric objects that represent real-world geographic features. Populated with data from socio-economic, environmental and topographic surveys, a GIS supports applications in a very wide range of subject areas. These range from largely academic fields, such as archaeology or oceanography, to applied commercial applications, including marketing or real estate.

A1.4. Inventory-type applications are found in the utilities sector, where, for example, a phone company manages and maintains its physical infrastructure using a geographic database. Land titling systems operated by local and regional government agency are another example. In some fields, GIS is used to support data collection. The use of digital mapping and geographic databases for census operations and data dissemination is, of course, the most pertinent example in the context of the present *Handbook*. More analytical applications are found in the academic sector, but also in many applied fields, such as natural resource management or marketing. Forest companies, for example, use GIS to optimize sustainable harvesting of trees, and marketing or retailing companies use sophisticated spatial analysis to target customers or locate a new facility.

Figure A1.1
Foundations of GIS



Source: Ebener/WHO

1. Hardware, software and data

A1.5. Hardware and software issues are discussed in the present *Handbook* in the context of the construction and maintenance of geographic databases for census operations. Generally, the required hardware is no different from that used in other graphics-oriented applications that are characterized by large data volumes: a high end PC-compatible computer or workstation, a large, high-resolution display monitor and the usual input devices—keyboard and mouse. Large format digitizing tables or a scanner are used to convert paper maps into digital databases. Such tools are also used by architects or graphic designers. Large-format plotters and desktop printers are used for producing map output for display and visual analysis.

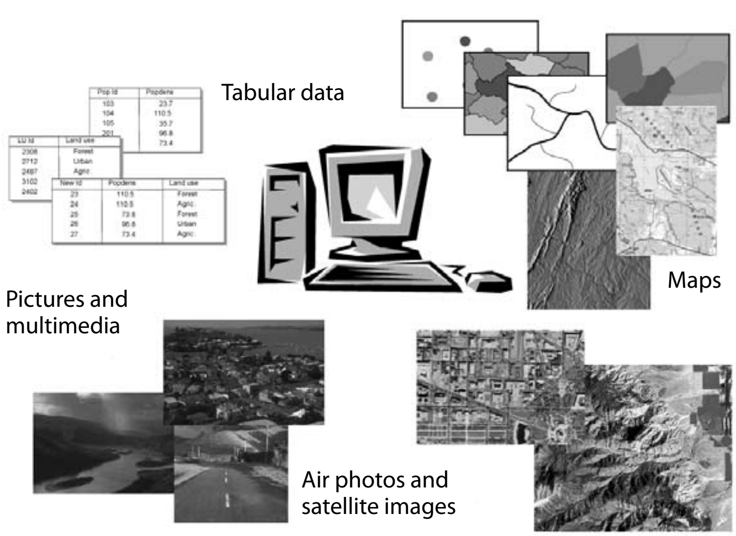
A1.6. GIS software has developed rapidly over its evolution from command line-oriented systems, which were very hard to learn, to menu-driven, easy-to-use packages that can be employed by anyone with minimum training. High-end packages are used by analysts who create new geographic databases and carry out advanced spatial analysis. At the medium level, there are a number of desktop mapping packages that combine a standard interface with a wide range of capabilities in terms of data input, management, analysis and output. Finally, at the low end, geographic data browser software is available. These do not allow the user to change the data, but provide many display functions. Such packages, some of which are distributed free of charge, are an excellent data distribution tool.

A1.7. Toolboxes of GIS-related software scripts or routines are a recent development sold by several GIS vendors. They allow the user to build tailor-made mapping applications within industry-standard object-oriented programming environments. These can be stand-alone systems, or they can be integrated into other software packages. Some of these products also include the tools for developing Internet-based mapping applications.

A1.8. Current GIS software trends are focused on Internet mapping and modular design that allows integration of geospatial functionality in any application. Users can now carry out data query and analysis on remote geospatial databases using their web browser and software that is downloaded on demand. For high-end applications, there has been a convergence of GIS and relational database management systems. Just as GIS packages use relational database management systems to store and manipulate the attribute information, some database management systems already include functions to store and manipulate geographic objects. The distinction between GIS and other information systems is disappearing.

A1.9. Data are what fuels GIS applications (see figure A1.2). Many of the most common geographic data sets are digital equivalents of paper maps, such as topographic maps showing roads, rivers, elevation contours and settlements as discrete layers. Thematic information includes socio-economic attributes referenced by administrative units; interpreted maps showing vegetation cover or land use; and derived indicators, such as catchment or watershed boundaries. Any geographic object shown on a digital map can be described in a data table linked to the digital spatial database. Sometimes a few attributes will be enough to characterize a set of features. In other cases, for instance for a census database, the attribute information stored in the system can be extensive.

Figure A1.2
Types of information stored in a GIS

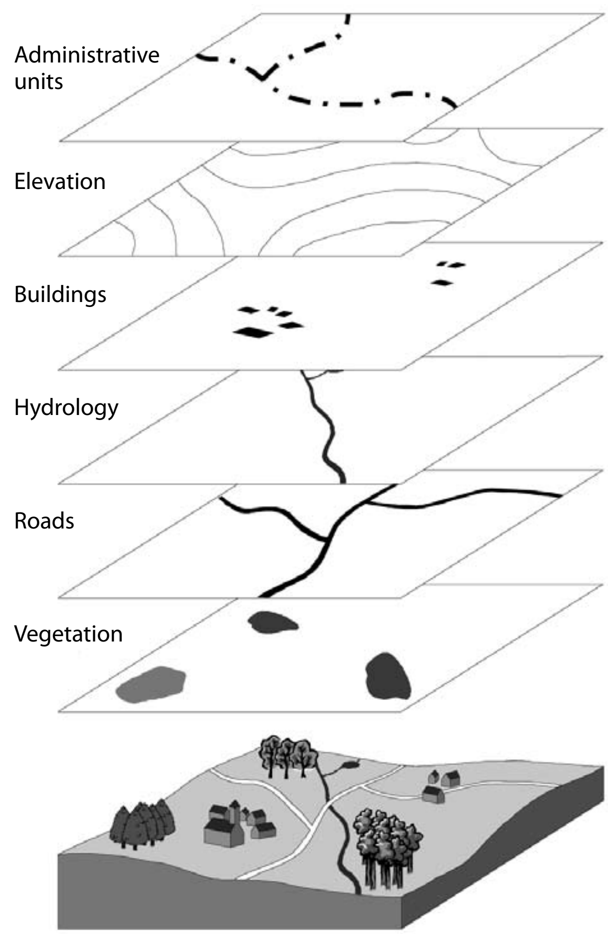


A1.10. Another source of geographic information comes from remotely sensed (i.e., aerial- or satellite-derived) imagery. Pictures or images taken from low-flying aircraft or from satellites are integrated with other spatially referenced information. Sometimes these images simply provide a backdrop for thematic or topographic map information. More often, however, information is interpreted and extracted from these images and stored as digital map information. Finally, multimedia information, such as photographs, video, text or even sound, is integrated in GIS also. Often, the integration is done by means of hotlinks. The user can interactively click on a feature to view photos or a video of the geographic location. This has been greatly facilitated by the emergence of easy-to-use Internet mapping programs, such as Google Earth.

2. Geographic data layers

A1.11. A GIS database is a computer-based representation of the real world. GIS software provides the tools for organizing information about spatially defined features. The basic organizational principle of a GIS is the data layer, as presented in figure A1.3. Rather than storing all spatial features in one place, as on a topographic map, groups of similar features can be combined in one of a number of these data layers.

Figure A1.3
Data layers—space as an indexing system



A1.12. A geographic database can include layers of physical features, such as roads, rivers and buildings, as well as layers of defined features, such as administrative boundaries or postal zones which cannot be observed on the ground. In addition, GIS software allows us to create new data layers based on existing ones. For example, a new data layer could show watersheds derived from digital elevation data or all areas that are within a specified distance of a hospital.

A1.13. In the process of creating a multi-layer geographic database, features might be extracted from a range of different topographic and thematic sources. In addition, field observations and remotely sensed data from satellites or air photos are often integrated with the map data. GIS provides the tools that integrate all these different data sets within a common reference framework which is defined by the geographic coordinate system. This allows the user to combine different types of data, create new information or execute complex queries that involve several data layers. The ability to integrate data from heterogeneous sources by using geographic location as the link is sometimes referred to as using "space as an indexing system". This is indeed one of the most important benefits of GIS.

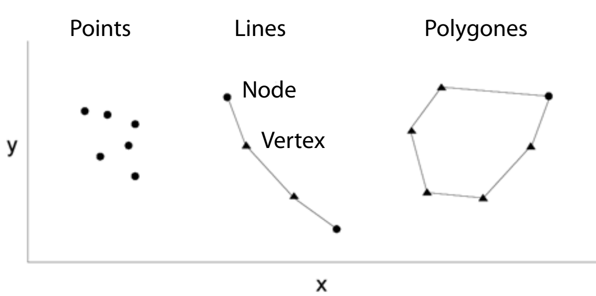
B. GIS data models

A1.14. Despite the heterogeneity of the information that can be stored in a geographic database, there are only a few common methods of representing spatial information. In developing a GIS application, real-world features need to be translated into simplified representations that can be stored and manipulated in a computer. Two data models—internal digital representations of information—currently dominate commercial GIS software, although many programs now handle both data models easily. The "vector" data model is used to symbolize discrete features, such as houses, roads or districts; and the "raster" data model is most often used to represent continuously varying phenomena, such as elevation or climate, but is also used to store pictures or image data from satellites and aircraft-based cameras. For census applications, the vector data model is usually more useful, although many auxiliary data sets are more appropriately stored using the raster model.

1. Vector

A1.15. Vector GIS systems represent real-world features using a set of what are called geometric primitives: points, lines and polygons (see figure A1.4). A point is represented in a computer database by an x, y coordinate. A line is a sequence of x,

Figure A1.4
Points, lines and polygons



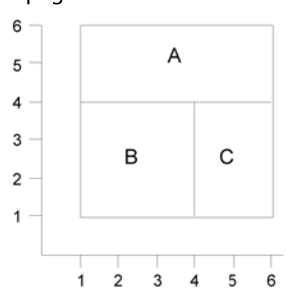
y coordinates, whereby the end points are usually called nodes and the intermediate points are termed vertices. Polygons or areas are represented by a closed series of lines, such that the first point equals the last point of the loop. Points might be used to represent houses, wells or geodetic control points; lines describe such features as roads and rivers; and enumeration areas or districts, for example, are represented by polygons.

A1.16. The simplest vector data models simply store the data without establishing relationships among the geographic features (see figure A1.5). This is sometimes called the "spaghetti model" (see, for example, Aronoff, 1991), since lines in the database overlap but don't intersect, like spaghetti on a plate. More sophisticated "topological data models" store relationships among different features in a database. For example, lines that cross will be split and an additional node will be added at the intersection. Instead of defining the boundary between neighbouring polygons twice—once for each closed-loop polygon—the line is stored only once together with information on which polygons are located to the right and left of the line, respectively. Information about the relationships between nodes, lines and polygons are stored in attribute tables.

A1.17. The advantage of the topological model becomes clear when we think about what questions we might ask from a spatial database. A topologically structured spatial database allows rapid queries on individual data objects and their relation to other data objects. For example, to quickly identify all neighbours of a particular EA, the system would simply go through the list of lines that define this EA and find all of the remaining EAs which are also bounded by these lines.

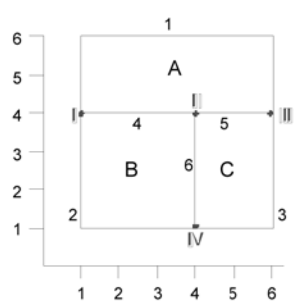
Figure A1.5
Vector data models: spaghetti versus topological

"Spaghetti" data structure



Poly	Coordinates
A	(1,4), (1,6), (6,6), (6,4), (4,4), (1,4)
B	(1,4), (4,4), (4,1), (1,1), (1,4)
C	(4,4), (6,4), (6,1), (4,1), (4,4)

Topological data structure



O = "outside" polygon

Node	X	Y	Lines
I	1	4	1,2,4
II	4	4	4,5,6
III	6	4	1,3,5
IV	4	1	2,3,6

Poly	Lines
A	1,4,5
B	2,4,6
C	3,5,6

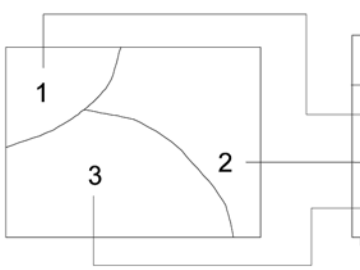
Line	From Node	To Node	Left Poly	Right Poly
1	I	III	O	A
2	I	IV	B	O
3	III	IV	O	C
4	I	II	A	B
5	II	III	A	C
6	II	IV	C	B

A1.18. Most GIS packages now employ fully topological data structures that allow complex operations, such as polygon overlay. In this operation, two vector data sets are combined—for example, administrative districts and watershed boundaries. New, smaller polygons are created by intersecting the polygons from both input data sets. Most desktop-mapping systems use simpler data structures. In these, all polygons are defined as closed loops, such that the lines that define the boundary between two districts will be stored twice in the database.

A1.19. Every feature in the database is labelled internally with a unique identifier that links the geometric feature with a corresponding entry in a data or attribute table (see figure A1.6). The user can add information about each feature to the corresponding database record. For points representing houses, the user might list the mailing address, the type of house, and whether electricity and sanitary facilities are available. In a database of enumeration areas, the user might add the official administrative code, the number of dwelling units and any census data that have been compiled for the EA. For practical purposes, most GISs use a relational database model to store the attribute or non-spatial information separately in a database. The attribute files are closely integrated with the digital geographic data and can be accessed through the GIS or through a relational database management system (RDBMS).

Figure A1.6
Spatial and non-spatial data stored in a vector GIS

Digital map



Geographic attribute table

Poly-Id	Area	Pop	TFR
1	297	4583	2.7
2	607	3927	3.1
3	806	9271	3.8

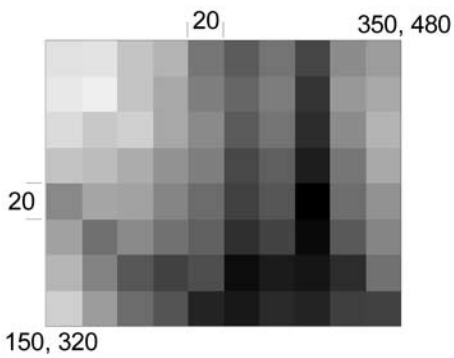
A1.20. Between the two extremes—the simple spaghetti and the complex fully topological model—some desktop-mapping packages have found a compromise. While not fully topological, these systems allow the quick computation of neighbourhood and connectivity information. They thus combine the ease of editing of the simple data model with elements of the powerful analytical capabilities of a topological vector GIS data model.

2. Raster

A1.21. Raster GIS packages divide space into a regular array of rows and columns. A cell in this array or grid is sometimes called a pixel, which stands for picture element. It reveals the origin of this data model in remote sensing and image processing. In most raster systems, the attribute value at a given location, for example its elevation, is stored in the corresponding cell of the raster. The raster database of elevation is thus simply a long string of elevation numbers. The only additional information required by the system is the number of rows and columns in the raster image, the size of the raster cells (which are usually square) in real-world units (e.g., metres or feet) and the coordinates of one of the corners of the entire raster (see figure A1.7). This information is usually stored in a header or in a small separate file. These pieces of information let the system calculate the grid dimensions. For instance, the x- coordinate of the upper right corner is $150 + 10 \times 20 = 350$. The system can use this information to register the raster grid correctly with other geographic data layers, for example to draw vector features on top of the grid.

Figure A1.7
Example of a raster data file

Raster map—elevation



ASCII raster file									
Number of columns		10							
Number of rows		8							
Lower left corner x		150							
Lower left corner y		320							
Cell size (square cells)		20							
219	313	407	462	681	783	689	877	595	540
297	274	407	501	642	744	650	955	556	501
336	391	368	501	603	783	689	994	595	462
414	430	485	579	642	861	767	1072	673	501
609	508	524	618	720	900	806	1267	712	579
531	703	602	696	759	978	884	1189	790	618
453	625	797	891	837	1173	1079	1111	985	696
375	547	719	813	1032	1095	1001	1033	907	891

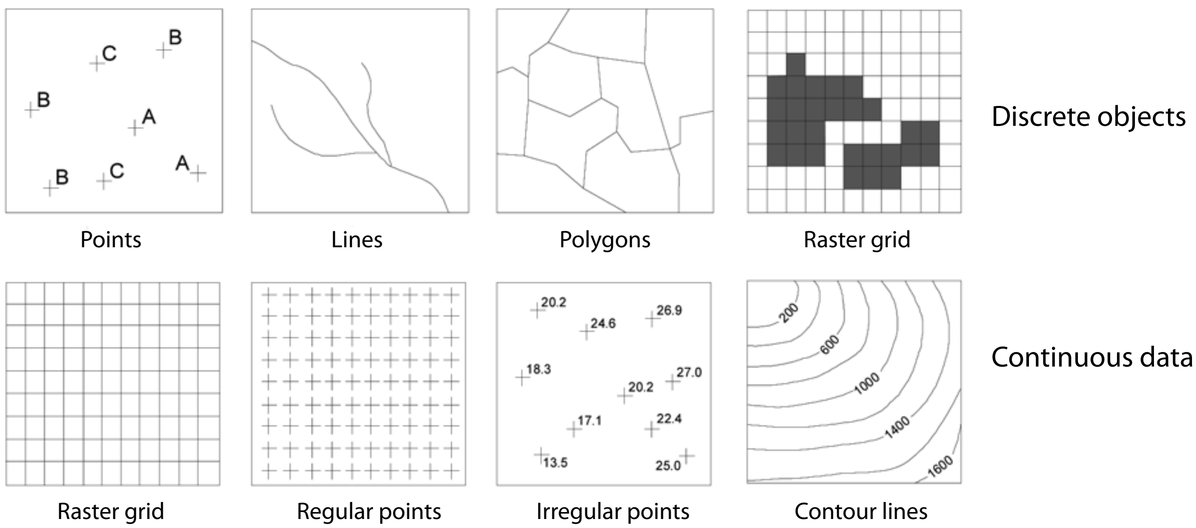
A1.22. This data-storage method is of course very inefficient if there are many cells with similar values in the raster. For example, discrete objects are also sometimes stored in raster format. A district map in raster format would show in each cell the district identifier or the total population of the district into which the cell falls. Obviously, there will be many contiguous cells with the same value. Most raster GIS programs therefore use some form of data compression. The simplest of these is run-length encoding, where the system stores pairs of two numbers: the data value and the number of times the value is repeated. This can reduce file sizes significantly.

A1.23. Raster data are most often used to store continuously varying data or images that show many continuous gray tones. Just as discrete objects can also be shown in raster format, continuous data can also be represented using vector data structures. The best example are contour lines which show elevation on topographic maps. Other examples are shown in figure A1.8.

3. Advantages and disadvantages of vector and raster data models

A1.24. The strength of the raster data model is its simplicity. Many operations on geographic data are easier to implement and execute faster in a raster GIS. Modelling of continuous data, as is often done with elevation or hydrological data, is usually performed with a raster GIS. One disadvantage is that there is a trade-off between the size of the resulting raster data sets and the precision with which spatial features can be represented. A very fine raster grid will represent all curves in a boundary with sufficient detail, but will require a large amount of disk space. For geographic analysis using "objects", such as points, lines and polygons, vector models work better.

Figure A1.8
Vector and raster data models can both be used to display discrete and continuous data



A1.25. Most GIS operations can be performed on both data models. Which data model is appropriate depends on the application. For census and many other socio-economic applications, the vector model is more appropriate. Vector data structures allow a more compact representation of points and polygons which define socio-economic objects. The close connection to database management systems supports socio-economic applications that are characterized by a large amount of attribute information—for example hundreds of census or survey variables—that is tied to a fixed number of spatial features, such as census districts, villages or survey clusters. Finally, printed output from vector GIS databases usually resembles more closely maps produced using traditional cartographic techniques.

A1.26. Even so, the capability to handle raster data is of increasing importance in population applications. Some of the input data that are useful for delineating enumeration are boundaries come in raster form. Chapter 4 of the present *Handbook*, for example, discusses the use of remote sensing images to create or update census cartography. Fortunately, the choice between data models usually does not have to be “either or”. Many GIS packages now support both types of spatial data. This, for example, allows the use of raster data as a background onto which line and polygon features can be drawn. Thus, remotely sensed images or elevation surfaces can be displayed on a computer screen together with other relevant information to aid the delineation of enumeration areas.

4. Precision versus accuracy

A1.27. The terms precision and accuracy are often used interchangeably, even though they have different meanings. Accuracy in a spatial context refers to the degree to which a representation’s location corresponds to the true location on the Earth’s surface. Precision, in contrast, refers to the ability to distinguish between small quantities or distances in measurement. For example, if our surveying tools measure coordinates only in metres, the point locations in our GIS will only be precise to the nearest metre. If we have a more precise measuring tool, we can obtain point coordinates that are precise to the nearest centimetre or millimetre.

A1.28. In practice, the precision by which coordinates can be stored in a vector GIS is virtually infinite, because they use double precision data types (8- bytes for each floating point number) for storing the geographic coordinates. The accuracy of spatial coordinates, however, depends largely on the data-collection tools. The best surveying instruments that are used for engineering applications or research on plate tectonics achieve accuracy of less than a millimetre. Most data used in GIS, however, come from data sources with much lower accuracy, such as paper maps, hand-held global positioning systems or even cartoon maps sketched during fieldwork. Here, the accuracy is likely to be measured in metres rather than millimetres.

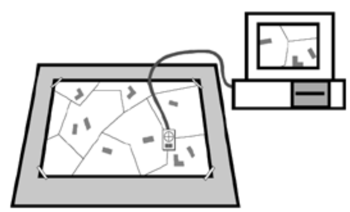
C. GIS capabilities

A1.29. An overview of GIS capabilities is set out in the table below. This list is by no means complete, since high-end GIS packages and even desktop-mapping packages offer numerous specialized functions for data entry, manipulation, analysis and display.

Data input and management

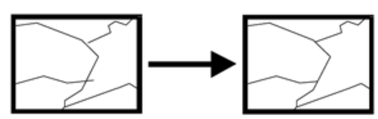
Line tracing, coordinate data input

The most common form of coordinate data entry is still by means of a digitizing table. Lines are traced on the paper map with a cursor and captured in the GIS or digitizing software. Alternatively, maps can be scanned to create raster bit-maps which are then converted into vector format.



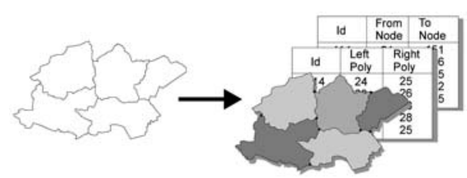
Editing

After lines have been digitized, the data have to be checked for errors. Common problems include unconnected lines (undershoots and overshoots), missing lines or lines that have been digitized twice. Some of these operations are automated in GIS.



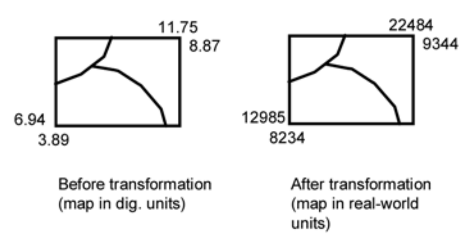
Building topology

Digitized or vectorized lines do not have any relationships to each other. GIS software can compute neighbourhood relations and connectivity between features in the data set.



Georeferencing and projection change

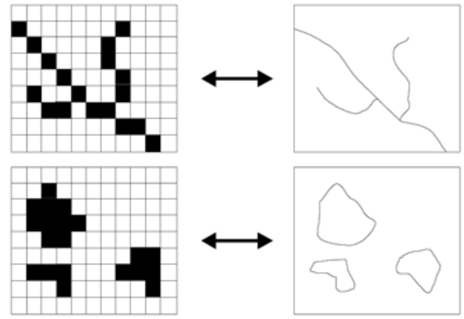
Digitized lines are in centimetres or inches. They need to be converted into real-world units corresponding to the coordinate system of the source map, such as metres or feet. For data integration, the projection of the digital maps may also have to be changed.



Raster-vector transformation

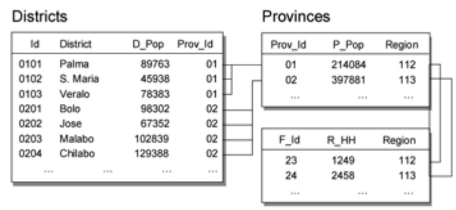
Most commercial GIS packages now support raster imagery in some form. Since each data model is appropriate for different tasks, functions to convert one into the other are needed. Raster-to-vector transformation is also used for the automatic conversion of scanned maps.

The opposite operation—vector-to-raster transformation—is required for analysis and modeling in a raster GIS.



Attribute data management

Each feature in the database is labeled with a unique identifier. This identifier is used as a link to external information about the geographic features. To enable manipulation and analysis of attribute tables, the GIS is usually integrated with a relational database management system.



Districts

Id	District	D_Pop	Prov_Id
0101	Palma	89763	01
0102	S. Maria	45938	01
0103	Veralo	78383	01
0201	Bolo	98302	02
0202	Jose	67352	02
0203	Malabo	102839	02
0204	Chilabo	129388	02
...	...	...	...

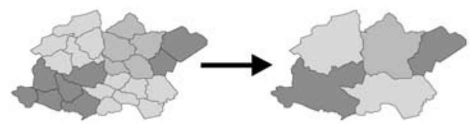
Provinces

Prov_Id	P_Pop	Region
01	214084	112
02	397881	113
...	...	...

F_Id	R_HH	Region
23	1249	112
24	2458	113
...	...	...

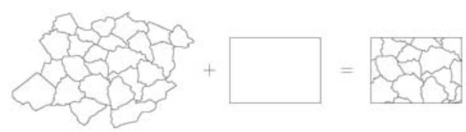
Reclassification, aggregation

GIS allows the aggregation of features based on a common identifier. For example, enumeration areas can be grouped into operational census areas of approximately equal population size.



Subset creation, clipping

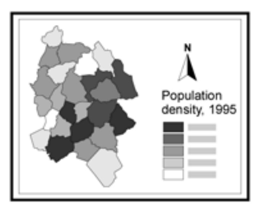
Apart from subset selection based on queries, GIS can also create custom subsets using clipping operations.



Display

Cartographic functions

Producing map output for presentation is only one application for cartography in GIS. Cartographic symbolization is also important to distinguish features in on-screen editing and analysis.

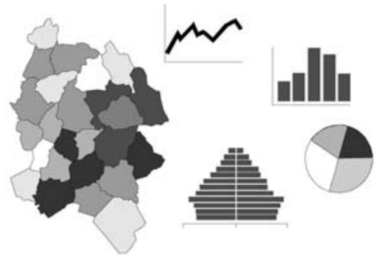


Combined display of image and vector data

Image or raster data come from various sources: scanned maps, remotely sensed images and raster GIS data are all stored in some form of grid format. Displaying vector and raster data in combination can provide valuable context for analysis and enables selective extraction of features from the raster data.

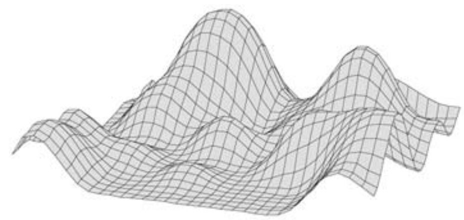
Link to statistical charting

Data-driven analysis of spatial data will usually be a combination of mapping and examination of attribute data. Statistical graphs are especially valuable if they can be displayed on the maps.



3-D display of surfaces

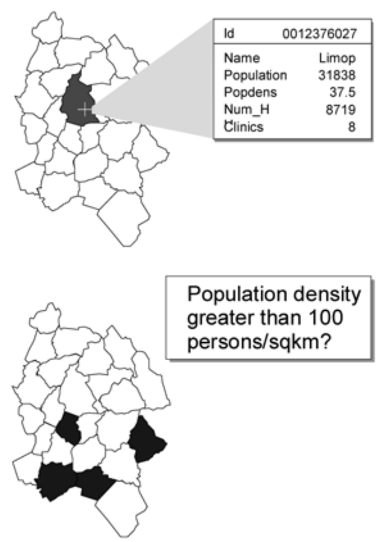
Continuous data, such as elevation or precipitation—and to some extent also population density—can be displayed in various formats: raster grids, contour lines or simulated three-dimensional visualizations using wire frames, onto which other features can be draped.



Query

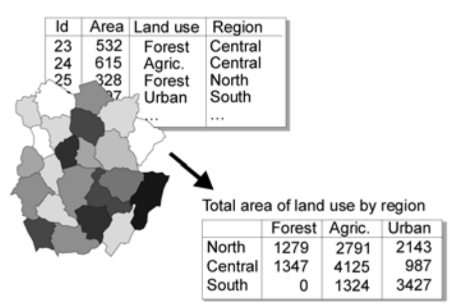
Spatial database query

"What is at …?" and "Where is …?" are the most fundamental geographic questions that GIS can answer. In simple browsing mode, a user can select features on a digital map and obtain information about them. Conversely, the user can select features that match a set of criteria and display those on the map. GIS are usually linked to database management software and query operations are based on the SQL concept. GISs also allow queries based on geographic relationships, such as distances ("What is within x km of this place?") or queries based on two or more GIS data layers ("Which buildings are located in this enumeration area?").



Summarizing attributes

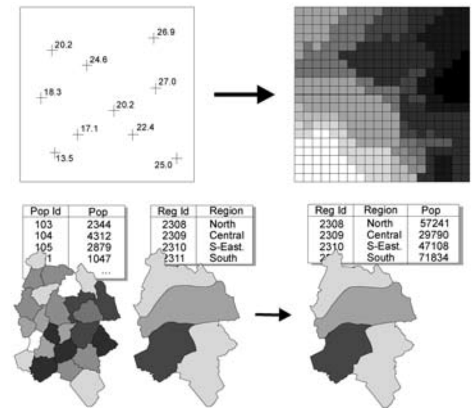
Database operations allow us to extract useful summary statistics or cross-tabulations from the geographic attribute table of a GIS data set. For instance, we can compute the minimum, maximum and average value of a field in the table. Or we can cross-tabulate two or more fields in the table and produce summary totals of a third field for each combination of attribute categories. This allows us, for instance, to compute the total area of each land-use class in the regions of a country. Cross-tabulations are often used after two or more GIS layers have been combined by a polygon overlay operation (see below).



Spatial data transformations

Interpolation

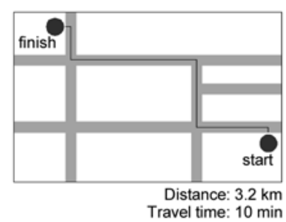
Interpolation allows us to create a modeled coverage from sample data. For example, based on a set of station precipitation surfaces, we can create a raster surface that shows rainfall in the entire region. More important for socioeconomic applications is areal interpolation. For example, using population by district, we may wish to estimate population for environmental monitoring regions whose boundaries do not match the districts.



Distance and time operations

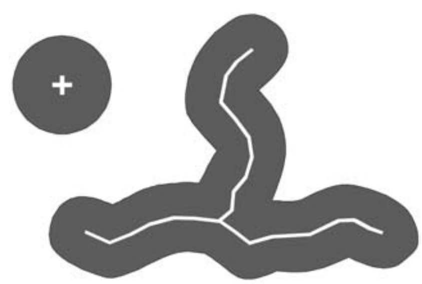
Simple distance computations

Distance computation is one of the fundamental GIS operations. Distances (and times) can be computed as straight line or as network distances. Based on a GIS roads database, for example, distances and travel times can be estimated.



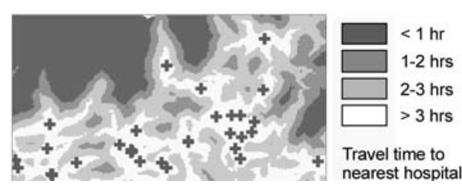
Buffer

A special type of distance operation is the creation of buffer regions. Buffers can be created around points, lines or polygons and can be weighted by attribute values. For example, surfaced roads could get a wider buffer than dirt roads. Buffers are often used in spatial queries. For instance, to identify the number of bilharzia cases within 3 km from a river, a buffer, point in polygon and database query would be performed in sequence.



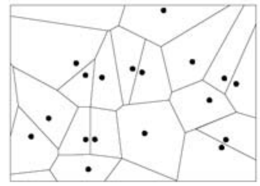
Finding the nearest feature

A combination of database query and distance computation is where we need to identify the closest of a number of features of a given category. For example, we may wish to compute the distance to the nearest hospital for all locations in a district. The resulting GIS data set is often called an accessibility surface.



Thiessen polygons

A variant of the "find nearest feature" function is an operation where the entire region is partitioned into polygons that are assigned to the nearest facility. The resulting area units are called Thiessen polygons. This function is often used to create simple catchment or service areas.

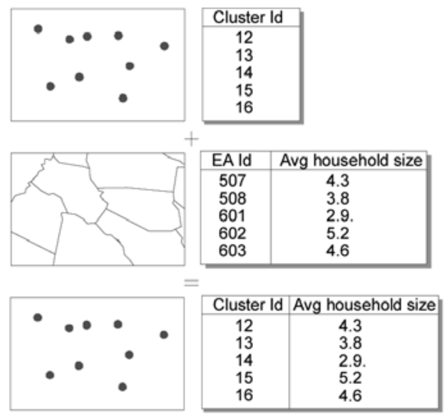


Combination of data layers

Point or line in polygon operation

Many questions that GIS can help answer require the combination of several data sets. For example, we may have a set of point coordinates representing clusters from a demographic survey and we may wish to combine the survey information with data from the census which is available by EA. GIS will identify for each point the EA into which it falls and will attach the census data to the attribute record of that survey point.

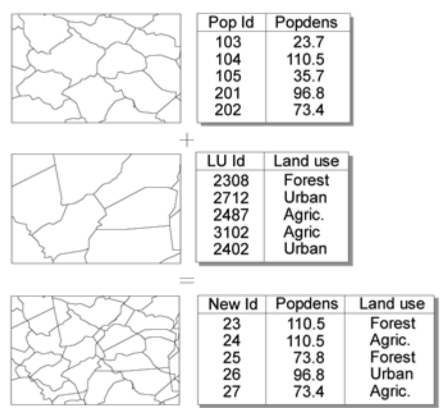
The same operation allows us to summarize an attribute of point or line features for a set of regions. For example, we can determine the average fertility rate for each health district using a sample of surveyed households (points).



Polygon overlay

Combining two GIS data sets of area features is called polygon overlay. The system will merge the data sets and create new area units from the areas of overlap. The resulting new data set will contain the attributes of both data sets. It depends on the data types whether the attribute should remain unchanged (e.g., categorical information or ratios) or should be divided over the new polygons (e.g., count data).

Polygon overlay is often used in combination with cross-tabulations, for example, to compute census data by land-use zone.



Annex II
Coordinate systems and map projections

A. Introduction

A2.1. The review of GIS concepts in annex I has highlighted the benefits of spatial data integration. By organizing different types of geographic information as data layers, measurements, queries, modelling and other types of analysis can be performed that makes use of data from many different subject areas. Thus, census data can be analysed in combination with land-use or agro-ecological data, or socioeconomic survey information can be linked to geographically referenced data on disease risk. This ability of linking data from numerous sources is made possible by the vertical integration of different data layers. That simply means that all geographic data sets are referenced using the same coordinate system, so that different data layers align correctly when overlaid on top of each other.

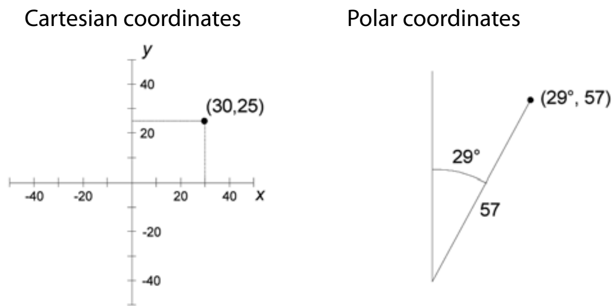
A2.2. In building a geographic database—for instance, for a census—the data developer must ensure that spatial coordinates and boundaries captured from hardcopy data sources, digital gazetteers or during fieldwork are registered in a proper coordinate system in a process referred to as “georeferencing”. This will also ensure that digital maps that were developed separately for neighbouring regions will match perfectly when displayed together on a computer screen or printed page.

A2.3. For census-mapping using traditional techniques, this was less of a concern since the paper maps—often sketch maps drafted in the field—were used for enumeration purposes only. They were not integrated with other data and not used for any type of spatial analysis. Knowledge of coordinate systems and map projections were thus much less important than they are when building a digital database that is meant to serve many different purposes. This annex provides a brief review of important cartographic concepts. Cartography textbooks, such as Robinson and others (1995), Kraak and Ormeling (1996) and Dent (1998) provide much additional information. More specialized treatments on the topic can be found in Canters and Decleir (1989), Snyder (1993) and Bugayevskiy and Snyder (1995).

B. Coordinates

A2.4. In cartography, the method by which positions of objects on the Earth’s surface are measured is called the geographic coordinate system. This is sometimes also referred to as the geographic referencing system. In two-dimensional geometry, the most common coordinate system is the so-called “Cartesian coordinate system” named after the French scientist René Decartes (1596-1650). Coordinates are given as perpendicular distances on two fixed axes (x and y) measured from a fixed origin. This is the system used in GIS and also in more general computer graphics applications. An alternative method for defining positions is the “polar coordinate system”, which measures the angle and distance from a fixed point of origin (see figure A2.1).

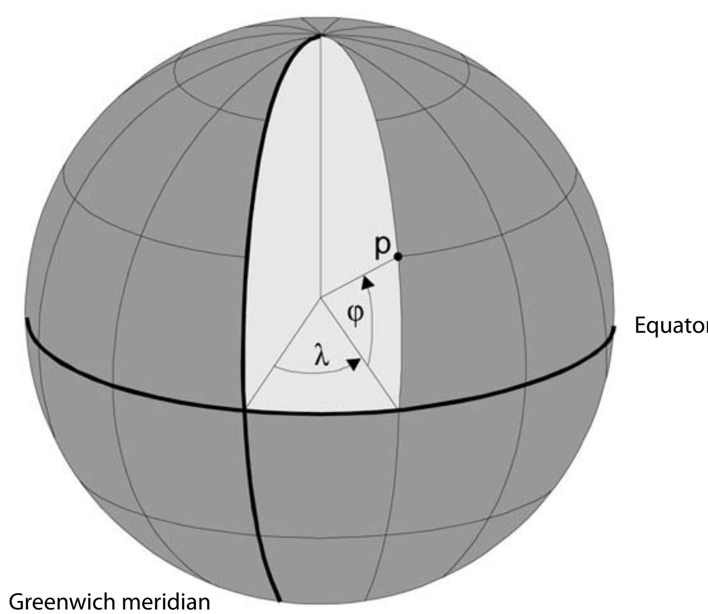
Figure A2.1
Planar and polar coordinate systems



A2.5. A flat map, whether on paper or a computer screen, shows coordinates in a planar, two-dimensional coordinate system where the coordinates are measured in standard units, such as metres or feet. The coordinates are usually termed "x" and "y" coordinates, although the terms "easting" and "northing" are often used in cartographic texts. However, the objects on a map are a representation of features that are located on the Earth's surface. Since the Earth is a sphere, coordinates on the Earth's surface are measured in a spherical coordinate system. More specifically, we usually use latitude and longitude coordinates to reference positions. This is a spherical polar coordinate system, where any point p is defined as the angle of latitude, ϕ, relative to the plane defined by the equator and the angle of longitude, λ, measured relative to the plane defined by the zero or Greenwich meridian (see figure A2.2).

A2.6. To produce paper maps of the world or some part thereof, these spherical latitude and longitude coordinates need to be translated in some way into a planar coordinate system. A recent book on map projections calls this process of producing a two-dimensional representation of a part of the three-dimensional globe as "flattening the Earth" (Snyder, 1993).

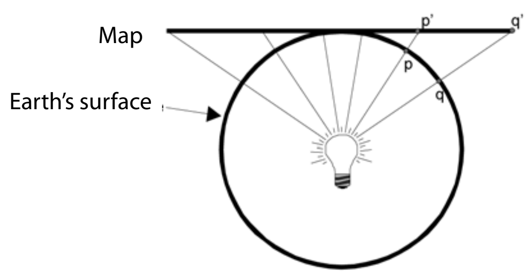
Figure A2.2
Coordinates on the sphere: the latitude longitude reference system



Map projections

A2.7. The mathematical procedure by which the spherical latitude and longitude coordinates are translated into planar coordinates is called cartographic projection. We can literally think of this process as a projection by imagining a light source that is located, for example, in the centre of the Earth. If the Earth's surface were transparent with only the features of interest outlined, we could simply place a flat piece of paper on top of the Earth and retrace the projected features on this so-called developable surface. For example, a feature located at point p on the Earth's surface would be placed on point p' on the map. As we see in figure A2.3, the further a point is located away from the location where the map touches the globe, the more its relative distance from points closer to the tangent point will be distorted. For example, the distance between p and q on the globe is much smaller than the distance between p' and q' on the map. Points located at the equator cannot be projected at all using this specific approach since the light rays passing through the equator run parallel to the map. This particular projection method is therefore only useful for areas that are relatively close to the tangent point.

Figure A2.3
Illustration of the map projection process (azimuthal projection)



A2.8. Over the centuries, cartographers have developed many different map projections, which can be classified according to the way in which the map is placed on or around the globe. Figure A2.4 provides an overview, showing how three types of map projections—cylindrical, conical and azimuthal—are constructed. As the map graticule on the right shows, each family of map projection gives rise to a characteristic pattern of latitude/longitude grid lines.

A2.9. A cartographer can also choose the location at which the "developable surface"—the cylinder, cone or plane—touches the globe. This tangent line or point is usually the area where distortions of size and shape are minimal. If we produce maps for a specific region of the world, we can thus choose the "aspect" of the map projection to optimize the map representation for our area of interest.

A2.10. The hypothetical light source is not always located at the centre of the globe (see figure A2.5a.), but may be located at the far pole (see figure A2.5b.), or we could imagine a series of light sources that emit light from a flat base parallel to the map rather than from a point source (see figure A2.5c.). In cartographic terminology, these projection methods are called "gnomonic", "stereographic" and "orthographic", respectively. As we can see from looking at where the projected points p' and q' end up on the map, each of these assumptions will lead to a different type of distortion of the relative position of locations that are represented on the map.

Figure A2.4
Map projection families

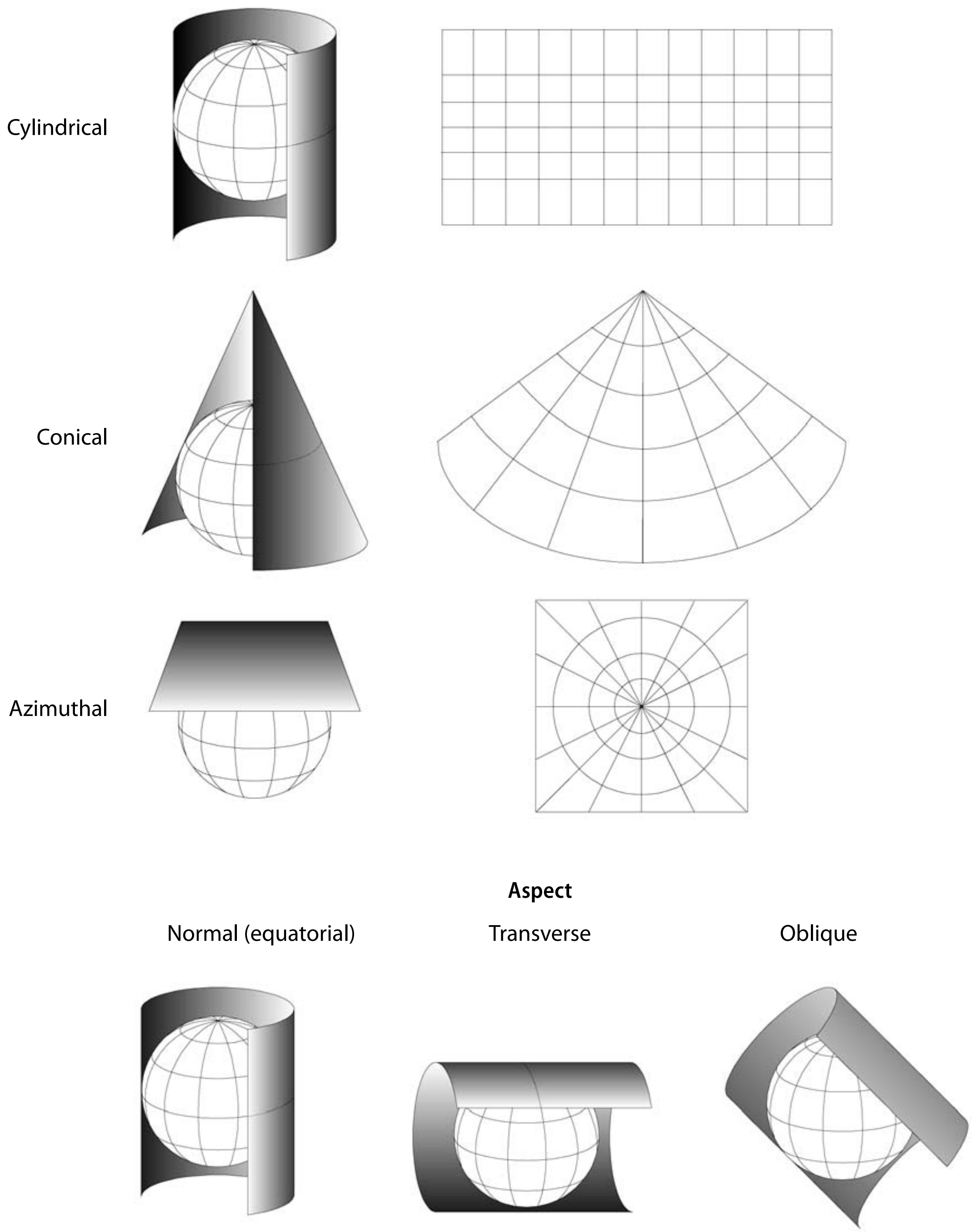
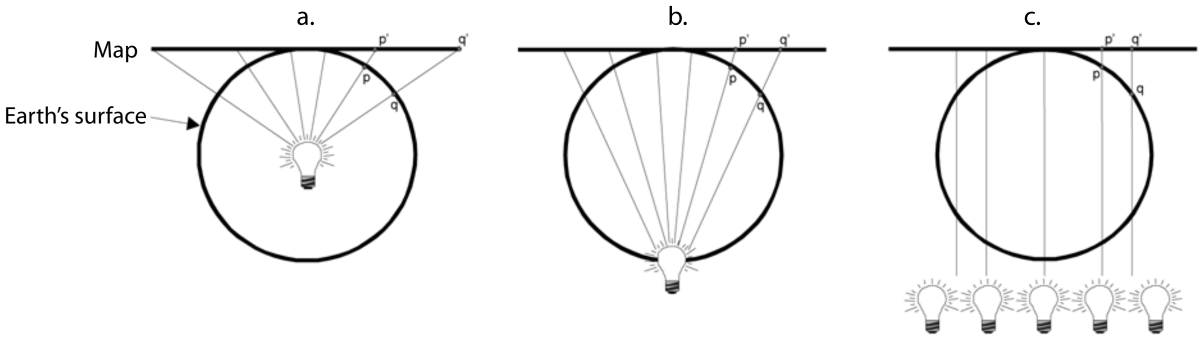


Figure A2.5
Different ways of constructing the projection



C. Map projection properties

A2.11. Although the imaginary light source is a good way of showing the principle of map projections, they are in practice of course defined mathematically. Given the latitude and longitude of a location, a formula is used to obtain the corresponding point in the projected planar coordinate system. The cartographer has many different options in creating a map projection that will have specific characteristics. The way in which the developable surface is arranged around the globe, the aspect and the position of the imaginary light source are only some of the possible parameters.

A2.12. Unfortunately, there is no perfect way of representing spherical coordinates on a flat map. Consequently, no map projection can serve all purposes. Each one is good at preserving some characteristics but bad at others. Depending on the projection method, different kinds of distortions will be introduced. Map projections are therefore classified according to which property they preserve. The most important are:

- **Correct areas.** Most projections stretch area features on the map. This stretching is usually not constant across the map, so that features close to the poles on a world map, for example, often appear relatively larger than features closer to the equator. For example, the Arabian peninsula is several hundred thousand square kilometres larger than the island of Greenland. On many maps, however, Greenland appears to be several times larger than the Arabian peninsula. Maps that show the relative area of all features correctly are called equal-area projections. *Example*: Mollweide projection.
- **Equal distance.** No map projection can represent distances between all points on the map correctly. This is important to remember since a common application of GIS databases is to compute distances. For large-scale mapping in a small geographic region, the errors introduced are usually negligible. For national or continental applications using small scale maps, however, the distances calculated by a GIS are not reliable unless the system compensates for the error introduced by Euclidean distance calculation at this scale. Even equidistant projections do not show all distances correctly, but they can accurately represent all distances from one or two points to all other points, or along one or more lines. *Example*: Equidistant conic projection. It should be noted that very accurate distance calculations are typically made using exact geometric formulae rather than simple Euclidean distance. These calculations are based on latitude and longitude coordinates to compute the so-called "great-circle distance".

- **Correct angles.** Conformal projections preserve the angles around all points and shapes over small areas. Meridians and latitudes intersect at right angles. These projections are most useful in navigation. *Example*: Mercator projection.

A2.13. Thus, all map projections represent a compromise between desirable cartographic characteristics. For any given application, there will therefore be map projections that are more appropriate than others. In addition to map projection properties, issues to consider are the size of the region to be mapped, its primary extent (e.g., north-south versus east-west), and the location of the area on the globe (e.g., polar, mid-latitude or equatorial).

A2.14. Cartography textbooks and many GIS manuals have comprehensive lists that show which applications are best served by which map projection. In some instances, the best choice may be a projection that does not preserve any property perfectly. The Robinson projection that is popular for global maps, for example, is a compromise projection that was designed mostly for aesthetic purposes, such as atlas mapping. In other instances, for example where only a relatively small area is mapped, the distortions introduced by any projection may be negligible for a given application.

A2.15. Figure A2.6 shows some popular map projections. At the top of the figure, the Earth is shown as a sphere and in unprojected latitude and longitude coordinates which are mapped as if they were planar coordinates. Incidentally, many GIS data distributors disseminate digital map data in unprojected, "geographic" coordinates because it is usually straightforward for a user to convert latitude and longitude coordinates into any map projection system, but sometimes more difficult to go from one map projection to another.

D. More precise mapping: geographic datums

A2.16. Complicating the conversion from spherical latitude/longitude coordinates into planar coordinates is the fact that the Earth is not a perfect sphere with a constant radius. Precise measurements show that the Earth's surface is highly variable and constantly changing. Most importantly, the Earth is flattened at the poles, so that the distance from the Earth's centre to the North Pole (the semi-minor axis) is smaller than that to the equator (semi-major axis). For precise mapping purposes, the globe is therefore more accurately described as an ellipsoid or spheroid with a specified relationship between the polar and equatorial radius (see figure A2.7). The parameters that describe the ellipsoid and the origin and orientation of the coordinate system used to reference map features is called a "geodetic datum" (after the science of the Earth's measurement: geodesy).

A2.17. The most appropriate parameters that approximate the ellipsoid vary across the globe. Consequently, hundreds of datums have been defined. Fortunately, each national mapping agency usually uses only one standard datum for all its mapping and geospatial activities and only a few are in use for regional, continental or global mapping. Complications occur where the standard datum is changed by a mapping agency. Datums have been refined continuously over the last two centuries, so that older maps for a place may be based on one datum while newer ones have been compiled using a newer and more accurate one.

A2.18. For small-scale mapping covering a large region or for the preparation of sketch maps in applications that do not require high accuracy, the issues introduced by different datums are negligible. For more precise mapping at large scales, however, the offset can be quite significant. Table A2.1 shows the coordinates of United Nations

Figure A2.6
Common map projections

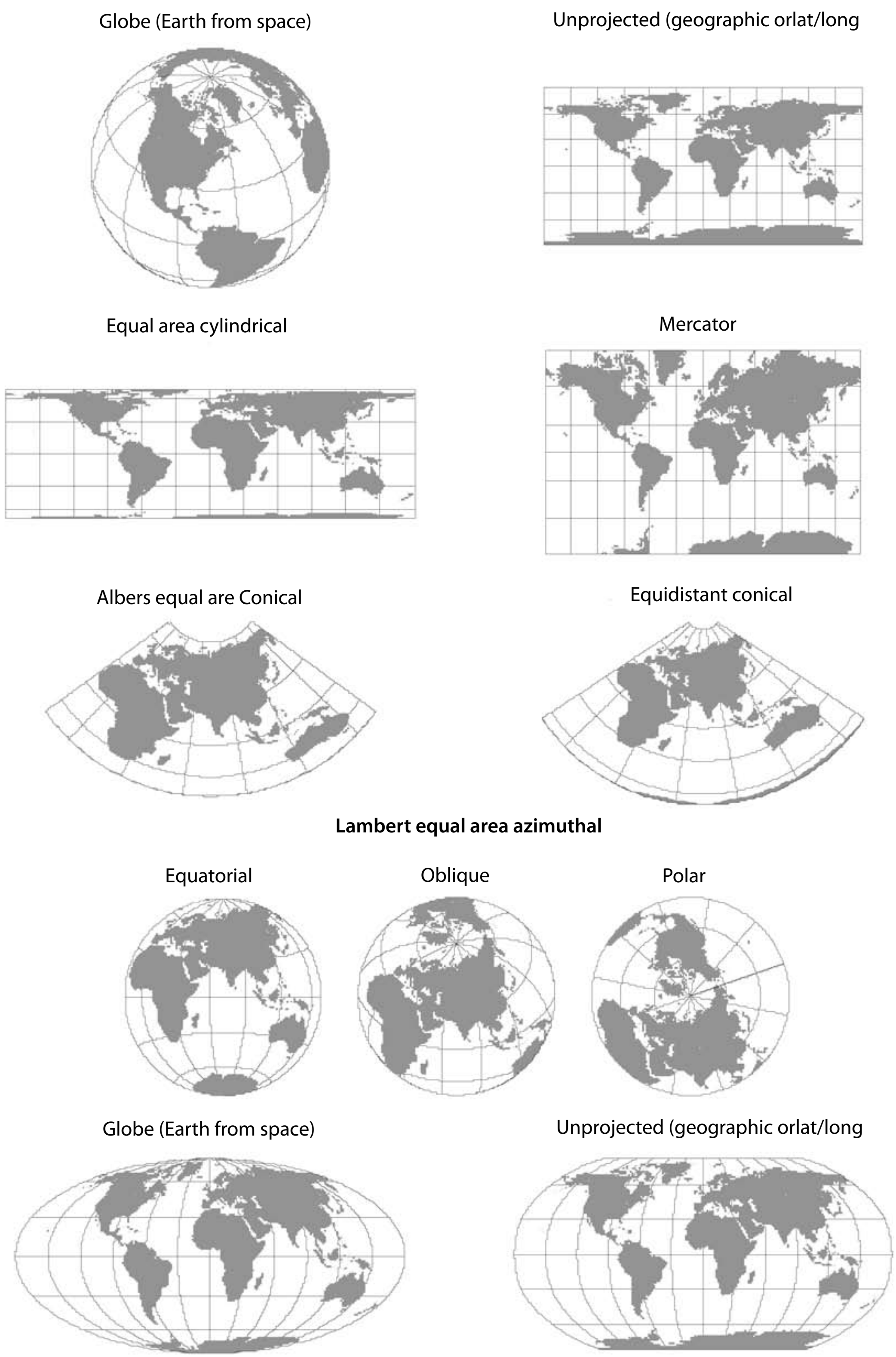
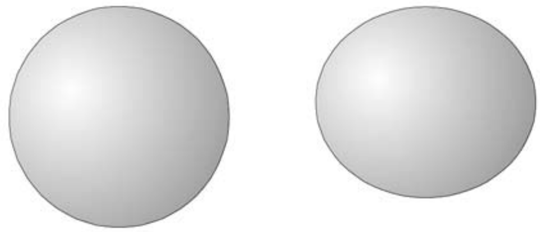


Figure A2.7
Sphere versus ellipsoid



Headquarters in New York in the UTM coordinate system, which will be discussed in more detail below. The latitude and longitude coordinates of United Nations Headquarters were projected into the same projection using different geodetic datums. The north-south shift between the older Clarke spheroids, which have been the standard in the United States until recently, and the newer World Geodetic System (WGS), is about 300 metres on the ground or more than 1 cm on a 1:25,000 scale map. Treating the Earth as a perfect sphere rather than as an ellipsoid would introduce an offset of more than 18 km.

Table A2.1
The projected coordinates of the United Nations Secretariat building in New York using different reference ellipsoids

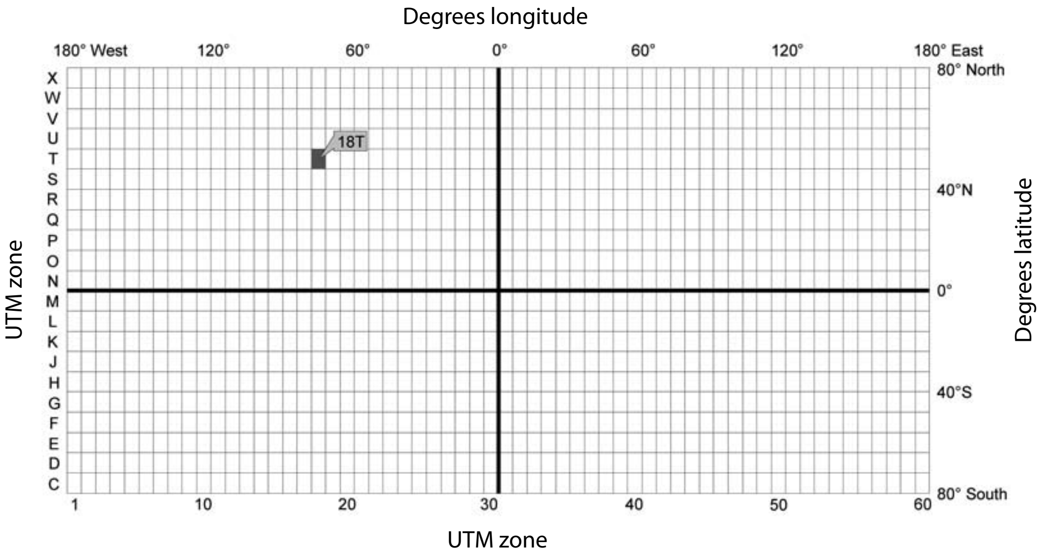
	UTM coordinates (metres)	
Reference ellipsoid	Easting (x)	Northing (y)
Clarke 1866	587 141.3	4 511 337.1
Clarke 1880	587 142.6	4 511 245.1
WGS84	587 139.0	4 511 549.7
Bessel	587 128.5	4 511 095.4
Sphere	586 917.2	4 529 920.6

The Universal Transversal Mercator (UTM) reference system

A2.19. One cartographic reference system that deserves more detailed discussion is the UTM system. It is one of the most common systems used for large-scale mapping around the world. It is based on a transverse cylindrical projection (Transverse Mercator), in which the cylinder touches the globe along a meridian. A different "local" meridian is chosen for different parts of the world. Distortions in scale, shape and distance along this tangent are very small. The global UTM system consists of sixty zones of longitude (see figure A2.8).

A2.20. Each zone has a width of six degrees longitude, three degrees in each direction from the tangent meridian. UTM zones are numbered sequentially from west to east, starting with 1 for the zone that covers 180°W to 174°W with central meridian 177°W. The zones are further divided into rows, with a height of 8 degrees (°). These are assigned letters from south to north, starting at 80° south with the letter C. Because distortion at the poles is very large, there are no UTM zones defined for regions beyond these limits.

Figure A2.8
The UTM system



A2.21. Coordinates are measured in metres (or feet) from the central meridian as eastings in the east-west direction and northings in the north-south direction. 500,000 is added to the easting to avoid negative numbers. For the same reason, 10,000,000 is added to the northing, but only for coordinates in the southern hemisphere. Such offsets are called "false easting" and "false northing".

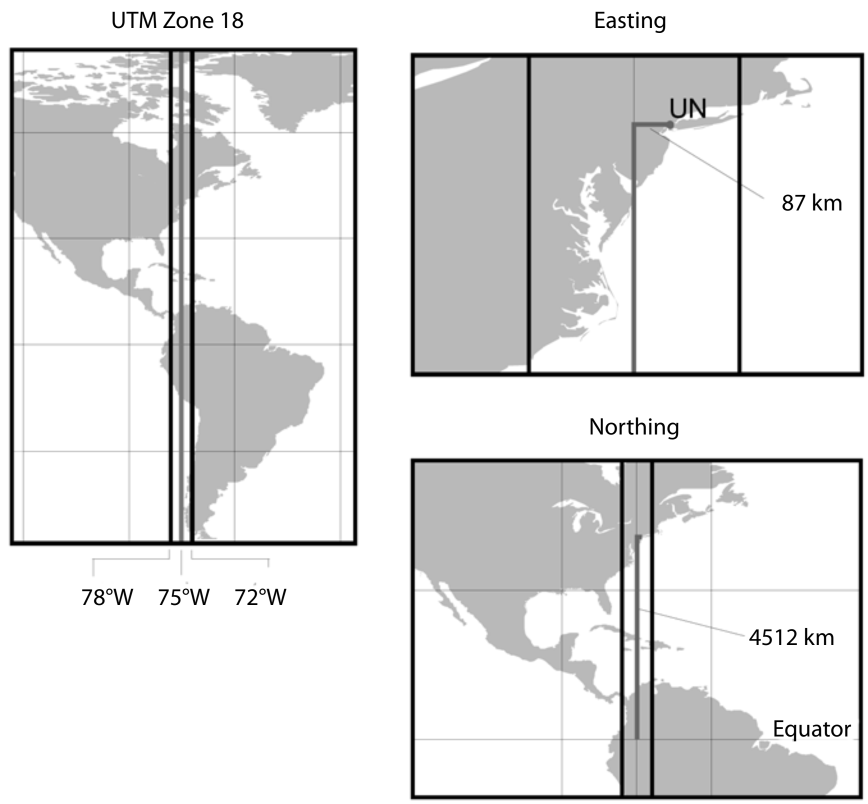
A2.22. To illustrate the use of the UTM system, an example is shown in figure A2.9. United Nations Headquarters in New York is located at 40°45'01" north latitude and 73°58'04" west longitude. This location falls into UTM zone 18T, which ranges from 72° to 78° west and from 40° to 48° north. The UTM x and y coordinates in metres are 587139.0, 4511549.7. This means that United Nations Headquarters is located approximately 87 km east of the central meridian of UTM Zone 18 (75°W) and about 4,512 km north of the equator.

E. Cartographic scale

A2.23. Published maps vary considerably in terms of the area on the ground that they cover. National or regional maps show only the most important features, while local maps show many details, such as individual houses or small creeks. The size or area that is covered on a standard map sheet or on a digital display is determined by the cartographic scale that is chosen to draw the map. This scale is represented by a fraction relating the distance on the map to the real-world distance on the ground. For example, on a 1:25,000 scale topographic map, 1 cm on the map represents 25,000 cm or 250 metres in the real world.

A2.24. Since map scale is a fraction or ratio, the larger the distance on the ground that is represented, the smaller the map scale. For instance, a 1:1,000,000 scale map is a "small scale" map since 1 divided by 1 million is a very small number (0.000001). A 1:5,000 scale map has a relatively "large scale" since 1 divided by 5,000 is a relatively larger number (0.0002). Thus, small-scale maps show large areas, while large-scale maps focus on small areas. In practice, the terms "small scale" and "large

Figure A2.9
The location of United Nations Headquarters in the UTM reference system



scale" are often confused because in colloquial use, "large" and "small" refers to the area covered or the size of the phenomena rather than to the fraction. Global climate models, for example, are often termed large-scale models. A useful convention to avoid misunderstanding is therefore to explicitly refer to "cartographic scale."

A2.25. The following are some common map scales:

Map scale	1 cm on the map represents	
1:5,000	50 m	larger scale
1:25,000	250 m	
1:50,000	500 m	
1:100,000	1 km	
1:500,000	5 km	
1:1,000,000	10 km	smaller scale

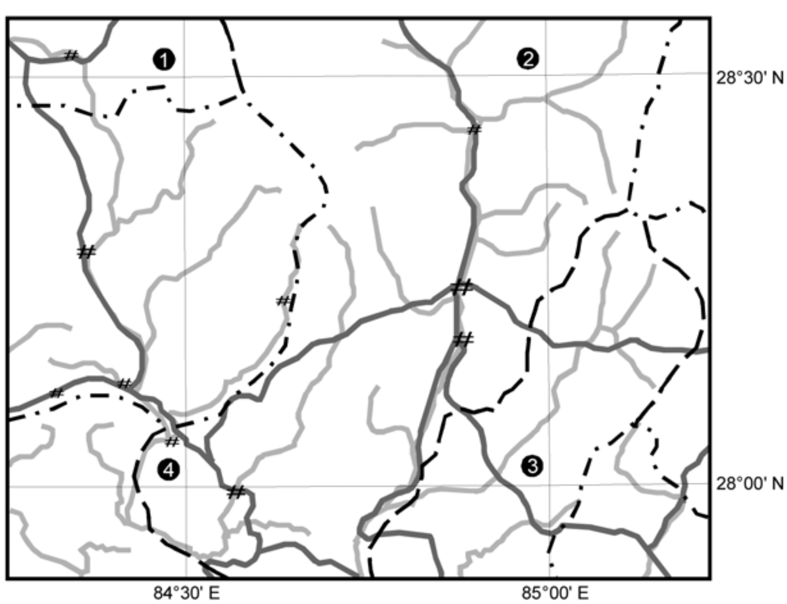
A2.26. Given the transition from analogue maps to digital geographic databases, it is very important to emphasize that digital geographic data are essentially scale-less. Once coordinates that define geographic features are entered into a GIS, they can be displayed at any specified scale. The user can zoom into and out of the map in exploring the data, thereby switching scales quickly and seamlessly. Nevertheless, it is important to keep in mind that the data were probably derived from source material (maps, images etc.) at a given source scale. Printed maps at different scales, for instance, will show varying degrees of detail. Individual buildings that make up a village will be shown on a 1:25,000 scale map. On a 1:500,000 scale map, the entire village will be displayed as a point if it is shown at all.

A2.27. The process by which map features are simplified or aggregated is called "generalization" and is an important component in map-making. Because of this generalization of features—meandering country roads become straight lines, details in district boundaries disappear—it makes little sense to print a map that was digitized from a 1:250,000 map sheet at a scale of 1:5,000 or to combine digital data sets that were derived from maps of very different scales. This shows that it is very important to indicate the source map scale in the documentation of a digital geographic data set. Also, owing to these map-scale issues, it is crucial in a large digital-mapping project to determine output scale requirements early on so that database development will be based on adequate source materials.

F. Georeferencing example

A2.28. The problem of georeferencing a map that has been digitized or scanned into proper map unit coordinates for storage in a GIS has been discussed in chapter II of the present *Handbook* in the section on digital map integration. To illustrate the process of georeferencing, the following paragraphs will describe a realistic example. Figure A2.10 shows a map that has been digitized into several layers. After digitizing, the coordinates are referenced in digitizing table units, in this case inches. In order to use the digitized map, together with other digital data for this geographic region, we need to convert the digitizer coordinates into the real-world coordinates that correspond to the map's original projection. Readers unfamiliar with coordinate systems and map projections may wish to review the material provided in paragraphs A2.--A2.7 above.

Figure A2.10
Control points on a map sheet



A2.29. The first step is to determine well defined control points. This is usually part of the digitizing process. Control points should be well distributed across the area of interest to improve estimation of the transformation parameters. That means they should not all be in one area or in the centre of the map. In addition to roads, rivers, administrative units and towns, the map also shows a regular grid of latitude and longitude lines spaced at half-degree intervals. The intersections of the graticule provide a good choice for the control points since their coordinates are easily determined. On the map, the four chosen control points are numbered from one to four. Their coordinate pairs are, respectively (84.5, 28.5), (85.0, 28.5), (85.0, 28.0) and (84.5, 28.0). Note that because GIS programs use planar coordinates, it is necessary to specify longitude/latitude (i.e., x/y) pairs rather than latitude/longitude pairs. For the same reason, we need to specify the coordinates in decimal degrees rather than in degrees, minutes and seconds as is common on paper maps or in gazetteers.

A2.30. Unfortunately, we cannot use the longitude/latitude coordinates directly for the transformation because the original paper map was not registered in geographic latitude/longitude coordinates; very few paper maps are, and often this is indicated by the fact that the latitude and longitude grid does not consist of straight lines. The map's original projection in this example is the Albers Conic Equal Area projection with the following parameters:

- Standard parallels: 27° and 30° north
- Central meridian: 84°
- Latitude of origin: 28°

A2.31. These map parameters are usually indicated on the map sheet. Before we can perform the coordinate transformation, we first need to convert the control point longitude/latitude coordinates into the correct real-world coordinates in the Albers projection. In most software programs, this can be done by listing the longitude/latitude pairs (since longitude is the x-, and latitude is the y-coordinate) in a text file or through a menu interface and specifying the relevant projection parameters in the system's projection change module.

A2.32. Of course, if we can read the real-world control point coordinates directly from the map, this additional step is unnecessary. This is possible, for example, on topographic maps referenced in the UTM projection. The same is true if the control points have been determined in the field using a GPS that automatically converts coordinates into a specific geographic projection.

A2.33. We now have the four control point coordinate pairs in the digitizer table coordinates, as well as in the real-world projection coordinates—in this case measured in metres. Both sets of coordinates are listed in table A2.2. The first control point, for example, is located about 49 km east of the central meridian (84°E) and 55.5 km north of the latitude of origin (28°N).

A2.34. The third step is the computation of the transformation parameters based on the two sets of coordinate pairs. Most GIS packages provide this option. Technically, the parameters are estimated using the following regression equations:

$$x' = a + bx + cy$$

$$y' = d + ex + fy$$

where x' and y' are the real-world coordinates, x and y are the digitizer coordinates of the control points and a, b, c, d, e, and f are the parameters to be estimated. The estimation errors in the transformation are the residuals of the regression.

A2.35. Table A2.2 shows for each control point the coordinate pair in the input coordinate system (digitizer units) and the output system (Albers projection in metres). In addition, the table shows the transformation errors (residuals) that the system has calculated in output units (metres). The transformation error may be seen to be about 7.8 metres in the x-direction and about 14.6 metres in the y-direction. These errors will rarely be zero. Sources of error include distortions in the paper maps due to shrinking and folding, as well as measurement error when digitizing the control point coordinates. A very large error in one or more of the control points usually indicates some significant mistake, such as switching of the x and y coordinates or control point identifiers. Overall, the process should be done with much care to ensure the accuracy and thus the usefulness of the resulting GIS database.

A2.36. The table also provides an indication of the overall error in the transformation. This is the root mean squared error, which is given in input and output coordinate units (inches and metres, respectively). A high RMS error indicates that the control point locations in the input and output map units do not correspond to the same relative locations. For a large-scale data conversion project, an acceptable maximum RMS error should be specified and maintained. What is considered acceptable depends on the map scale of the original paper maps and the accuracy requirements of the application. While census-mapping may not require a very large degree of accuracy, cadastral applications, for instance, have to conform to much higher standards.

Table A2.2
Transformation parameters

Control point	Coordinates in digitizing units (inches)		Coordinates in projected real-world units (metres)		Calculated errors in real-world units (metres)	
	x	y	x	y	x	y
1	11.777	19.660	48936.2	55529.6	-14.59	7.80
2	26.670	20.661	97871.5	55835.2	14.60	-7.81
3	27.696	3.824	98333.0	409.3	-14.55	7.78
4	12.751	2.810	49166.9	102.3	14.54	-7.77
RMS error (input, output)			0.005034,	16.524		

A2.37. The system will convert all coordinates in the map database into the output coordinate system in the same step. The output database is then properly referenced in the original paper map's coordinates. Subsequently, this map can be projected into a different cartographic projection, for example, to integrate it into a comprehensive database in a different standard projection. This description was intended to outline the general principles of transformation. Although the actual implementation is software specific, an understanding of the steps involved in georeferencing helps to appreciate the importance of this step.

G. Practical considerations

A2.38. Any large digital geographic database project (for example, in support of census operations) requires that map information from many different sources be integrated. For that reason, a standard projection and coordinate system need to be chosen. Ideally, the reference system that is chosen should match the system used in other mapping activities in the country. Most countries use a standard projection and coordinate system that is optimal for their national territory for the national map series at different scales.

A2.39. Almost all GIS packages provide functions for transforming coordinates from one reference system into another (e.g., from metres to feet or from digitizing units to map units), for converting digital maps from latitude/longitude into a map projection or to change between projections. They also allow the user to select a geodetic datum and any other relevant parameters. In some rare instances, a particular projection may not be supported and specialized projection software needs to be used. Global positioning systems (GPS), which are discussed in detail in chapter IV of the present *Handbook*, also support selected map projections and the most common geodetic datums. Coordinates collected during fieldwork can thus be captured as latitude and longitude pairs or in a projection system.

A2.40. Projection and datum information are usually included on topographic maps. A problem with digital cartographic data sets is that standard GIS formats do not necessarily store projection information explicitly. For example, a census agency may obtain a geographic data set of roads or hydrology without information about their map projection. If such data are combined with the digital census maps, they may not match perfectly. Vertical integration is thus impossible unless the two data sets are brought into the same projection system. If the map projection cannot be determined by tracking the lineage of the data set back to the source maps, the only option is to reconcile the two digital maps in an ad hoc manner, which may introduce significant errors. It is therefore important that all data sets be properly documented and that the metadata—information about the data—are kept with the digital map data set.

A2.41. A final practical consideration discussed here pertains to the conversion between different formats of storing latitude and longitude coordinates. These are usually expressed in degrees, minutes and seconds (DMS). The location of United Nations Headquarters in New York, for example, is 40°45′01″ north latitude and 73°58′04″ west longitude. To enter these latitude and longitude coordinates into a GIS or cartographic projection system, we need to convert the coordinates into decimal degrees first. Basically, this makes them look like normal Cartesian x and y coordinates. To convert degrees, minutes and seconds into decimal degrees we calculate, for example, the latitude and longitude of United Nations Headquarters as:

$$40+\frac{\left(45+\frac{1}{60}\right)}{60}=40.7502778$$

$$73+\frac{\left(58+\frac{4}{60}\right)}{60}=73.9677778$$

A2.42. Since the longitude of United Nations Headquarters is west of the Greenwich meridian, it is specified as a negative number in decimal degrees (i.e., -73.97). Similarly, latitude values in the southern hemisphere are also expressed as negative numbers.

A2.43. To convert, for example, the latitude back to degrees, minutes, seconds:

Degrees: 40

Minutes: $0.7502778\times60=45.016668=45$

Seconds: $0.016668\times60=1$

Annex III
Data modelling

A. Introduction

A3.1. This annex reviews geographic data-modelling issues and an example of the content of a detailed data dictionary that may be used by a census office to document the geographic databases that are produced for census purposes. A simpler data dictionary to accompany geographic census products disseminated to the public is continued in annex IV.

B. Definition of key terms

A3.2. A "spatial data model" is the description of geographical entities, such as houses, administrative units or rivers, and the relationships between those entities. In object-oriented data models, the definition usually also includes the operations that can be performed on the entities. A data model is independent from any specific software package. The user can therefore implement the data model in any comprehensive GIS package.

A3.3. The "spatial data structure" implements a specific data model. It consists of specific file structures that are used to represent different types of entities. For example, administrative units or water bodies would be represented as polygons—i.e., a series of coordinates where the first and last coordinate are the same. A data structure enables software operations that define the relationships between geographic entities. For example, a road may coincide with a part of a boundary of a polygon that defines an administrative unit.

A3.4. "Data format" is a more general term that is usually applied to a specific set of data structures within a software system. Some commercial data formats have been used so widely that they have become a de facto standard. DXF (Drawing eXchange Format) format, for example, was initially developed for the AutoCad software package. It is now supported by virtually all commercial GIS software packages.

A3.5. A "data dictionary" is a master document that describes the data model in detail as well as any codes used to identify the entities and their attributes.

A3.6. Finally, a "database schema" is a description of the logical relationships between spatial entities, attribute tables and integrity rules that define a complete and comprehensive spatial database.

C. Example template

A3.7. The following example template is adapted from the very comprehensive definition of geographic entity definitions in the *Canadian National Topographic Database—Data Dictionary* (Geomatics Canada, 1994).

Table A3.1
Information compiled to define a spatial data model

Entity name	The concise name of the geographic feature.
Definition	Detailed description of the geographic entity.
Fixed domain attributes	Attributes that can have only a limited number of pre-defined values. For example, the type of administrative unit (district, province, etc.) or the surface type of a road. These predetermined codes are the "domain" of possible values.
Variable domain attributes	Attributes that have a potentially infinite number of possible values. Their domain can therefore not be defined. Examples are the unique identifier of the administrative unit, the unit's population or the name of a river.
	Each attribute is described by the following information: Name Type *e.g., alphanumeric (A), integer (I) or real (R)* Number of characters or digits allowed The *domain* of values—i.e., a list of all possible values and their definitions—for fixed domain attributes, or the attribute *definition* for variable domain attributes.
Authorized combinations of attribute values	For fixed domain attributes, all allowable combinations of attributes are listed. For example, in the case of administrative units, only districts and provinces may have an official administrative capital. So, if the administrative unit type is not district or province, another attribute that lists the capital name should be empty. Information about authorized combinations of attribute values is useful for automated consistency checking. If the entity has no fixed domain, "none" is entered. If there is only one fixed domain attribute, all authorized values are listed. If there are several fixed domain attributes, all authorized combinations of values are listed.
Relations	A description of the relations that the geographic entity may have with other spatial features. This is useful, for example, to define how rivers or roads may coincide with administrative unit or enumeration area boundaries. Relations are defined by the following characteristics: — *entity name* and *geometry of entity*—e.g., point (P), line (L) or area (A) — *relation*—e.g., *connect*, for a line connecting to a point; or *share*, for an area sharing a border with a line. — *cardinality*, defined by a pair of values defining the minimum and maximum number of times an entity can be involved in a relation. For example, a road intersection is related to road features. The intersection must have at least one road connected to it, and can, in theory, be connected to an infinite number of roads. If the maximum number cannot be determined, it is represented by N. The relation of road intersection to road is therefore (1, N). — name and geometry of related entity *Note*: This definition refers only to relations between geographic features; relationships between fields in the geographic attribute table and external tables need to be defined separately.
Geometric representation and minimum size (metres)	The geometric feature used to represent the entity. For administrative units, this will almost always be polygons. However, for other features the geometric representation of a spatial entity may depend on the cartographic scale. For example, a village may be represented as an area representing its perimeter at large cartographic scales (e.g., 1:25,000), while it is shown as a point at small cartographic scales (e.g., 1:250,000). At the same map scale, a larger village or town may be represented as an area but a small village as a point. Depending on the type of feature, the minimum size of entities can refer to their "surface area", "width", "length" or "height".
Notes	Any additional information required to define the entity, as well as footnotes pertaining to any of the other descriptive fields.
Diagram	To illustrate the way in which an entity is modeled, a graphic illustrates the relations of that entity with various other entities.

A3.8. The most important information in the database template is the definition of each entity and the detailed description of all attributes stored for the geographic features. For many census-mapping projects, these basic database descriptors may be sufficient. However, especially if the census database is to be incorporated in a national geographic database, it is advisable to spend more time and effort in developing a database design that ensures compatibility with information from other agencies. In this case, the relationships between administrative or census units and other geographic features should be clearly defined.

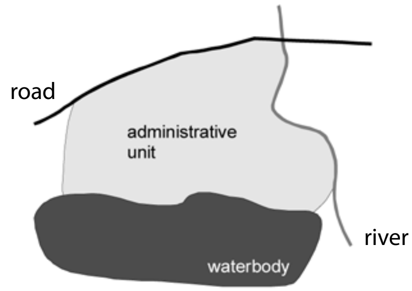
A3.9. To clarify the contents of a data dictionary, table A3.2 gives an example that describes a definition of an administrative unit data layer. This example is for illustration only. The exact specification will vary depending on the implementation in each country. It should be noted that "fixed domain attributes" are attributes that are discrete and not continuous.

Table A3.2
Example: administrative units for a country with three subnational levels

Administrative unit	
A geographic area with legally defined boundaries created for the purposes of implementing administrative and other government functions.	
Fixed domain attributes	
Administrative unit type I (1):	
1—Province	a first-level administrative unit
2—District	a second subnational-level administrative unit
3—Locality	a third-level administrative unit
Rural/urban indicator I (1):	
1—Not applicable	only localities are classified as rural or urban
2—Rural	an administrative unit consisting of a town or city
3—Urban	an administrative unit with predominantly rural characteristics
Variable domain attributes	
Administrative unit identifier I (14)	
Note: in this example database, all attribute information (e.g., name, alternative name, number of households, population etc.) is stored in separate data tables that are linked to the geographic attribute table through the administrative unit identifier.	
Authorized combination of attribute values:	
Province (not applicable)	
District (not applicable)	
Locality (urban)	*Note*: only these combinations are possible; for example, there are no urban provinces or rural districts.
Locality (rural)	
Relations	
Administrative unit (P) Share (0,N) Road (L)	
Administrative unit (P) Share (0,N) River (L)	
Administrative unit (P) Share (0,N) Water body (P)	
Note: Roads and rivers are represented as lines (L) and may coincide with parts of the boundary of an administrative unit which is represented as a polygon (P). Similarly, an administrative border may coincide with the shore of a water body, such as a lake, which is represented as a polygon. (0,N) refers to the cardinality of the relationship. It means that, for example, at a minimum zero roads may coincide with an administrative unit boundary and the maximum is indeterminable (indicated by N, meaning any number).	
Geometric representation and minimum size	
The administrative unit is represented as a point feature if its surface area is less than or equal to 1 km^2 and as an area feature if it is larger than 1 km^2.	

Note: Administrative units must coincide with enumeration area boundaries. Administrative units must cover the national territory exhaustively. In other words, there cannot be any part of the territory of the country that is not assigned to an administrative unit.

Diagram



Annex IV
Example of a data dictionary for distribution

A4.1. This is a sample data dictionary for the distribution of a census geographic database of localities for the hypothetical country of Poplandia. The term data dictionary is sometimes used interchangeably with the term metadata, although in practice the two are different. Data dictionaries predate metadata, and usually the term "data dictionary" refers to the information that is distributed with data that an agency disseminates. ISO metadata standards provide specific guidelines. Each country can use "profiling" to adapt a common metadata format for its own national uses. The example below is meant for illustration only. The actual content of the data dictionary should be carefully designed by the national census office to consider specific issues relevant to the country.

Data dictionary: census GIS database of localities

Database title	Digital geographic census database of localities in Poplandia.
Source	National Statistical Office (NSO), Census Branch, Cartography Section (1996), National Population and Housing Census of Poplandia 1995.
Database content	The database consists of a geographic data layer of localities for the entire country. The GIS database is distributed in Arc shape file format (Environmental Systems Research Institute, Inc.), MapInfo Interchange format (MapInfo, Inc.) or as a plain text file of coordinates. This documentation refers to the Arc Shapefile version. The geographic attribute data table of the locality GIS layer (LOC.DBF) contains basic information only, including the locality code (LOC_CODE), and the names of the administrative units into which it falls. Two external data tables are distributed with the GIS database, one containing population characteristics from the census (POP.DBF) and one for household attributes (HH.DBF). These data tables can be linked to the locality GIS database by using the common field LOC_CODE. Unless otherwise indicated, all data refer to the census date on 1 July 2005.
Administrative and reporting units	The database contains information for 1,291 localities in 9 provinces and 123 districts.
Software and hardware requirements	The database can be viewed with any GIS or desktop-mapping package that can import Arc shape files or MapInfo Interchange format files. Minimum systems configuration depends on the software used to access the data. Generally, a 486Mhz or faster IBM-compatible personal computer with at least 8 mb of RAM will be sufficient. The database can be accessed from the CD-ROM or installed on the computer's hard-drive. It will require 16 mb of hard disk space.
Database distribution format	The database is distributed in uncompressed form on the CD-ROM and can be accessed directly.
Projection	Equidistant Conic
Standard parallels	20° north and 60° north

Database title	Digital geographic census database of localities in Poplandia.
Central meridian	140° west
Coordinate units	Metres
Coordinate offset	None
Source map scale	Varies. Most urban localities were delineated on 1:25,000 and larger-scale maps; rural localities were delineated on 1:50,000 and smaller-scale maps.
General accuracy information	According to national mapping agency information, the estimated average coordinate accuracy is +/-100 metres in rural areas and +/-30 metres in urban areas.
Disjoint reporting units	Some of the localities consist of more than one polygon. The attribute table contains a field (FLAG) which has a value of 1 for the major polygon (the only one for localities consisting of only one polygon) and zero for any minor polygons. To avoid double-counting when census data are aggregated, the aggregation should be done only after selecting those localities with a FLAG value of 1.
Related products	The NSO has published similar digital GIS databases for enumeration areas. Because the number of enumeration areas is very large, there are separate GIS databases for each province. Please contact the National Statistical Office for more information.
References	National Statistical Office (2005), *Technical Report on the Census-Mapping Activities of the National Population and Housing Census of Poplandia 2005, Census Branch, Cartography Section* National Statistical Office (2005), *Methodological and Administrative Report for the National Population and Housing Census of Poplandia 2005, Census Branch* National Statistical Office (1996), *Results of the National Population and Housing Census of Poplandia 2005, Census Branch, Cartography Section*
Contact information	National Statistical Office, Census Branch Cartography Section, User Services P.O. Box 9999 Tarota, Sambas Province Tel: 99-99-99999 Fax: 99-99-99998 E-mail: geog@census.gov.xx Website: www.census.gov.xx

Geographic data files

LOC.SHP—GIS database of locality boundaries

File name:	LOC.SHP	
File type:	ESRI Arc Shape File	
Feature types:	Polygons	
Associated files:	LOC.DBF	Polygon attribute table (part of the shape file)
	POP.DBF	Census population indicators
	HH.DBF	Census household indicators
	LOC.SHX	Internal geographic indexing file used by ArcView

Attribute data files

LOC.DBF: locality characteristics

Field name	Description	Field definition	Range	Codes	Missing values
LOC_CODE	Official locality code. Provides the link to the external data tables pop.dbf and hh.dbf. The geocode is constructed by concatenating administrative identifiers: 2-digit province + 3-digit district + 3-digit locality	Int, 8	Positive value	None	-999

Field name	Description	Field definition	Range	Codes	Missing values
AREA	Area of the locality in km²	Real, 6.1	Positive value	None	-999
FLAG	Indicates whether the polygon is the major one for the locality. For localities that consist of two or more polygons, only the biggest or most important will have a value of 1	Int, 1	0-1	0-minor 1-major	
URBAN	Indicator whether the locality is classified as urban or rural.	Int, 1	0-1	0-rural 1-urban	-1
LOC_NAME	Name of locality	Char, 25	None	None	"n. a."
DIST_NAME	Name of district	Char, 25	None	None	"n. a."
PROV_NAME	Name of province	Char, 25	None	None	"n. a."
AREA_TOTAL	Total area of locality in km²	Real, 10.3	Positive value	None	-999
AREA_LAND	Area of locality covered by land in km²	Real, 10.3	Positive value	None	-999
AREA_WATER	Area of locality covered by waterbodies in km²	Real, 10.3	Positive value	None	-999

POP.DBF—census population indicators

Field name	Description	Field definition	Range	Codes	Missing values
LOC_CODE	Official locality code. Provides the link to the GIS attribute data tables loc.dbf and hh.dbf.	Int, 8	Positive value	None	-999
POP_TOT	Total enumerated population	Int, 7	Positive value	None	-999
POP_DENS	Population density in persons per km² (POP_TOTAL / AREA)	Real, 5.1	Positive value	None	-999
…	…	…	…	…	…

HH.DBF—census household indicators

Field name	Description	Field definition	Range	Codes	Missing values
LOC_CODE	Official locality code.	Int, 8	Positive value	None	-999
HH_NUM	Number of households.	Int, 7	Positive value	None	-999
HH_HEAD	Sex of head of household.	Int, 1	0-1	0-male 1-female	-1
…	…	…	…	…	…

Annex V
Thematic map design

A. Introduction

A5.1. The present annex presents a brief overview of design considerations for making thematic maps. This overview cannot cover all the issues surrounding information content of maps, however; if necessary, a textbook should be consulted. Cartographers distinguish between several types of maps. General purpose maps serve as a reference frame for orientation. They show mostly real geographic features that can be observed on the ground. These features are either natural—rivers, mountains, coastlines—or man-made, such as roads or settlements. Reference maps also show features that are not visible on the ground. The best example are political boundaries and the reference grid showing latitudes and longitudes. Topographic maps fall into this category of general-purpose or reference maps. They play an important role in the mapping of enumeration areas as they provide information about features that an enumerator uses for orientation in the assigned work area.

A5.2. More relevant to the mapping of census results are "thematic maps". These display the geographic distribution of physical or cultural phenomena that cannot easily be observed directly on the ground. Thematic maps can be based on qualitative or quantitative information. An example of the former is a map showing the distribution of people by mother tongue or religion. Quantitative thematic maps, sometimes called statistical maps, by contrast, provide some information about the relative size of the features that are mapped. An example is a map in which the symbols representing the cities in a country are scaled according to the size of each city. Another example is a map in which reporting areas, such as districts, are shaded according to their population density. Most of the maps produced for a census atlas will be of this nature.

B. Map design principles

A5.3. Despite their frequent use in analysis, maps are not good at showing exact data values. On a map, data values are translated into symbols. The cartographer must assign data values into class intervals to obtain a manageable number of categories that are represented as colours or symbols. This means that some information is lost in the map display. While maps are strong at showing trends, relative magnitudes and distribution of indicator values, data tables or digital maps whose database can be queried are more appropriate if the exact values are of interest.

A5.4. The production of presentation maps is a design process in which the cartographer communicates an idea or concept to the reader (Monmonier, 1993). This is similar to other forms of communication of qualitative or quantitative information in graphical form, using charts, pictures or other visuals. The same design principles that guide graphic design therefore also apply to cartography.

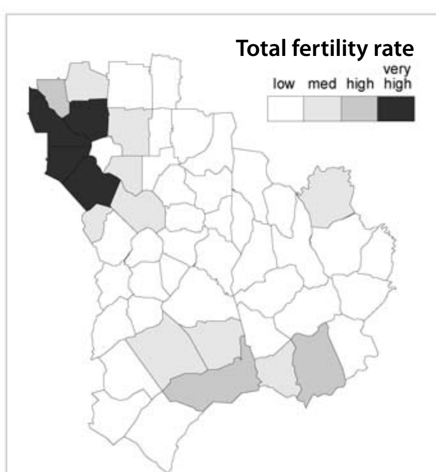
A5.5. The most important design principle is simplicity and clarity. Many maps end up being cluttered because the cartographer tried to present too many things in a small space. A useful concept is Tufte's (1983) maximization of the "data to ink ratio": adapted to map-making, this means that most of the ink used should be devoted to representing geographic data rather than to drawing extraneous information. Superfluous information should thus be left out. Titles beginning "Map of …" or "Legend" are unnecessary, as are many boxes, neat-lines and often, though not always, north arrows and scale bars. Like most principles, this one also has its exceptions. Some map elements, such as the legend itself, a concise title and source information, are clearly required for the understanding of the map.

A5.6. Simplicity also implies that no space should be wasted. With high-resolution laser printers available almost anywhere, maps do not have to be printed in very large format to show all details. The better the design of the map, the smaller it can be printed. Using space parsimoniously also means that oversized fonts, legend symbols or insets should be avoided.

A5.7. Achieving "visual hierarchy" is another important concept. It applies to the elements within the map itself, as well as to the arrangement of all components of a map. On the map itself, the choice of colours or symbols reflects the ordering of data values. In a map of child mortality, for example, the reporting units with the highest values could be shaded in the strongest colour or darkest gray shade. These are the hotspots that should catch the attention of the viewer immediately. In figure A5.1, for instance, the low to high classes are deliberately shaded in light gray tones to highlight the "very high" category. It is the contrast between the dark colours and the surrounding subtle tones that creates visual hierarchy. A relatively light area surrounded by dark tones would stand out just as much. Colour choice will be discussed in more detail below.

A5.8. The cartographer can also use other techniques to guide the attention of the viewer towards a particular area on the map. A crisp boundary around the most important features on the map, for instance, makes them stand out from the background. Annotation or arrows pointing at specific features are also sometimes used but often clutter the map.

Figure A5.1
Establishing visual hierarchy through choice of colours or gray shades



A5.9. For the overall map composition, the same principles apply. The most important part of a map is the cartographic information itself, the title and the legend explaining symbolization. These should be the most prominent features on the map page. Any other map elements should be added with caution.

A5.10. The subject of colour in statistical thematic maps is a multifaceted one (see also para. A5.4 above). Cartographers should be aware of the trade-offs between visual presentation using colour and the extra cost in ink and other supplies. One issue to consider is colour blindness. Maps should avoid the use of the colours of red and green for data values, because colour-blind people often cannot distinguish them. More generally, a colour map should be reproducible on a black-and-white-printer with no loss of information content. This means respecting the shading density and substituting pattern fills, when appropriate. Finally, cartographers should take care not to offend any part of the population by their design choices. Cartographers need to be aware of the sensitivities of different regions or population groups. Some symbols or colours may have certain negative or positive connotations to different ethnic or racial groups in the country. Map design should avoid using symbols that are associated with stereotypes concerning any population subgroup. Above all, cartographers should be aware of the overall need to use colour judiciously and not for gratuitous effect.

1. Elements of a thematic map

A5.11. A thematic map is made up of several components. The map itself consists of a base map that shows the boundaries of the area of interest, such as the country's borders, and possibly some reference features, such as major rivers or cities. These provide orientation for the reader who wishes to compare the magnitude of a variable in one part of the country with that in another. The second main element is the thematic map overlay that presents the geographical distribution of the variable.

A5.12. In addition to the actual map information, a publication-quality map contains additional elements. In particular:

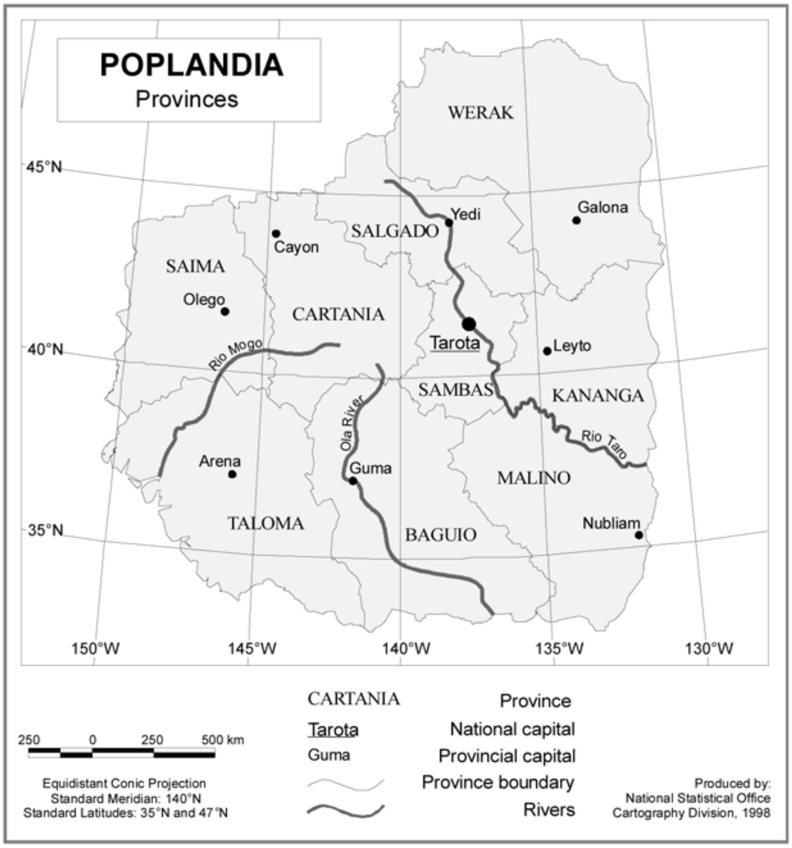
- **Titles and subtitles** should be short and descriptive.
- **Data source, credits and production date** give the user information about the reliability and credibility of the map. Copyright information should also be included. Some agencies that regularly produce maps also add reference and version numbers for internal use. Any other explanatory information relevant to the understanding of the map's content should be added. For maps printed in large format, the cartographic projection parameters should also be indicated.
- A **map legend** describes how the values of the mapped variable were translated into map symbols, for instance which colours are used to map a given range of population density values. It is important to always include the units of measurement in the legend, such as "persons per km^2".
- A **map scale** allows the user to measure distances on the map. For a series of thematic maps, such as a census atlas, where all maps are drawn at the same scale, this information does not have to appear on every page. The same is true for relatively small maps of well-known areas, where it is unlikely that the reader would wish to perform distance measurements. Adding a scale bar is usually better than specifying the scale numerically (e.g., 1:1,000,000). If the map is reduced or enlarged during photocopying, the scale bar will still be applicable. The nominal map scale used to draw the original map, by contrast, will then be incorrect.

- A **north arrow** is not absolutely necessary on a reference map as long as all maps are oriented towards the north. This is especially true if the map shows a well recognized geographic area, such as the entire country. If maps are rotated to obtain a better fit on the page after rotating, a north arrow must always be included.
- **Map borders and neat-lines** serve to separate different elements of the map; the use of such graphic elements is largely a design issue. Too many lines and boxes make the map look cluttered. Additional borders should only be used if the map elements are not well separated.
- **Place names and labels** that support the identification of geographic features or statistical areas.
- The **graticule** is the grid of latitudes and longitudes (parallels and meridians) that facilitate orientation on the map. These should be included on small-scale maps.
- **Locator maps** are used to show the location of the area covered by the main map. For example, a district level map of population density could be accompanied by a small map showing the location of that district in the country or province.
- **Inset maps** are similar to locator maps. But instead of showing the location of the area covered by the main map, they show some small part of the map at a larger cartographic scale. For example, a province level map might be accompanied by a small inset map showing the capital area or the information in a small district in more detail.
- **Text and annotation** give background information or explanations that should be short and to the point.
- **Additional graphic elements** could include a histogram showing the statistical distribution of the variable or the logo of the office that produced the map.

A5.13. Figures A5.2 and A5.3 show two examples of maps that incorporate many of the thematic map elements. Figure A5.2 is a map of first-level administrative units in the hypothetical country of Poplandia. Draped over the map is a grid of latitudes and longitudes that provides the geographic reference. The national capital, provincial administrative capitals and major rivers are added for reference. All features are labelled properly, using different fonts for different types of features. The margin below the map area shows a scale bar, the legend describing the types of thematic features shown and the source of the map. If the statistical office has a logo, this could be added to each map as well. A north arrow has been omitted for two reasons. One is that the map does not have an unusual orientation and the longitude lines make it quite clear that north is at the top of the map. The other, less obvious reason is that the cartographic projection used for the map has the longitudes converging towards the north. This implies that north is in a slightly different direction at different longitudes.

A5.14. The thematic map in figure A5.3 shows population density in one of the provinces of Poplandia. A map of this kind could, for example, accompany tables that show population characteristics by province in a census publication. Such a map could also include charts or other graphical representations. The design of the map itself is kept fairly simple. The title describes the theme of the map and the subtitle shows the geographic area. Rather than the standard legend that shows the colours in equal sized boxes, the legend in this map shows the population density categories in the form of a histogram. This serves the purposes of a traditional legend—relating values to shade

Figure A5.2
A sample map of administrative units and major urban centres



colours—and in addition presents the frequency distribution of the district values. For more complex maps, consisting of more areas, one could add the actual number of districts falling into each category. In order to keep the map clear and simple, that was not done in this case. Below the legend and data source, a small locator map shows the location of the province of Cartania in the country. It is often not necessary to add labels to a locator map that shows the country since the country's shape is usually recognizable to readers.

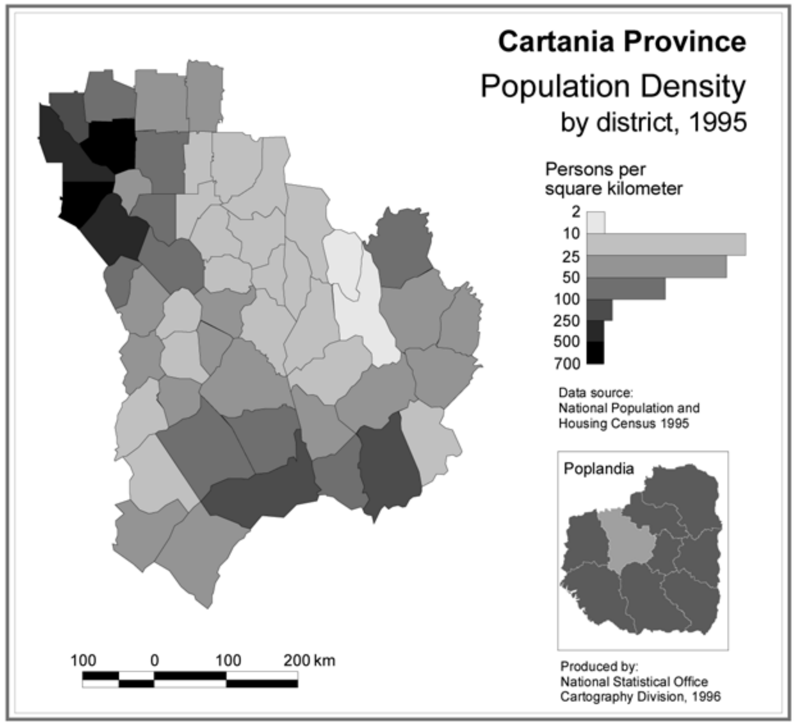
2. Measurement levels and graphic variables

a) Spatial dimensions

A5.15. Thematic maps not only show the location of a feature but also provide some information about that feature—the value of a variable at each geographic location. A thematic map is thus composed of the geographic elements and some attribute for these elements. That means in designing a thematic map, the spatial dimension of the geographic features must be considered, as well as the level of measurement of each variable. Both will determine cartographic options that are available for producing a map that is visually appealing, easy to interpret and accurate.

A5.16. Geographic features are represented in a geographic database by geometric primitives: points, lines and areas. Further categories, though less often used in cartography, add a third and fourth dimension: volume and space-time. Which geo-

Figure A5.3
Example of a thematic map of population density



metric shape is used for a real-world feature depends sometimes on the spatial scale of the map or data set. For instance, a village or town can be represented as an area in large-scale maps but will be shown as a point in maps with a smaller cartographic scale at the province or country level (see figure A5.4). A road might be shown as a line on a province map, but as an double line—i.e., an area feature—on a city map.

A5.17. It is important to keep in mind that boundaries and locations are not always as clearly defined as they appear in the discrete representation of a map or geographic database. Complex real-world features often need to be generalized, simplified or abstracted to be represented in a computer database. For instance, many real world features do not have clear boundaries. There is often a transition zone between forest and non-forest. If the forest is represented as an area feature (rather than as points for each tree), there will necessarily be some loss of information (see figure A5.5).

Figure A5.4
Effect of generalization on the display of spatial features

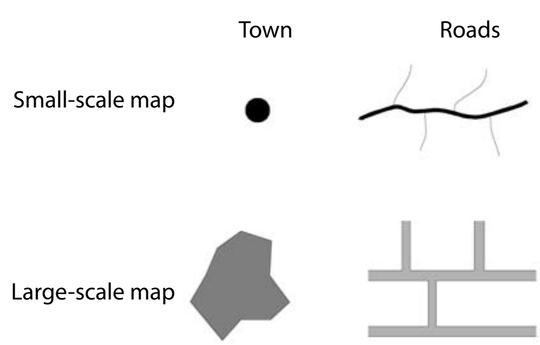
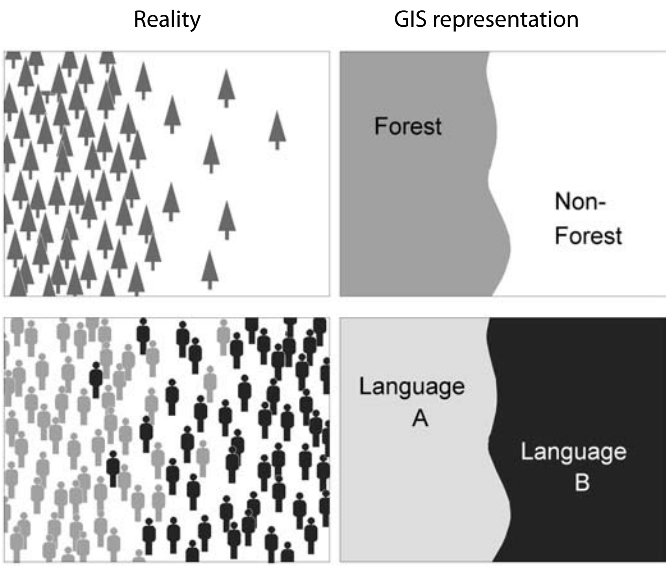


Figure A5.5
Real-world complexity sometimes needs to be simplified for GIS representation

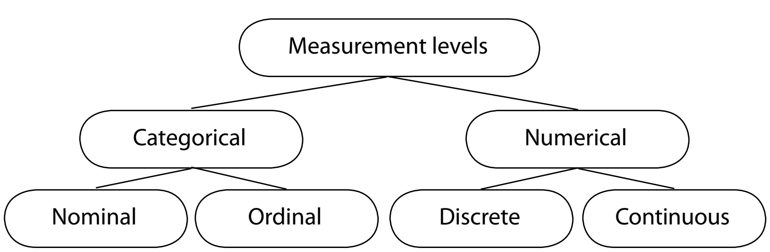


A5.18. One example in the socioeconomic realm where ill-defined boundaries are encountered is the distribution of ethnic or language groups. Despite the sometimes very distinct distribution patterns of such groups, there are probably areas at the outskirts of each region where people of different ethnic or language groups live interspersed. Cartographers sometimes use dashed lines to represent such ill-defined boundaries, although this does not resolve the issue of where to place the boundary on the map.

(*b*) Measurement levels

A5.19. Equally important is how the variable we want to map is measured. The main distinction is between categorical and numerical information (see figure A5.6). Categorical data, in turn, can be classified as nominal or ordinal. Nominal or qualitative data simply describe a type of feature, but there is no natural ordering between the categories. An example are types of houses, such as stone or wood-frame houses. Ordinal data, on the other hand, implies a ranking between the categories, although we do not know the interval between the categories. For instance, based on survey responses we might classify households as having a low, medium or high level of well-being. We do not know, however, whether the difference between low and medium is the same as that between medium and high.

Figure A5.6
Measurement of variables



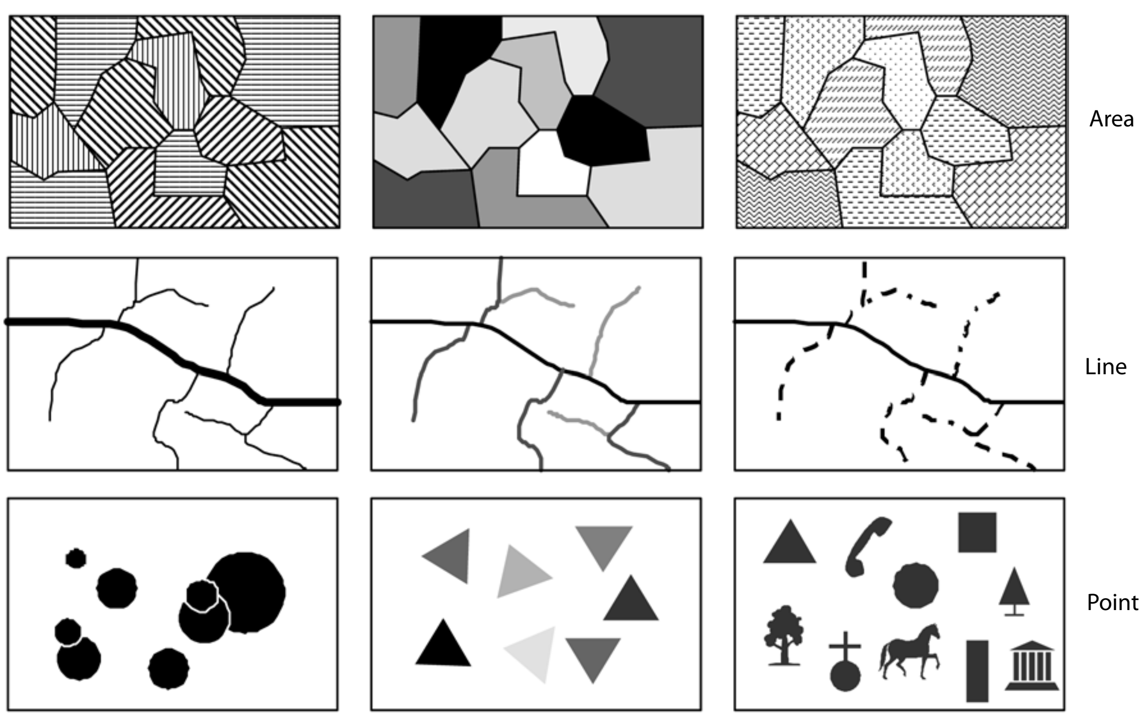
A5.20. If we can quantify the difference between the categories, we have numerical data. Discrete data are counts, such as the number of bedrooms in each household or total population. Continuous or ratio variables can take on any desired value. They can thus be measured with high precision. For census data, continuous variables are usually indicators that are calculated for aggregate census units, such as population density, the proportion of the population with access to safe drinking water or the total fertility rate.

(*c*) Graphic variables

A5.21. On a thematic map, graphic symbols reveal the differences in values or categories among geographic features to the viewer. The concepts of symbolization which are applied in cartography are similar to those developed for graphic design applications by Bertin (1983; see also MacEachren, 1995). He distinguishes between the following graphic variables;

- **Size** is an indicator of ordinal or numeric differences. Size is most important for point or line features, such as to show the size of towns and cities, using graduated circles, or the magnitude of migration between regions, using lines or arrows of varying thickness.
- **Orientation** is used, for example, in cross-hatching of area features. Also, geometric point features can be shown in varying orientation. Orientation does not imply any differences in the magnitude of a variable and are therefore useful for showing nominal data.
- **Texture** refers to the density of a constant pattern that varies between areas. It can be used to represent ordinal or numeric differences. This is a useful

Figure A5.7
Graphic variables for polygons, lines and points



short cut if output devices have limited capabilities in printing colours or gray-shades. Texture is also very useful in showing layered information, in which two variables are drawn on top of each other. However, it is not easy to preserve clarity in such maps and they are therefore most useful for exploratory analysis applications.

- **Shape** is most important for point features. Symbol and font sets in commercial GIS and desktop-mapping packages provide a large number of distinct symbols. Best known in cartography are symbols that represent public buildings, such as places of worship or hospitals.
- **Colour** is well-suited to show numeric and to some extent ordinal differences. Colour choice is one of the most important issues in cartographic design, and this topic is discussed in more detail below.

A5.22. In principle, each of these dimensions is applicable to every geographic feature type—i.e., points, lines and areas. But in most instances, only a subset of graphical variables is used for different feature types. Some examples are given in figure A5.7. The graphic variables for a thematic map are chosen to match the type of measurement in the indicator that is mapped. For instance, size and colour are most important for representing numeric values. Shapes of point symbols or texture of area features represent different nominal values.

3. Types of thematic maps

(*a*) Mapping discrete features

A5.23. Census data compiled for public release consist of numbers aggregated for a reporting unit, such as a district or enumeration area. Such data are best represented cartographically, using choropleth maps. The term "choropleth" is derived from the Greek words for "choros" (place) and "pleth" (value). Choropleth maps show data for discrete reporting units, which are often established independently from the actual spatial distribution of the data (e.g., administrative boundaries). The symbol—i.e., colour or pattern—used to shade each reporting unit is determined by the value. Choropleth maps are different from so-called "area-class maps" where the reporting units are determined by the data. For example, on a map showing forest cover, the reporting units will be determined by the boundary between forest and non-forest areas.

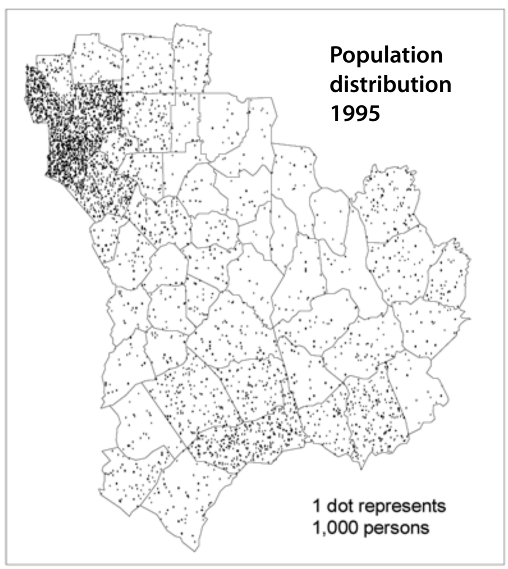
A5.24. An example of a choropleth map has already been shown in figure A5.3. Choropleth maps are constructed by first dividing the complete range of data values for the reporting units into a set of categories. Each category is then assigned a colour or shade pattern. Since enumeration data have a natural ordering, there is usually some logic in the choice of colours or tones, such as from light to dark colour shades or from coarse to dense patterns. The goal is to give the user an intuitive sense of the magnitude of the value in each reporting unit. There are many different ways of determining the symbols used to shade choropleth maps. The choice depends on the type of the variable, the range of data values and also on the output medium that is used to present the map. The choice of symbols is very important and this topic is therefore discussed in detail below.

A5.25. Choropleth maps are good at showing the overall distribution of data values on a map and for comparing the distributions across different maps. It is usually not possible to obtain the exact value for each reporting unit since the colours or shades only represent ranges of similar values. Such exact information is better presented in data tables or obtained using interactive query in a GIS.

A5.26. The values used for producing choropleth maps are almost always ratios, proportions or densities. These can be geographic ratios, where a data value such as population is divided by area to compute population density. Or they can be general ratios, where the denominator is a value other than area, such as the crude birth rate as the number of births per 1,000 persons. Most often when we map socioeconomic variables, the size of reporting units is not constant. For example, districts or provinces often vary drastically in size and population. If we would map a count variable, such as total population, rather than a ratio, the largest districts would likely be shaded in the darkest colours even if their population is small in relation to their area. Choropleth mapping is therefore not suitable for mapping absolute values.

A5.27. An alternative method for displaying count data are "dot maps". Dot maps were first used in France in 1830 to map the country's population distribution. On dot maps, a point symbol is used to represent one or more units of a variable that is mapped. For example, each dot might represent 1,000 people or households. The magnitude of the variable is then represented by the varying density of dots in the reporting unit. Figure A5.8 presents a typical dot map, showing population distribution.

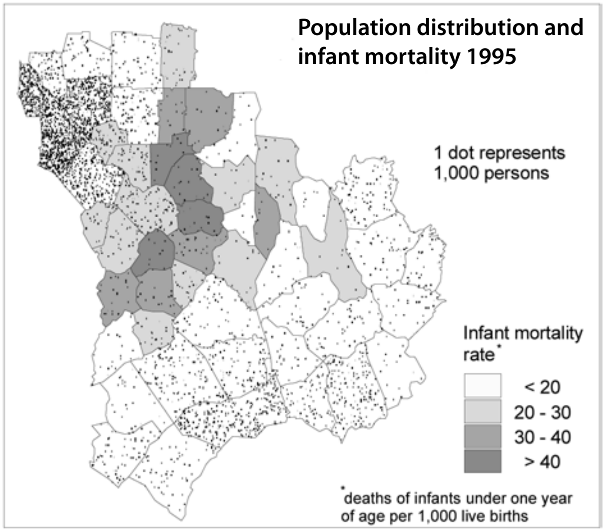
Figure A5.8
Dot-density map



A5.28. For the placement of the dots, two approaches are possible. The cartographer may select the location of dots based on knowledge of actual population distribution within each district. For instance, more dots would be located in and around urban areas than in less populated rural regions. In some applications, land-use or land-cover maps have been used to assist in the determination of dot densities within each reporting unit. Also, virtual masks may be employed so that no dots are placed in areas that are known to be uninhabited, such as water bodies, very dense forests or protected nature reserves.

A5.29. The alternative is to place the dots randomly within each district. In this case, dot density reflects overall value density. GIS and desktop-mapping packages that provide dot-density mapping functions usually use random placement of dots. The user has control only of the size of each dot and the symbol used for the dots. This symbol could be chosen to reflect the variable mapped, although a simple dot usually provides the clearest display.

Figure A5.9
Combination of dot-density and choropleth maps



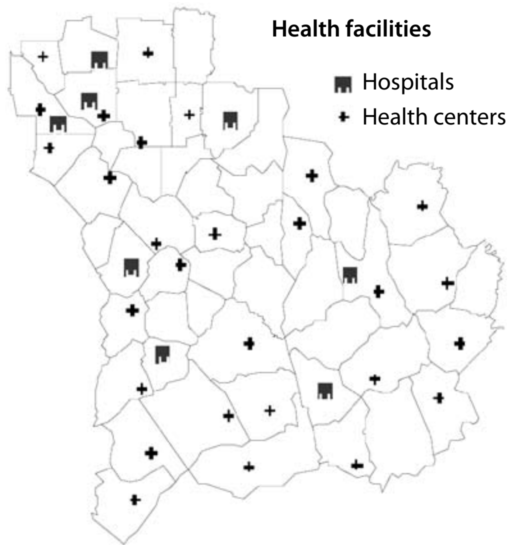
A5.30. Some specialized programs exist that allow dot placement assisted by some other data layers. Manual placement of dots to incorporate the cartographer's knowledge of variable distribution is, of course, very tedious.

A5.31. Dot maps are an effective way of representing density information provided that dot placement is guided by actual geographic distribution of the mapped variable, or that the distribution within each reporting unit is largely homogeneous. A great advantage of the method is that dot-density maps reproduce very well when photocopied or printed since they are essentially monochrome (black-and-white) maps. Dot-density maps can also be used in combination with choropleth maps to show two variables at the same time—for example, the map in figure A5.9 shows that there is no relationship between high-population densities and high rates of infant mortality. In this case, the density of dots should not be very high so that the colours or shades for the underlying districts can be easily determined.

(*b*) Nominal point data

A5.32. The simplest case of a dot map is where each dot represents one discrete element, such as a farm or hospital. Such nominal point data represents categories of features rather than a count or size attribute. In simple point symbol maps, the dot location correctly represents the location of the element. The size, colour or symbol used could reflect different types of features, such as health service centres versus hospitals, as in figure A5.10. One could use simple geometric figures, such as circles, squares and triangles, to represent different types of point features. Alternatively, desktop-mapping or GIS packages allow the user to specify a symbol that matches the type of feature mapped. For example, the map in figure A5.10 shows the distribution of two types of health facilities with easy-to-interpret symbols. The symbols used are typically text font characters or bit-maps. Most packages have their own font sets that provide a large number of cartographic symbols ordered by topics, such as transportation, public utilities or facilities. Some systems also allow the user to import self-designed bit-map symbols. The best practice, however, is to use standard symbol sets wherever possible to reduce the specialization needed for reproduction.

Figure A5.10
Mapping discrete point objects

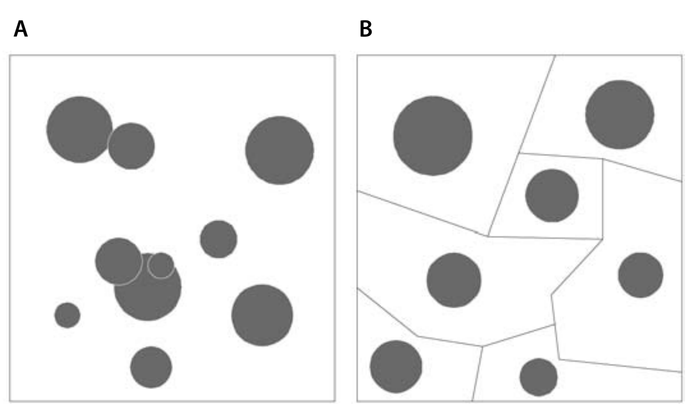


(*c*) Proportional point symbols

A5.33. Point symbols can also be used to map a quantity at a specific location. One popular type of census map, for instance, shows the location and size of major cities using circles or squares that are scaled according to each features numeric values. Such maps are called proportional or graduated symbol maps. Graduated symbol maps are suitable to show the absolute value of a variable. They are less appropriate for a relative value, such as a density or ratio.

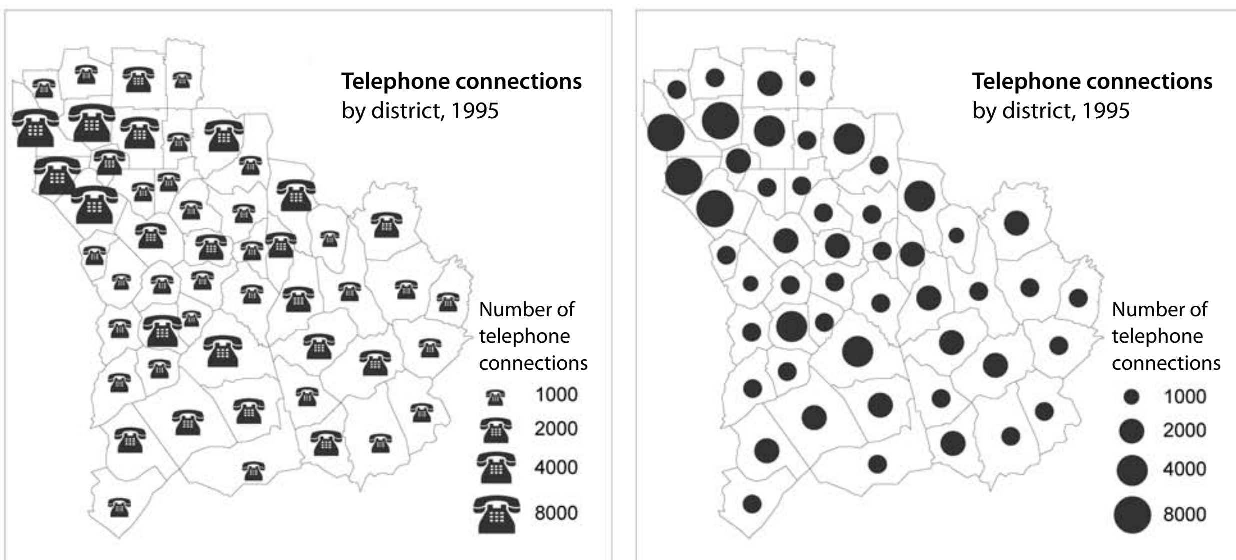
A5.34. There are two types of graduated symbol maps. In one instance, the data refer to a point feature, such as a city or household. In this case, the symbol location corresponds to the feature's location (see figure A5.11A). In the second instance, symbols are used to represent values of area features, such as districts. In this case a representative location within each reporting unit must be chosen (see figure A5.11B). It should be noted that most systems draw a halo around each circle, so that circles that are very close to each other can still be distinguished. The program draws the largest circles first to prevent smaller ones from being covered.

Figure A5.11
Proportional symbols for point and area features



A5.35. As noted above, a computer-mapping package will allow the selection of a symbol that reflects the theme of the map. Such figurative symbols can make the map more interesting to look at. However, if the symbols are too complex, there is a danger of distracting the viewer from focusing on the main information that is to be communicated: the relative magnitude of the variable in different regions. Compare the two versions of a map showing the number of telephone connections in figure A5.12. Even though the telephone symbol is quite simple, it is more difficult to judge the size of the variable in the left map than on the simpler map on the right. The cartographer must create a balance between showing information in a clear and easy-to-understand way, and, on the other hand, making the map attractive. In almost all instances, better results are achieved with simple symbols that do not distract the reader's attention away from the relative magnitude of the variable studied.

Figure A5.12
Pictograms versus simple graphical symbols



A5.36. Proportional symbols can also be used to present two variables at the same time. For instance, the size of the circles could represent the number of households in a reporting unit, while the colour or gray-shade of each circle indicates the percentage of the households that have a telephone connection. Again, the cartographer needs to avoid overloading the map with information. If the number of reporting units is very large or the units are very small, it may be preferable to show the two variables in separate maps.

A5.37. Besides circles, other commonly used geometric symbols include squares and triangles. By varying the orientation of triangles, we can show the magnitude of diverging variables, such as in- and out-migration from each reporting unit (see figure A5.13). Different gray-shades or colours ease interpretation further.

A5.38. Related to graduated symbol maps are maps in which differences in value are represented by the number of standardized symbols drawn for each geographic unit. For instance, total population can be represented as in figure A5.14. This type of map used to be popular in thematic cartography. But, as with figurative symbols, such maps easily become cluttered and difficult to interpret. The magnitude of different values is better represented by proportional symbols.

Figure A5.13
Showing magnitude and direction of flows using simple graphical symbols

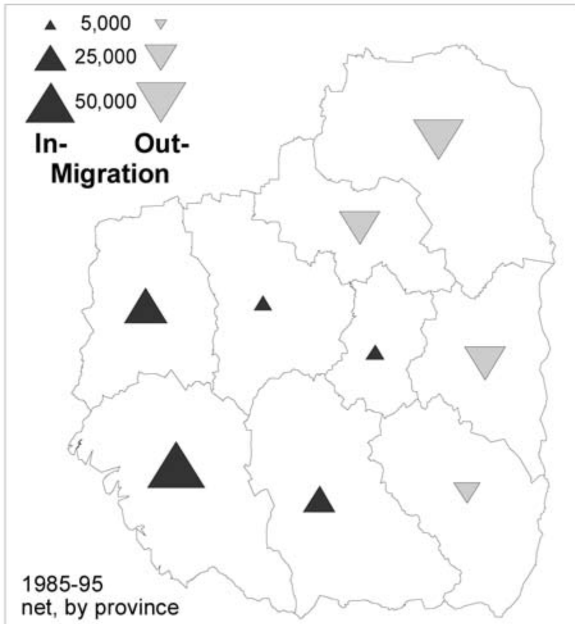
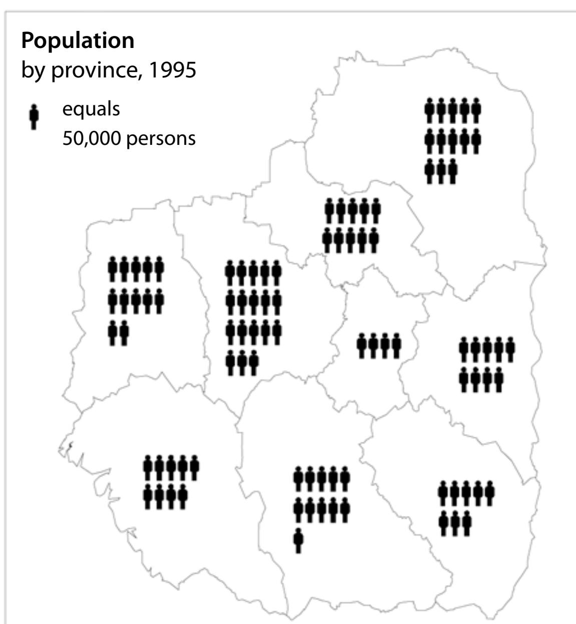


Figure A5.14
Representation of data values by varying the number of map symbols for each feature

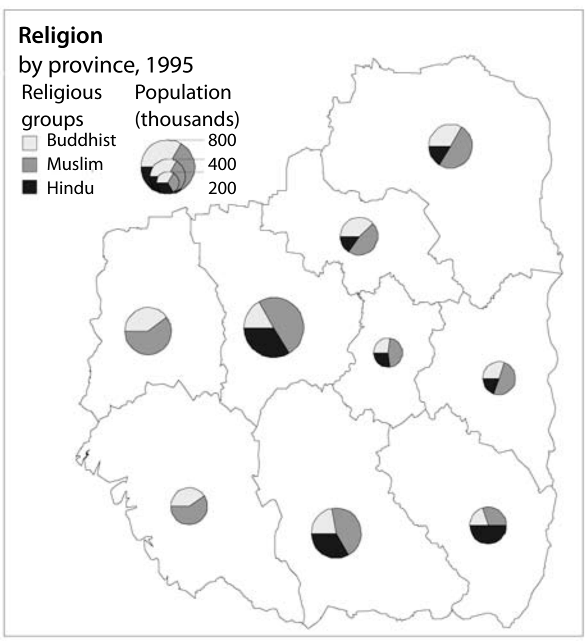


(*d*) Chart or diagram maps

A5.39. Maps that show statistical information in a chart or diagram for each geographical observation have become very popular thanks to their availability in commercial desktop-mapping and GIS packages. As with several of the map types discussed before, diagram maps easily become overloaded with too much information. Unfortunately, there are many published examples of such maps, from which it is hard or impossible to extract useful information.

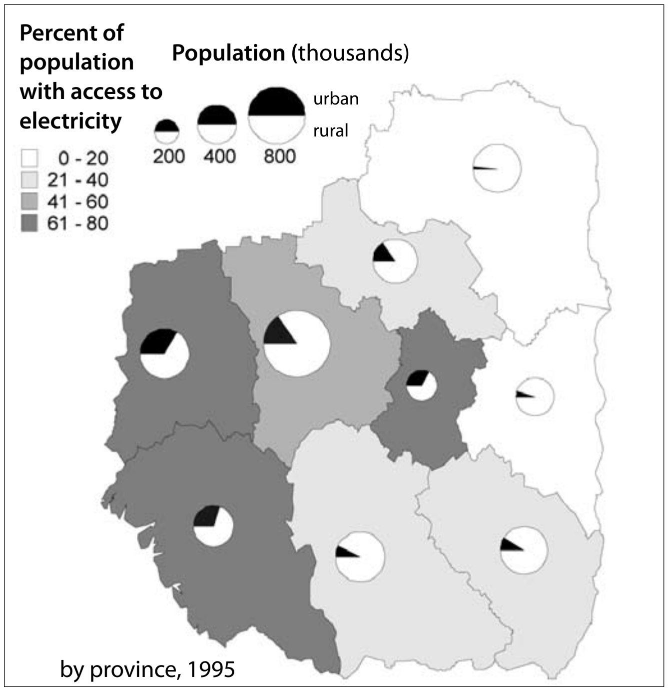
A5.40. The most common types of diagram maps use pie, bar or column charts. The charts are usually scaled so that the size of each pie chart, for example, reflects the magnitude of the denominator. For example, figure A5.15 shows the geographic distribution of the proportion of major religious groups. The pies are scaled according to total population. We therefore need to show two types of information in the legend: the colour that refers to each religious group and the population totals that correspond to a given pie size.

Figure A5.15
Pie chart map



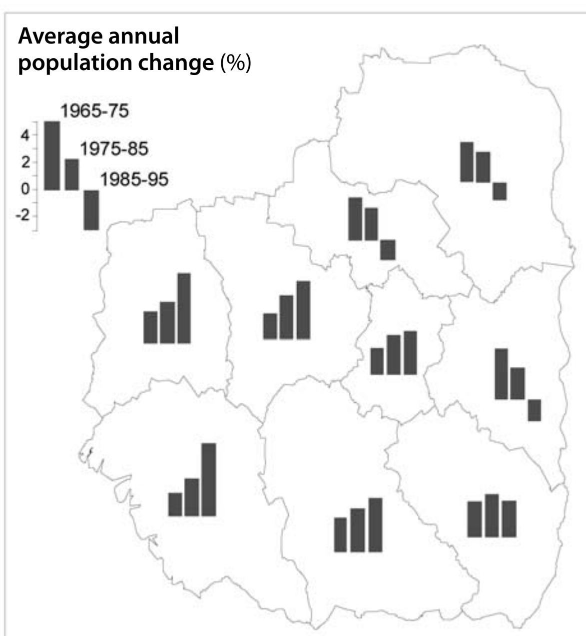
A5.41. Diagram maps work best if there are relatively few geographic observations and very few groups represented. For example, a pie chart map that has only two categories can be very effective in combination with a simple choropleth map to show several variables at once (see figure A5.16): the spatial distribution of different levels of access to electricity, the total population in each province, and the proportion of the population that is rural versus urban. In this map, we can see that there is some indication that provinces with a high proportion of urban population also have a higher percentage of access to electricity. A well designed map that is not overloaded with symbols, colours and shades can support the multivariate analysis of several variables. However, pie chart and similar maps can easily become difficult to interpret, and their use should be limited to cases where the cartographic message does not become obstructed by too many symbols and categories.

Figure A5.16
Combination of choropleth and pie chart maps



A5.42. A further use of chart or diagram maps is to show trends over time. The map in figure A5.17, for instance, shows the average annual percentage change in the population of each province between the last three censuses. The bar charts are very simple, without a border and without a base line, since for these data it is apparent which bars represent an increase or a decrease of population. As before, what we want to convey are relative changes over time, not the exact values which are better presented in a table.

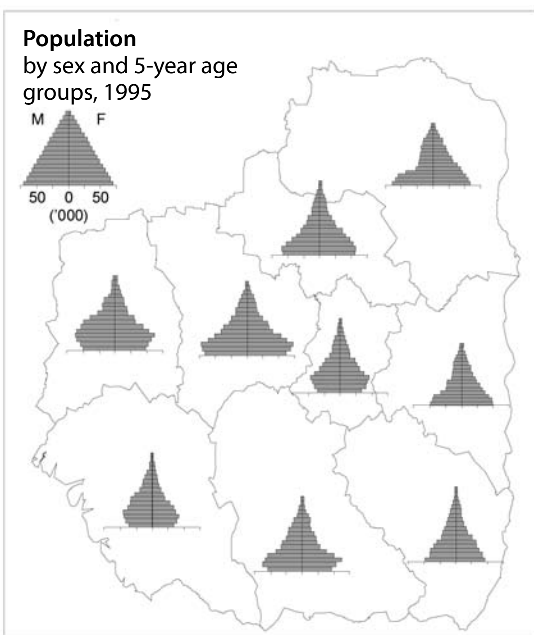
Figure A5.17
Map showing changes over time, using histograms



A5.43. One type of chart that is of great relevance to population census data is the population pyramid. These can be combined with a base map of reporting units to show how the age and sex distribution in the country varies by region (see figure A5.18). Population pyramids are very complex charts. This implies that they can be represented reasonably only if the number of regions on the map is relatively small. Usually, this means that they will be presented in a census atlas at the first subnational level only. A practical problem is that commercial GIS and desktop-mapping packages do not produce pyramid charts automatically. They therefore have to be created externally, for example, in a spreadsheet package, and added to a base map in a graphics package or in the layout module of a desktop-mapping program.

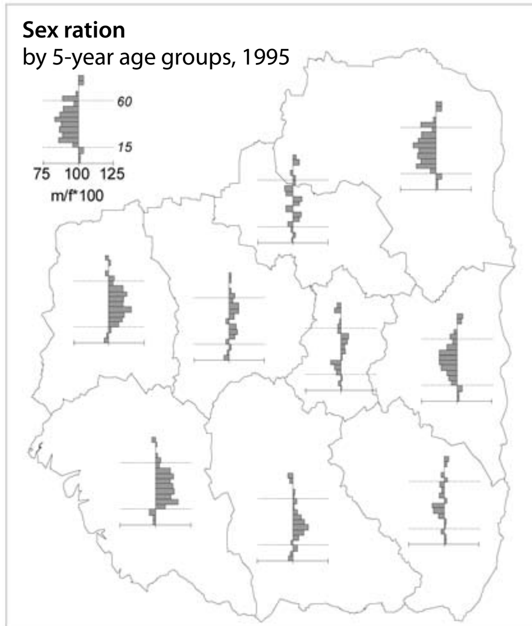
A5.44. Population pyramids shown for several regions are meaningful if there is some variation in the shape of the pyramids. If age and sex distributions are fairly constant across the country, the resulting maps will not provide much insight. In figure A5.18, there is some indication that provinces in the south-east have been experiencing a fertility decline over the past 15 years, while provinces in the north have not. Furthermore, it appears that the provinces in the north-east show a skewed sex ratio distribution. There seem to be more females than males in the age groups corresponding to the economically active population. In the south-west, the situation appears to be reversed.

Figure A5.18
Combination of maps and population pyramids



A5.45. Variations in the sex ratio can be highlighted using a different type of bar-chart, as shown in figure A5.19. These charts show the surplus or deficit of males and females within each province. The trend that was visible in the population pyramid map is much clearer here. Yet the map is fairly complex and visually not very appealing. An alternative way of depicting sex ratios is presented below.

Figure A5.19
Display of sex ratios on a map



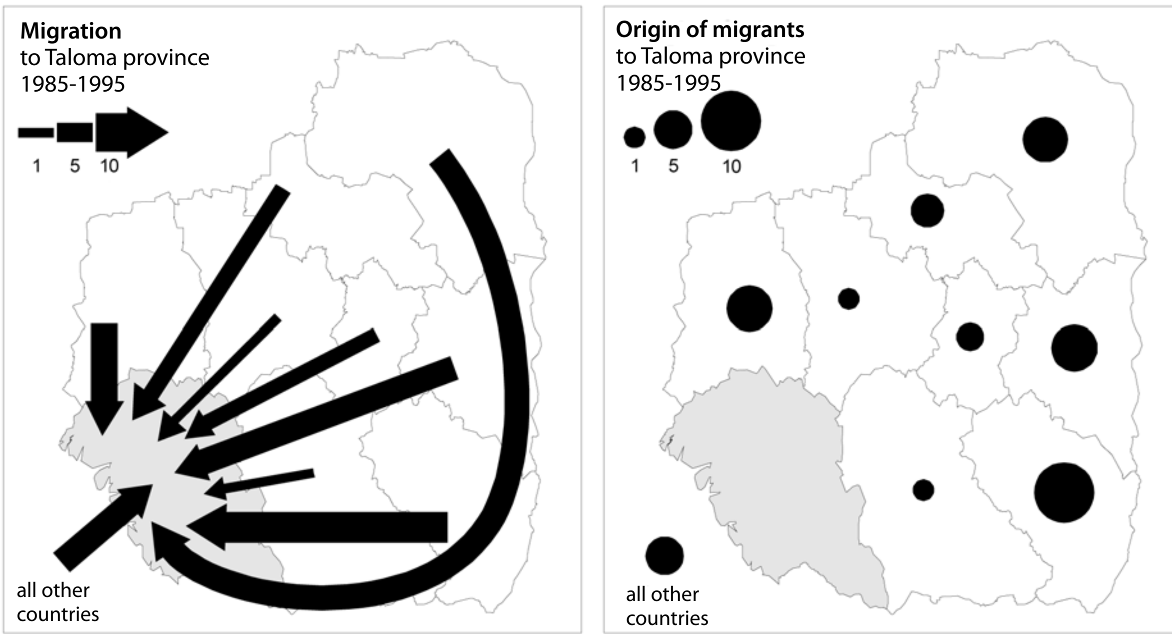
(*e*) Flow maps

A5.46. Migration is a demographic variable that represents the movement of people from one part of the country to another (internal migration) or between the country and the rest of the world (international migration). Migration can be portrayed on maps in several ways. Migration rates are shown using choropleth maps of in-migration, out-migration or net migration rates. The volume of in- or out-migration can be shown using graduated symbol maps (see figure A5.13). Alternatively, if complete migration information is available, flow maps—also called flow-line maps—can be used. These show several aspects of migration: the route of the migration flow, the direction (from-to) by using an arrow symbol, and the magnitude of the flow by varying the thickness of the line.

A5.47. Migration maps can get very complex very quickly. Even with our sample province map of only nine reporting units, there are 72 possible flows—not counting international or within-province migration. Complete flow maps showing all possible migration routes within the region or country are therefore rarely produced. There are several alternative options. One option is to ignore the smallest migration flows and to represent the largest, most significant ones only. Another possibility is to produce separate maps for each province that show only in- or out-migration into or from this province (see figure A5.20). For our sample provinces, this would result in a series of nine pairs of map. Even these simpler maps can get quite crowded. The cartographer often has to create snake-line arrows that wind around the map if origin and destination region are far apart.

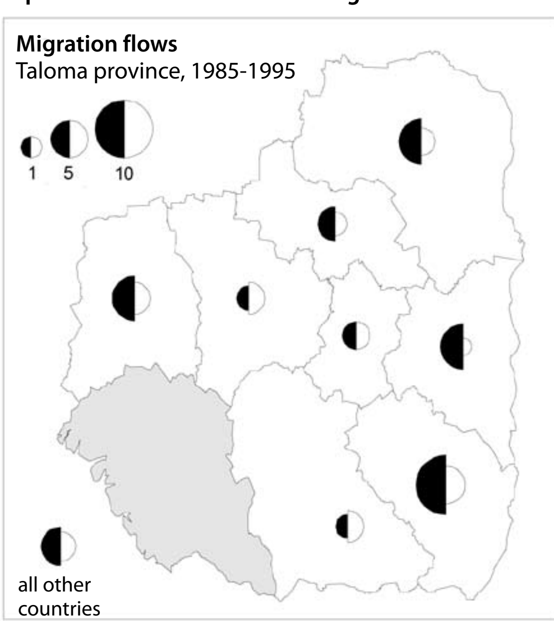
A5.48. In flow maps that use arrow symbols, the visual impression is guided by the length and thickness of the arrow. A longer, narrow arrow may be visually more dominant than a shorter thicker arrow due to its larger surface area. Although a cartographer may sometimes wish to use this fact to point out an interesting migration flow from a remote region, the reader will often have some difficulty assessing the relative magnitude of flows that are represented by arrows of different length. If the focus is on the absolute level of migration from each region of origin, alternative displays are thus more appropriate. For example, instead of arrows one can use graduated symbols to show the magnitude of migration flows by origin or destination (see figure A5.20).

Figure A5.20
Alternative ways of representing flows between regions



A5.49. Using special types of graduated symbols, each map can show both migration into and from the province, as shown in figure A5.21. Here, half-circles of different colours or gray-shades are used to distinguish between in and out migration.

Figure A5.21
Representation of in- and out-migration



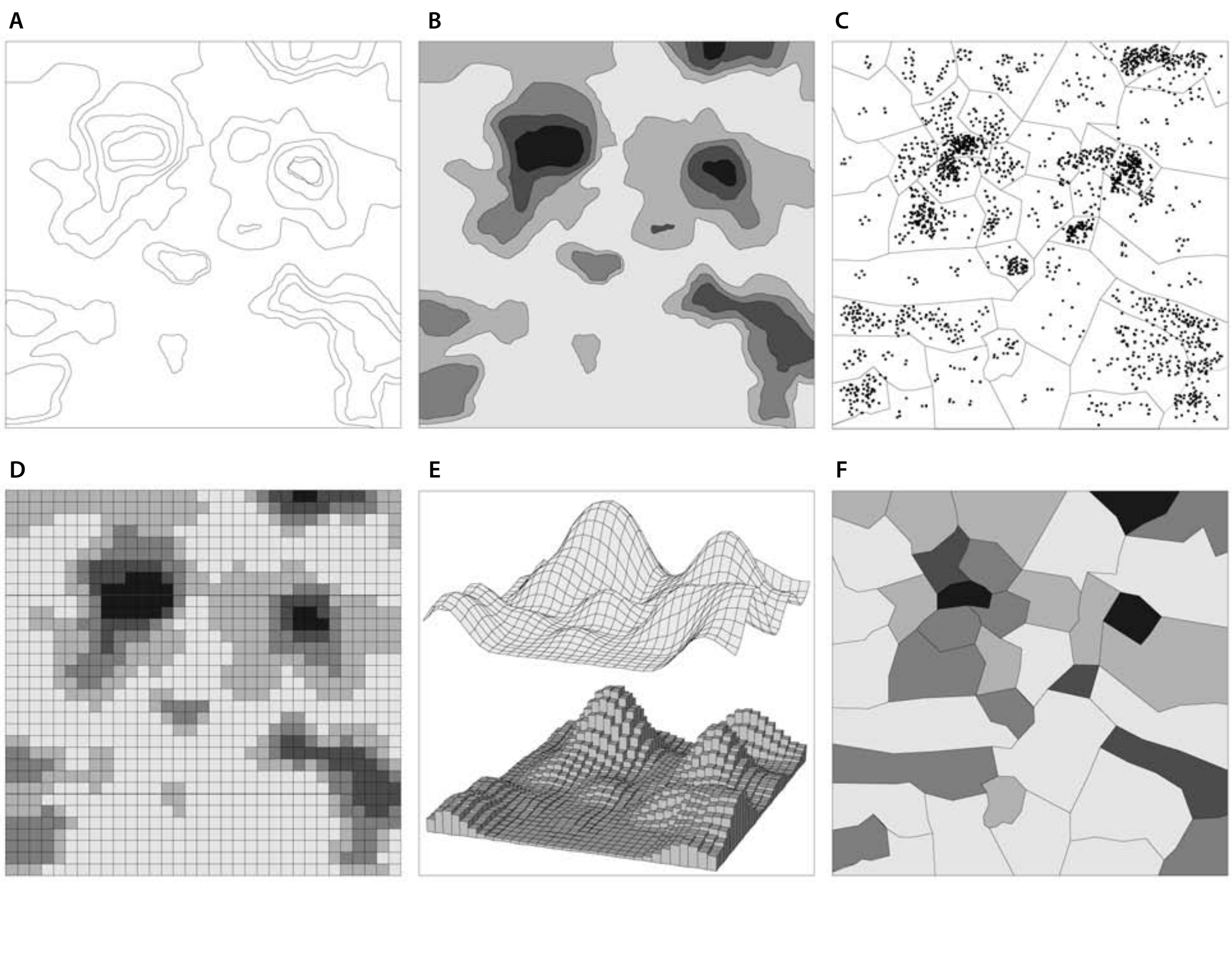
(*f*) Mapping continuous phenomena

A5.50. The map types above are appropriate for data that are referenced for discrete geographic features, such as point locations or areas. Some geographic phenomena, however, are continuous. Temperature or elevation, for example, vary smoothly across space. However, it is also possible to view population distribution as a more or less continuously varying variable. Reporting areas are fairly arbitrary and the aggregate values tabulated for these units hide spatial variation within each unit. Population atlases and, increasingly, GIS data sets, therefore sometimes show population density and distribution as continuously varying.

A5.51. True continuity can not be represented on a paper map or in a computer database easily. Even if we could theoretically derive a different value for each exact point in the country, we need to classify the data in some way for mapping purposes. Several of these are shown in figure A5.22.

A5.52. The most common way of representing continuous data is by means of isolines or regular raster grids. Isolines—the Greek word "iso" means equal—are lines of constant value and are also called contours (see figure A5.22A). They are used on topographic maps that shows elevation. Contour maps can also be shaded which causes them to look more like choropleth maps (see figure A5.22A). Colours represent values in the data range between two contour intervals. Dot maps can also be used to provide a more continuous view of the distribution of population or a similar vari-

Figure A5.22
Alternative cartographic methods for displaying continuous data



able. As described above, most GIS packages produce dot maps by randomly drawing points within each reporting unit. In this case, we do not gain any additional information compared to a choropleth map. But if the dots are placed according to additional information on land cover or village locations, for example, a more continuous image of the variable's distribution is possible (see figure A5.22C).

A5.53. For modelling and analysis in a GIS, continuous data are typically stored as regular raster grids (see figure A5.22D). The grid cell size is chosen to preserve the variability in the data set, although a very fine grid leads to very large file sizes. Finally, computer mapping packages as well as general graphics software, provide various ways of showing continuously varying data sets as a surface. In figure A5.22E, two examples are shown: a wire-frame model and a two-dimensional bar chart. Such techniques are very useful in showing terrain information based on a digital elevation model. Sometimes, such maps can also show population distribution very well. In such maps, hills and peaks represent clusters of very high population density, while valleys indicate sparsely populated areas. For population and similar socioeconomic information, however, it is often difficult to assess the true spatial distribution on surfaces. While we are intuitively able to interpret elevation heights, it is much more difficult to quickly associate surface heights for other variables with their respective values. Additional standard-mapping techniques are therefore generally more appropriate. For comparison, figure A5.22F shows a choropleth map, in which the reporting units are not determined by the data distribution.

C. Data classification

A5.54. The tools available to the cartographer to display thematic information on maps were described above. The map designer must choose the graphic variables and thematic map types that are most appropriate for the variable that is mapped. In some instances, there will be a one-to-one match between symbol types and variable values. This would be the case where a small number of nominal categories are represented, for instance, with point symbols of similar size but different shape. Even with categorical data, however, there is often a need to represent several features that have similar values by the same graphical symbol. For instance, single-family and multi-family households might both be represented by the same point symbol. Numerical data almost always need to be categorized before they can be matched to symbol sizes or colours.

A5.55. The process by which observations with similar values are lumped together to be represented by the same graphical symbol is called classification. It is similar to classification methods in statistics that group values into categories so that the variance of the observations in the same category is minimized and the variance between different categories is maximized. Computer-mapping packages provide default methods for assigning symbols to values or value ranges. Although graphical user interfaces (GUIs) have improved, defaults may not be appropriate for the variable that is mapped—most often they are not. Indeed, automated classification tools often lead to inappropriate or even misleading map designs. Some of the classification options are discussed in more detail below.

A5.56. Classes for numerical data are usually contiguous value ranges. The number of classes is determined by several factors: the data distribution (i.e., the variation of values in the data set), the intended accuracy of data representation, and—last but not least—the ability of the output device to show small differences between colours and texture. More classes do not necessarily improve a thematic map since it is increasingly difficult for the viewer to distinguish between classes. It is more important to determine the class ranges in a way that accurately reflects the variation in the data set.

A5.57. Which classification technique is appropriate depends on the data distribution of the variable. A method that yields an accurate and visually appealing map for a data set that is uniformly distributed (e.g., there is an approximately equal number of high, medium and low values) may not work well for a very skewed data distribution—i.e., one that has many low values and only a few very large values.

A5.58. For the preparation of publication-quality maps, the data should therefore always be evaluated using statistical graphs. Some GIS and desktop-mapping packages, unfortunately, have only limited charting capabilities, but they do allow the exporting of data to spreadsheet or statistics packages which provide extensive charting functions.

A5.59. The most useful type of chart for determining class ranges is a rank-order plot. All data points are sorted according to their values from low to high. They are then plotted next to each other—the x-axis shows the rank of each observation and the y-axis shows the data value. Vertical gaps or "natural breaks" between neighbouring data points are good candidates for class boundaries, although there may often be more or fewer gaps than the desired number of classes.

A5.60. Examples are set out below of common classification methods for three variables with different statistical data distributions. The population density variable has a skewed distribution. There are many small values in the range of 21 to about 110 persons per km^2 and only a few very large values. The largest value (791) is nearly two and a half times the second largest value (320). This is not unusual for population density. The very high district, for instance, might contain the capital of an otherwise rural province. The second variable is the literacy rate for the districts. The values are quite uniformly distributed, which is indicated by the nearly straight line that the observations form in the rank-order plot. There are no extreme values.

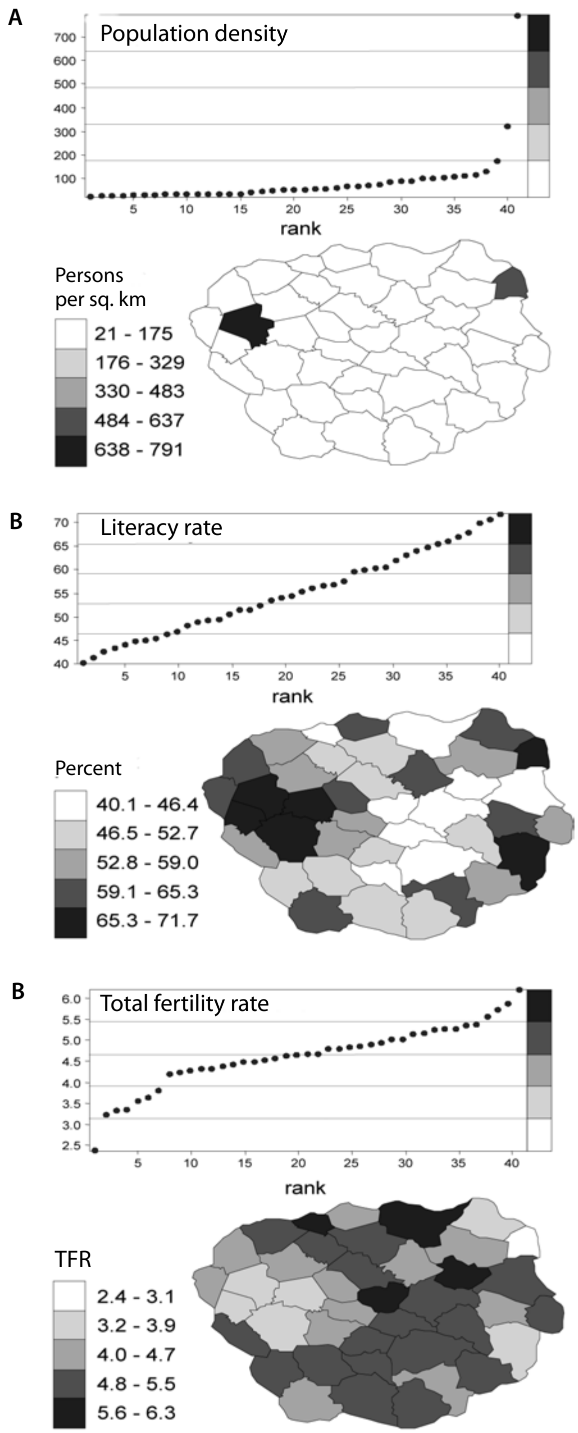
A5.61. The third example variable is the total fertility rate (TFR). The rank-size plot shows a fairly steep increase in values for the lowest observations, a large middle section with a much less extreme increase, and again a more rapid increase of values for the very high observations towards the right. This indicates a normal distribution, which is characterized by a smaller number of extremely low and high values, and many observations in the middle ranges. Of course, the examples here are for illustrative purposes only. The same variables for other geographic areas may show very different distributions.

A5.62. The examples will show that the appearance of a map depends crucially on the choice of classification method, which may or may not be appropriate for the data distribution. This confirms that automated classification methods that are provided in GIS packages should be used with some degree of caution.

1. Serial data classification

A5.63. One of the simplest methods of classification is to divide the range of data values into "equal intervals" (see figure A5.23). The cartographer first determines how many classes will be used. The range of data values—the highest value minus the lowest—is then divided by the number of classes to obtain the increment, also called common difference. The first class then ranges from the lowest value to the lowest value plus the increment, and subsequent classes are determined by adding the increment to the previous upper range value. Some rounding may be necessary if the numbers are shown with low precision in the legend.

Figure A5.23
Equal intervals



A5.64. For the population density variable, the lowest value is 21 and the highest is 791. The range is therefore 770. Given the intent to use five categories, the common difference is thus 770/5 which is 154. So the first class would be from 21 to 175, the next from 176 to 329 and so on.

A5.65. The map for population density illustrates why this can lead to problems. The range of the values is influenced by one very large value. The common difference is therefore so large that the first class range includes all but two of the observations. Clearly, the resulting map is not very informative.

A5.66. The method works much better for literacy rate, which is more uniformly distributed. The data set is divided into approximately equal numbers of observations in each class and the resulting map gives a good impression of the distribution of literacy across the districts.

A5.67. Finally, the map for TFR shows similar problems as the one for population density, although much less extreme. There is only one observation in the lowest class range and the map is somewhat dominated by values in the middle class ranges. By coincidence, however, the class breaks between the second and third and between the fourth and fifth categories capture the breaks in the data distribution quite well.

A5.68. In addition to equal intervals, there are other options for serial data classification. One is to use a steady geometric progression, such as 0-2, 2-4, 4-8, 8-16, and so on. This can work well for skewed data distributions, such as the population density variable.

2. Statistical classification

A5.69. One classification method is to have an approximately equal number of geographic observations in each category. This can be implemented by using the statistical concept of quantiles that divide the data set into classes with the same number of observations. If there are four classes, they are called quartiles, if there are five they are called quintiles, and so on.

A5.70. To determine the quantiles, the number of observations is divided by the number of desired categories and, if necessary, rounded to the nearest integer. In the rank-order plot, the first "n" observations are then assigned to the first category, the next "n" to the second, and so on. Any odd number is assigned to the first or last category.

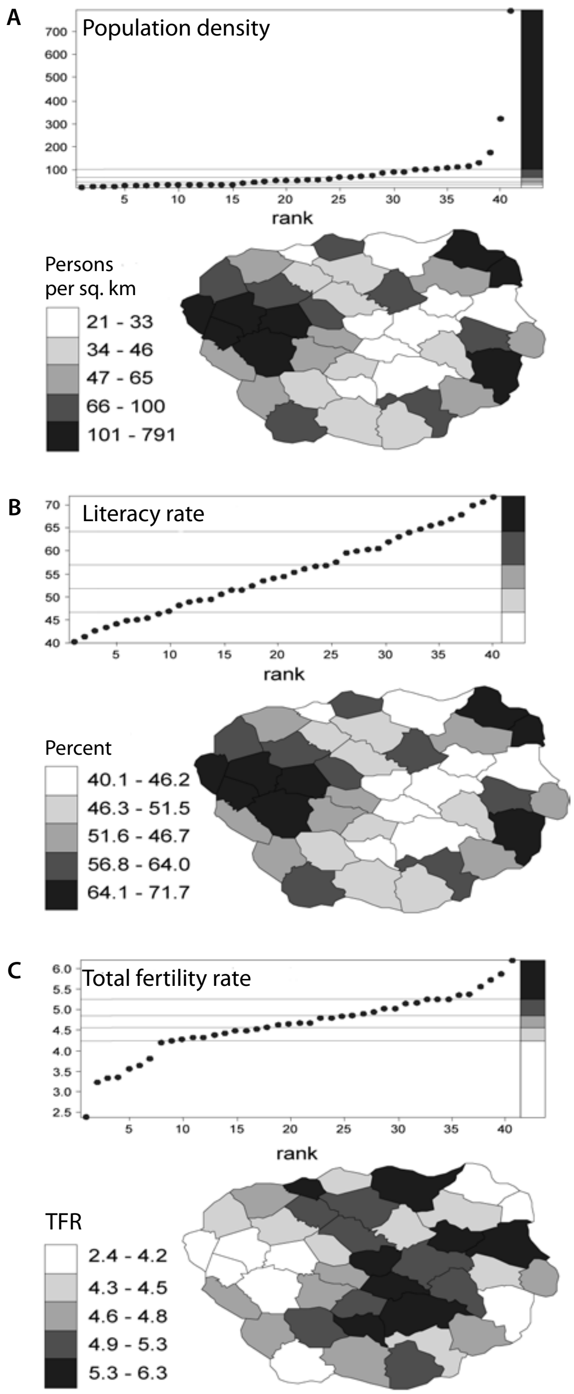
A5.71. Quantile mapping is implemented in many desktop-mapping packages and this method has therefore become very popular for map production.

A5.72. The three sample maps all look very good. There is, by definition, a good distribution of observations across classes so that all maps make good use of the full gray-scale range.

A5.73. Looking at the data distribution, the classification for the literacy rate variable seems quite appropriate. In fact, the map does not look much different from the one that uses equal intervals.

A5.74. However, in the population density and TFR maps, it is clear that the method groups very similar values into different categories. For TFR, for instance the two observations with the largest values in the lowest data range (2.4-4.2) are much more similar to observations in the second category than to the observations in the first category. Even worse, there are three observations with a value of 5.3, one of which is assigned to the fourth-class range and two to the fifth (some desktop-mapping packages relax the criterion of equal number of observations to avoid such cases).

Figure A5.24
Quantile (equal frequency) mapping



A5.75. Quantile maps should therefore be used with caution. Quite often, similar values will be assigned to different categories and dissimilar values will be grouped in the same class. Although the resulting maps are visually appealing, the impression may be quite misleading.

A5.76. Another statistical classification technique is based on summary measures of data distribution. One option is to determine class ranges using the standard deviation of the variable distribution. The standard deviation is computed as the square root of the variance. The variance is calculated as the mean of the squared differences between the data values and the overall average value. For example, for the literacy rate variable, the standard deviation is 8.9.

A5.77. Map categories based on standard deviations therefore show how individual observations—such as districts—compare to the average value for the entire province or country.

A5.78. Categories are determined by subtracting and adding the standard deviation to the mean (55 for the literacy rate). The class ranges are therefore constant, similar to the equal interval method.

A5.79. For the literacy rate, the first data range (40.1-46.2) corresponds to values that are more than one but less than or equal to two standard deviations below the mean. Since the data distribution is quite compact, all values are within +/– two standard deviations and only four categories are needed. As is clear in figure A5.25B, the method divides the literacy rate values into approximately even numbers of observations in each class, which yields a map with good visual contrast.

A5.80. For the population density variable, however, the approach is much less appropriate. Because of the many small values, the mean population density is quite low (85.4) and the standard deviation is quite high (124.8). The first category—corresponding to values that are within one standard deviation from the mean—should therefore actually range from -39.5 to 85.4. On the other hand, the highest value (791) is more than five standard deviations from the mean. We would therefore need to use many more classes, most of which would not contain any observations. Instead, the largest class for the map presented here includes all values larger than one standard deviation from the mean. Clearly, standard deviations are not a good choice for this variable.

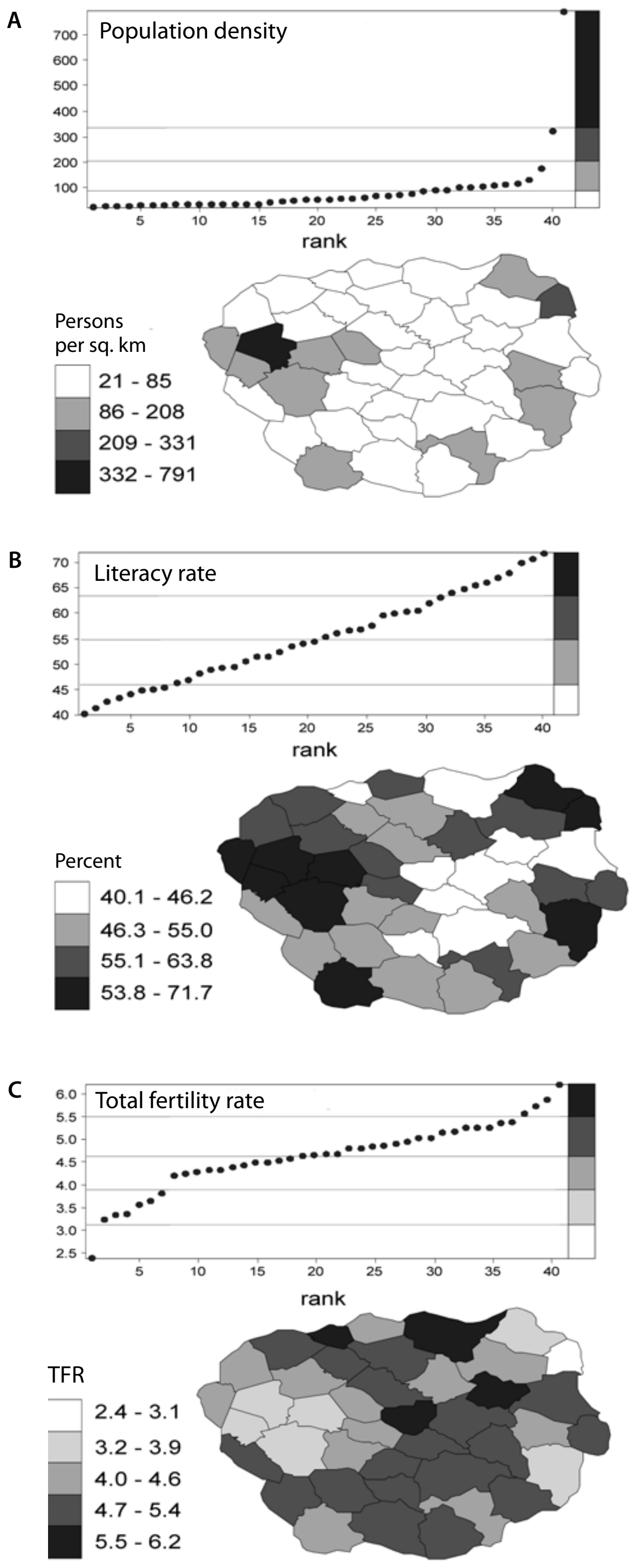
A5.81. Standard deviations work a little better for TFR, with a mean of 4.6 and a standard deviation of 0.8. However, only the very low value of 2.4 falls into the lowest category, which is more than two standard deviations below the mean.

A5.82. The standard deviation classification method has intuitive appeal because of its close relationship to descriptive statistical techniques. It works well if the data are normally distributed with relatively low variance so that, at the most, six categories will include all values.

A5.83. Standard deviations can be used to represent different types of trends in a data set (see figure A5.26; see also Dent, 1999). In the examples given in figure A5.25, a gray-scale from light to dark is used. The maps highlight the progression from low to high values of population density, literacy and TFR corresponding to a categorization, as shown in figure A5.26A. This is, in fact, the least common application of standard deviation classification.

A5.84. The method is more commonly used to highlight diverging trends. For instance, to show income levels, the poorest and richest districts may need to be highlighted. In that case, strong colours or texture would be assigned to the districts with values of more than one and two standard deviations from the mean, while relatively

Figure A5.25
Standard deviation



muted shades would be assigned to those located in the centre of the data distribution (se figure A5.26B).

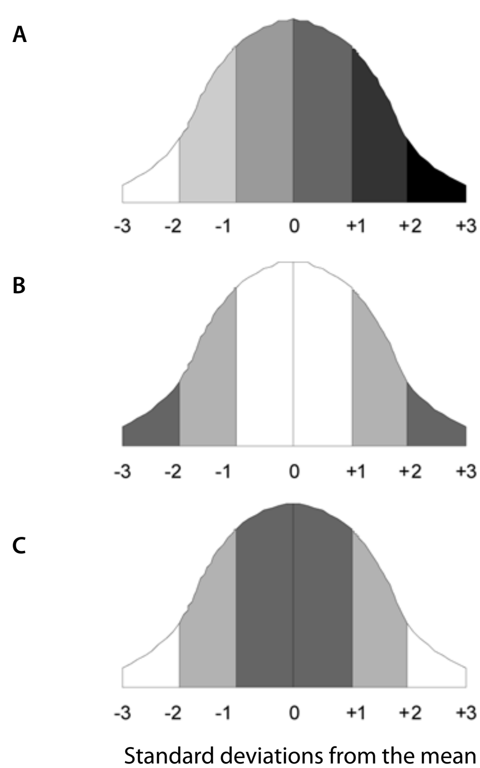
A5.85. If only the distance from the mean is of interest—regardless of whether the values are above or below the mean, then the same colours may be used on both sides. If the interest is also to determine whether the values are above or below the mean, then different colours or textures should be used on either side. For example, on a map printed in colour, the classes below the mean could be assigned red shades from light to dark, and the ones above the mean could be shown in corresponding blue tones.

A5.86. In other cases, we may wish to highlight the middle ranges (see figure A5.26C). MacEachren (1994), for example, discusses a map of Northern Ireland (United Kingdom) published in Fothergill and Vincent (1985) that shows the proportion of Protestants and Catholics. In this map, values around 50 per cent, which indicate an approximately equal balance of Protestants and Catholics, are highlighted by assigning the middle classes a strong colour (yellow). Areas where either Catholics or Protestants have a clear majority are shown in more muted colours (green and orange, respectively).

3. Natural breaks

A5.87. As noted above, most methods produce maps that are somewhat misleading for variables that do not have a very uniform distribution. It often happens that similar values are assigned to different classes or very different values are grouped

Figure A5.26
Assignment of shades for classes determined by standard deviations



together. A logical approach to data classification in cartography is therefore to find a grouping that optimizes the assignment by minimizing the differences between values within each category and maximizing the variation between groups.

A5.88. This objective can be implemented by visual inspection of the data distribution and subsequent choice of class breaks. Examples of this approach are presented in figure A5.27. For the TFR variable, this is quite straightforward, since it shows several distinct break points in the distribution.

A5.89. It is somewhat more difficult for the two other variables. For population density, a strict application of the method might assign all low values to the same category and the higher values into a number of separate classes. This needs to be balanced with the desire to preserve the subtle variation in the lower value ranges.

A5.90. Similarly, for the uniformly distributed literacy variable, the class breaks are not very distinct since the value difference between observations does not vary much.

A5.91. But because classification according to natural breaks explicitly considers the data distribution, the method usually results in accurate cartographic representations of the data and good visual contrast.

A5.92. Rather than relying on somewhat subjective judgment, one can also let the computer determine natural or optimal break points. Many GIS and desktop-mapping packages provide functions that determine natural break points based on an automated evaluation of data distribution (Jenk's optimum classification method). Classification or clustering functions in statistical software packages can also be used.

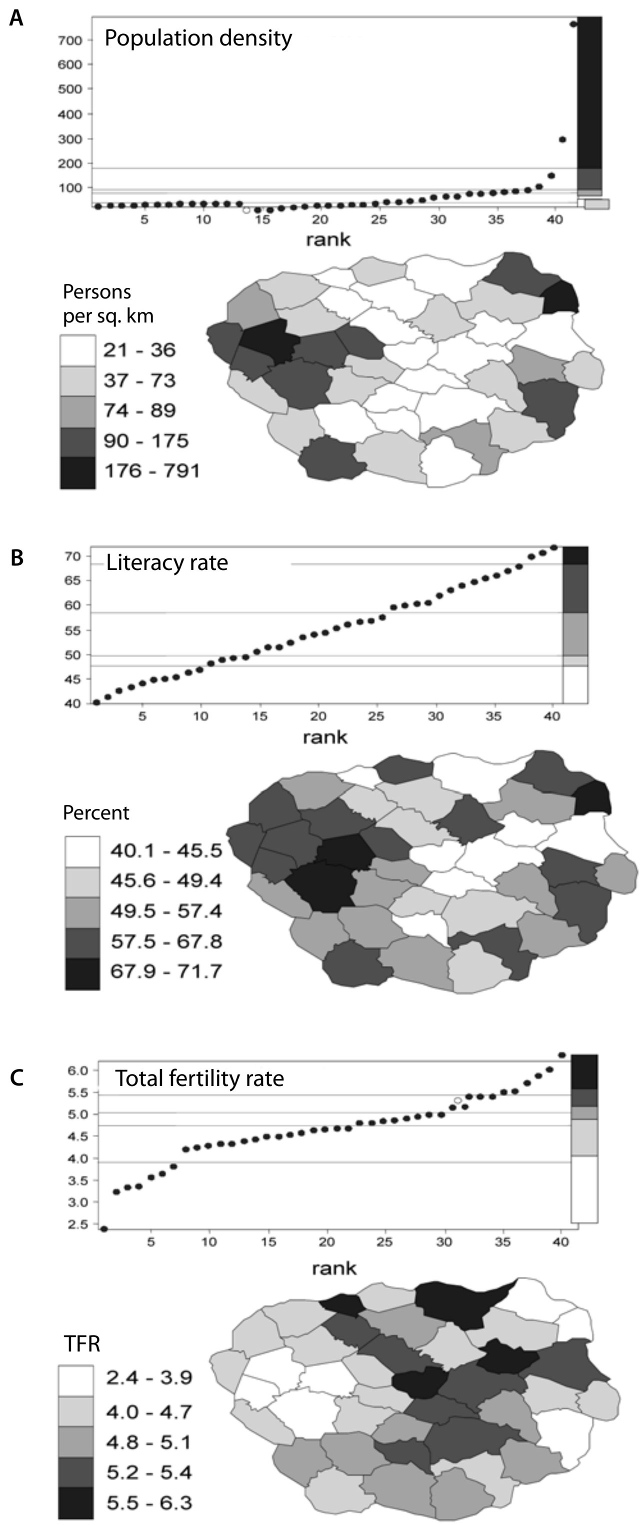
4. Choropleth maps without class intervals

A5.93. So-called unclassed choropleth maps do not require any choice of classification method by the cartographer. Thanks to improved display and print technology, computer screens and printers can produce a large range of different colour tones or gray-shades. For an unclassed or "n" class map, the data values would directly determine, for instance, the percentage of gray level. For example, for a percentage variable, the corresponding gray level may be selected, on a scale from 0 per cent gray (white) to 100 per cent gray (black), that corresponds to each observation value. Although, if the reproduction method can produce a sufficient number of distinguishable shades it is advisable to avoid white as a shade colour since it is usually also the page background colour.

A5.94. In practice, this may not yield optimal results, however. One reason is that many variables do not range from 0-100 but instead have values concentrated in a smaller range. The map may thus end up with very light or very dark gray shades only. We can avoid this problem by "stretching" the data distribution: using the lightest colour for the standard lowest value and the darkest for the highest will produce maps that are easier to interpret for the viewer.

A5.95. In general, however, there is a limit to the number of gray shades or colours that can be distinguished easily. While a continuous shading scheme is useful for analytical purposes, the classification of data values into a small number of categories is generally preferable for presentation maps.

Figure A5.27
Natural break points



5. External data classification

A5.96. In some instances, the classification scheme is given externally. For instance, to prepare a map of poverty by district, the cartographer uses a given threshold value of average income—a so-called poverty line—below which a district is considered poor. Another instance where a classification scheme is given is where comparisons are made with existing printed maps for which the original data are not available. For example, to allow for accurate comparisons between maps of the fertility rate in the provinces of a country, the classification needs to be identical.

6. General remarks

A5.97. This overview has shown that there are many methods for assigning data values to categories. Most GIS and desktop-mapping packages support equal intervals, quantiles, standard deviations and natural breaks. Additionally, all packages allow the user to define a custom data classification.

A5.98. Each method has strength and weaknesses, which are highlighted in table A5.1. Which one is appropriate depends on the data distribution and on the purposes of the map. In general, the data distribution should always be evaluated using a statistical chart, such as the rank-order plots shown above. Such tools are included in some GIS programs. The optimal number of categories and the best break points will then often be quite obvious.

A5.99. Note that natural breaks are not appropriate if several maps are presented together for comparison—for example, for a time series of sex ratios by district or maps of access to safe drinking water for two provinces in the country. In such cases, the class breaks need to be kept constant. A user-defined classification scheme based on an evaluation of all data series needs to be chosen for this purpose. Quantile maps can also sometimes be used if the objective is solely to compare the ranking of different observations over time or space rather than the actual data values. For example, two quartile maps could highlight the 25 per cent of districts with the highest literacy rates as determined in the last and previous census.

Table A5.1
Evaluation of different classification techniques

Classification method	Advantages	Disadvantages
Equal intervals	Easy to implement Appropriate for uniformly distributed data	No relationship between classification scheme and data distribution Since class intervals are fixed, similar values may be assigned to different classes, dissimilar values to the same class Not appropriate for skewed data distributions or data sets with outliers
Geometric progression	Easy to implement Appropriate for data with a very skewed distribution (e.g., many small and few very large values)	Proper geometric progression must be determined by the user Since class intervals are fixed, similar values may be assigned to different classes, dissimilar values to the same class
Quantiles (equal frequency)	Good visual contrast is ensured Appropriate for fairly uniformly distributed data	Similar or identical values may end up in different categories

Classification method	Advantages	Disadvantages
Standard deviations	Good for showing diverging trends centred around the average value Relates individual categories to overall mean value Appropriate for data with a normal distribution	Skewed distributions or data sets that contain outliers (few very large or very small values) will lead to a large number of categories (i.e., several standard deviations above or below the mean)
Natural breaks	Similar values will be assigned to the same category Number of categories is often suggested by the number of break points	Resulting class ranges may be very uneven Requires subjective judgment (visual determination) Does not support comparison of maps over time
Unclassed choropleth maps	No category break points need to be defined Gray-shade or colour tone is determined directly by data value Highlights continuous value distribution in the data set	Most output devices only support a limited number of distinguishable gray-shades or colour tones Maps with subtle gray or colour differences do not reproduce (e.g., photocopy) well Not easily implemented in most GIS and mapping packages

D. Colour choice

A5.100. The map examples presented in the present *Handbook* all use gray-scales for symbolization. Black-and-white publications are less expensive to produce and gray-scale maps retain legibility when duplicated on a black-and-white photocopier. The use of colour, on the other hand, gives the cartographer many more options for map design. Colour printers and plotters continue to drop in price. Also, many more maps will be presented electronically on web sites or electronic publications. Here, colour can be used extensively in map design although it should still be used judiciously.

A5.101. Knowledge of how a computer interprets colours is useful when a colour scheme needs to be defined for a choropleth map. Colours are defined in a computer by using one of several colour models. Two of the most common are the hue-value-saturation (HVS) and the red-green-blue (RGB) models. The term "hue" refers to what we usually mean by colour, such as "red" or "blue". Physically, hue relates to the spectral range of the reflected light and ranges from violet with a low wavelength to blue, green, yellow and orange to red, which has the highest wavelength in the visible spectrum. "Value" is also sometimes called lightness (i.e., hue-lightness-saturation (HLS)). It determines the difference between, for example, a light pink and a dark red, which would both have the same hue. Finally, "saturation" is a measure of brightness or intensity. A colour with a lower saturation will appear more pale or gray, while one with a high saturation value will look more pure.

A5.102. RGB is a model in which new colours are defined additively by combining different levels of red, green and blue. Computer or television screens use the RGB method. Equal levels of the same three colours result in gray shades. The lowest levels of red, green and blue combined produce black, the highest values produce white.

A5.103. Colour choice depends on the measurement level of a variable, the type of map used and the message that the cartographer wishes to communicate. Colour hues may be easily differentiated, which makes them very useful for distinguishing between discrete categories. For example, blue circles can be contrasted with red circles to show different types of schools. One consideration when choosing colour hue to solely distinguish between map symbols, however, is colour blindness. Colour-blind persons may be unable to distinguish between red and green—the most common form of colour blindness, affecting about 1 per cent of males—or between blue and yellow. Some people are unable to see the green part of the colour spectrum. In general, it is good practice not to rely on red-green differences in map composition.

A5.104. Continuously measured variables, such as population, income or ratios and percentages, are presented using graphic variables that show a distinct ordering. Differences in colour value (e.g., from light to dark shades of the same hue) are easily associated with magnitudes of a variable, whereby darker shades are typically associated with higher data values. For instance, levels of population density are often represented with red tones, reaching from very light reds for low population density to dark reds for areas of high population density. For skewed data distributions, colour values cannot be correlated in direct proportion to the values of the data categories. For the population density variable example used above, for instance, this would result in many very light and barely distinguishable shades of light red being used for the many low values, and a very dark shade or colour for the few high values. Instead, equal steps of colour value are used to represent classes of a geometric or similar progression.

A5.105. If a classification consists of many categories, a cartographer may end up with more categories than can be clearly distinguished on a printed page. In that case, colour hues that are adjacent to each other may be combined—a so-called part-spectral colour range. To use the population density example again, the range would start from light yellow shades and progress through oranges to dark reds. The important thing to consider is that there must be a clear progression from less dominant to more dominant colours. Maps that use several dominant bright hues for low and high values of a continuous or ordinal range of categories do not communicate a clear message and are quite confusing to the viewer.

A5.106. One application where different hues are appropriate for a continuous data range is a diverging data scale. For instance, a map of net-migration by administrative unit would have categories ranging from high negative numbers for large out-migration through zero to large positive values which reflect large in-migration. To highlight the large negative and positive values—areas in which migration has the most significant impact on population dynamics—we can use a colour scheme ranging, for instance, from bright red to light red or pink through white for the near zero net-migration rates and through light blues to a bright blue for the highest in-migration.

A5.107. A final comment relates to multivariate mapping, where two variables are shown in combination. For instance, a map could show a combination of different levels of literacy rates and fertility, using a legend that is essentially a matrix of possible combinations of literacy and fertility categories. The cartographer must find an appropriate colour scheme that would, for example, indicate differences in literacy by adjacent hues in a partial-spectral scheme, and differences in fertility rates through variations in colour value. Unfortunately, such maps are not easy to interpret. The viewer constantly has to consult the legend to match colours to data values for the two variables. Generally, multivariate maps should therefore be avoided. Section A5.6 will present some alternative approaches to presenting multivariate information geographically.

A5.108. Returning to the types of measurement levels that were discussed above, table A5.2 summarizes guidelines concerning the use of gray shades and colours (see also Brewer, 1994).

Table A5.2
Choice of gray shades and colour

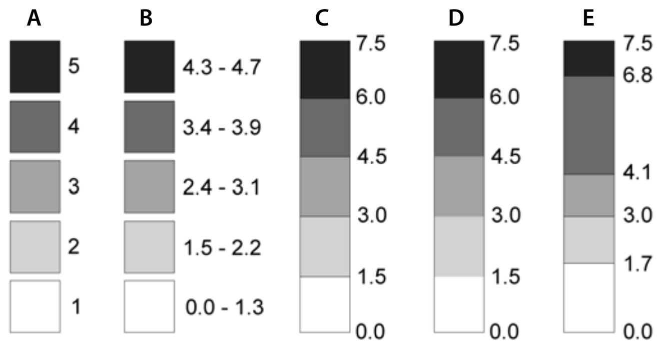
Measurement levels		Example	Black-and-white maps	Colour maps
Nominal	Binary	Access to safe drinking water (yes/no)	White versus black or light gray versus dark gray	Strong contrasting colours of different hues, such as blue and red or yellow and green
	Categorical	Dominant language (French, English, Spanish etc.)	Variations in patterns with similar visual dominance	Different hues with similar levels of value and saturation that do not imply any ordering, for example, blue, green, yellow, violet
Ordinal		Educational attainment (primary school, secondary school, etc.)	Ordered gray shades with relatively strong differences between the gray levels; differences in texture highlight the ordinal nature of data even better	Same hue or partial spectral colour range with relatively large differences between the categories. For example, light yellow, orange, medium red, dark red
Discrete		Household size (1, 2, 3, … persons)—but not average household size!	Similar to ordinal data, but smaller differences between gray shades are acceptable	Similar to ordinal data, but smaller differences between gray shades are acceptable
Continuous	Sequential	Literacy rate (any value between 0 and 100 per cent)	Continuous range of gray shades; level of gray may or may not be proportional to data values; subtle but distinguishable differences in gray levels are acceptable	Continuous colour range within the same hue or within a partial spectral range; subtle variations in colour value are acceptable
	Diverging	Sex ratio (below one = more women than men; above one = more men than women)	Texture/pattern differences must be used; solid fill shades on one side and texture differences on the other side can work to good effect	Neutral colour (white or gray) in the centre, with continuous range of two different hues on either side; for example, from light to dark oranges for values below one, and light to dark greens for values above one

E. Map legend design

A5.109. The measurement level can be reflected in the design of the legend, which provides the reference between data values or value ranges and the graphical symbols used. GIS and desktop-mapping packages provide a built-in design for legends that satisfies most applications. For more careful cartographic design, however, we can modify the default legend either in the layout module of the mapping packages or in external graphics software.

A5.110. Figure A5.28 shows some examples. For categorical data, individual legend boxes should be kept separate (see figure A5.28A). Similarly, class ranges that are not contiguous—e.g., there is a gap between the upper limit of one class and the lower limit of the next one—can be emphasized this way (see figure A5.28B). In general, however, the use of such legends should be avoided. Contiguous legend boxes highlight the continuous nature of variables, such as ratios or densities (see figure A5.28C). Continuity of data values is emphasized even more if the individual category boxes are not enclosed by an outline (see figure A5.28D). A legend for a classification of a continuous variable with irregular class ranges, finally, is shown in figure A5.28E.

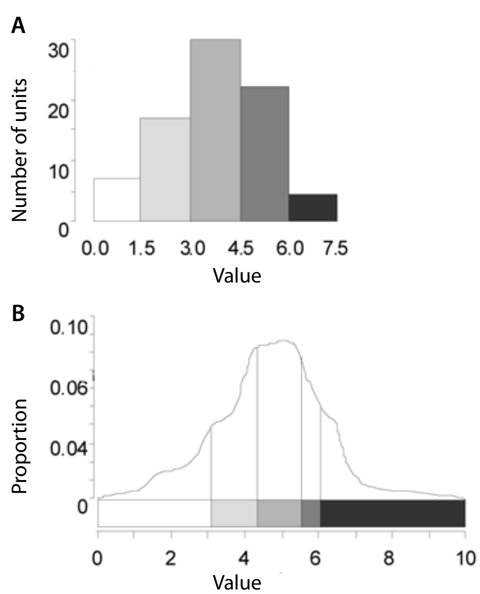
Figure A5.28
Different types of legends for shaded maps



A5.111. The legends in figures A5.28C, D and E show break-points rather than separate data ranges. When using data ranges for a continuous distribution, we encounter the problem of showing a data value for two categories: e.g., 0-10, 10-20, 30-40. This problem can be overcome by using the "less than" symbol to assign each value to only one category: e.g., 0-<10, 10-< 20, 20-30. With open ended classes, the "greater or equal than" symbol can be used: <10, 10-< 20, >20.

A5.112. The legend can also be integrated with a statistical chart that summarizes the data distribution of the variable. "Histograms", in which the bar colours correspond to the shade colour are often used for this purpose (see figure A5.29A). If the class ranges are not constant, the bars of the histogram can be drawn with varying width. If the mapping package does not support histograms, they can be designed in a graphics program or imported from a spreadsheet or statistical package. There are two options for determining the height of each bar. The more conventional approach is to use the number of geographic units whose values fall into each category. Some

Figure A5.29
Legends showing statistical data distribution



desktop-mapping packages display the number of units that fall into each class in the legend. The problem with this is that the units—for example, districts—may be of very different population size. Instead of the number of units, the height of the histogram bars could thus be determined by the size of the underlying population. For a map of population density, for instance, this will be the number of people living in each density range. Of course, the shape of the histogram will be very different and the procedure used should be indicated clearly on the map or the accompanying text.

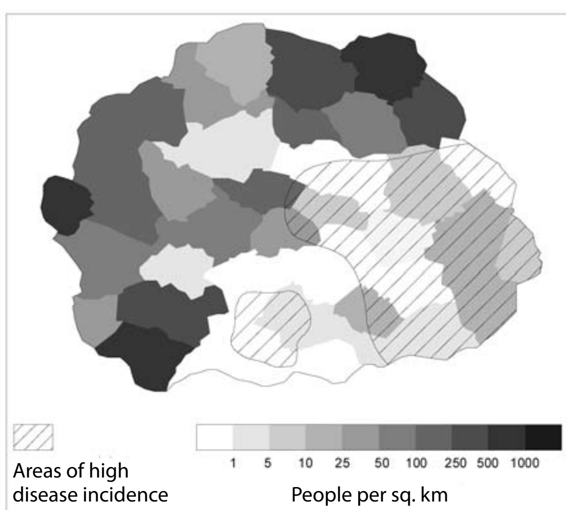
A5.113. Statistical software also allows the computation of density plots showing the data distribution in a more continuous form compared to a histogram (see figure A5.29B). The surface under the density curve sums to one so that the approximate frequency of any individual data value can be read from the graph. Legends of this kind have been used, for example, in the *Atlas of United States Mortality* (United States, National Health Center for Health Statistics, 1997).

F. Maps that tell stories

1. Multivariate maps

A5.114. With few exceptions, the examples provided above have only shown one variable at a time. This is the most common type of display used in a census atlas. For analytical purposes and to illustrate relationships between variables, there may be a need to display more than one variable at a time. In the section on colour choice, it was argued that multivariate maps that use a complex colour scheme to show both variables in the same map tend to be difficult to understand. One alternative, as noted above, is to use a pattern with a transparent background colour on a shaded choropleth map. This works well if the overlaid variable has only a few classes or if it is a binary variable (e.g., presence versus absence) (see figure A5.30).

Figure A5.30
Combinations of solid and hatched shade symbols to display two variables on the same map

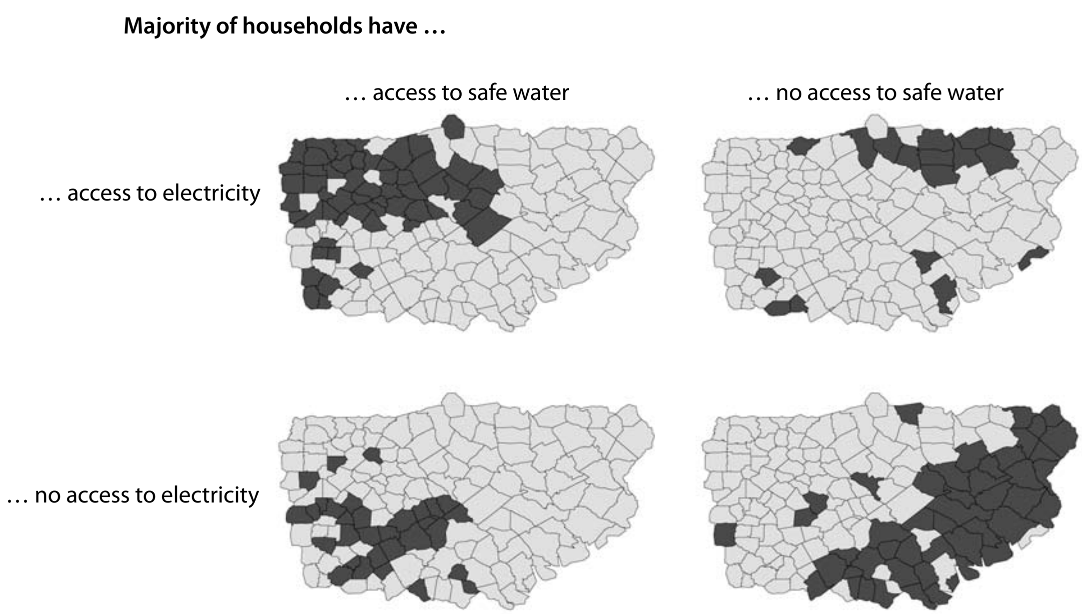


A5.115. In statistical data analysis, two categorical variables that take on only a small number of values are analysed using cross-tabulations. Such tables are also called “contingency tables”. The rows and columns of a two-way table show the categories of the two variables and the cells show the number of observations that take on the corresponding values for each variable. This arrangement allows the quick evaluation of relationships. For instance, we might have converted two variables from a housing census—percentage of households with access to safe water and percentage of households with access to electricity—into two binary variables that indicate whether or not the majority of the district’s households have access to these public services. The cross-tabulation may look something like this:

Majority of households have …			
	… access to safe water	… no access to safe water	Total
… access to electricity	55	17	72
… no access to electricity	31	48	79
Total	86	65	151

A5.116. In order to present this information geographically, a map with four classes could be produced, one for each of the cells in the cross-tabulation. But because the four classes do not have any natural ordering, it will be difficult for the viewer to detect patterns in such a map. A better approach is to translate the concept of a two-way table directly into cartographic language. Figure A5.31 shows a map-equivalent of a two-way table. Each map indicates the districts corresponding to the corresponding cell in the two-way table. The maps do not require a comprehensive legend since the dark shade clearly highlights the districts of interest.

Figure A5.31
The map equivalent of a two-way table



A5.117. Patterns are immediately apparent even on a small map that covers only about one third of the page. Most districts in the north-west have both access to safe water and electricity, while the majority of households in districts in the south-east have neither. In cross-tabulations, the off-diagonal cells are often the most interesting. In some districts in the north-east, most households do not have access to safe water but do have access to electricity. In a cluster of districts in the south-west, the situation is reversed.

A5.118. The approach could be extended to more complex tables, for instance where one variable takes on three values (e.g., low, medium and high) and another has two categories. The maps do not have to be drawn large. Even with many geographic units—in this example 151 districts—small maps are sufficient since only two contrasting colours or gray shades are required.

2. Small multiples

A5.119. Arranging data in several maps can also effectively present dynamic information. Figure A5.32 shows population increase over time based on figures from four successive censuses. The population density maps show where population growth has been highest. To allow comparisons over time, the class limits need to be the same on all maps. This means that classification schemes that are based on the data distribution (e.g., natural breaks) are not appropriate. The density maps are complemented by three smaller maps that show the average annual population growth rates between the censuses.

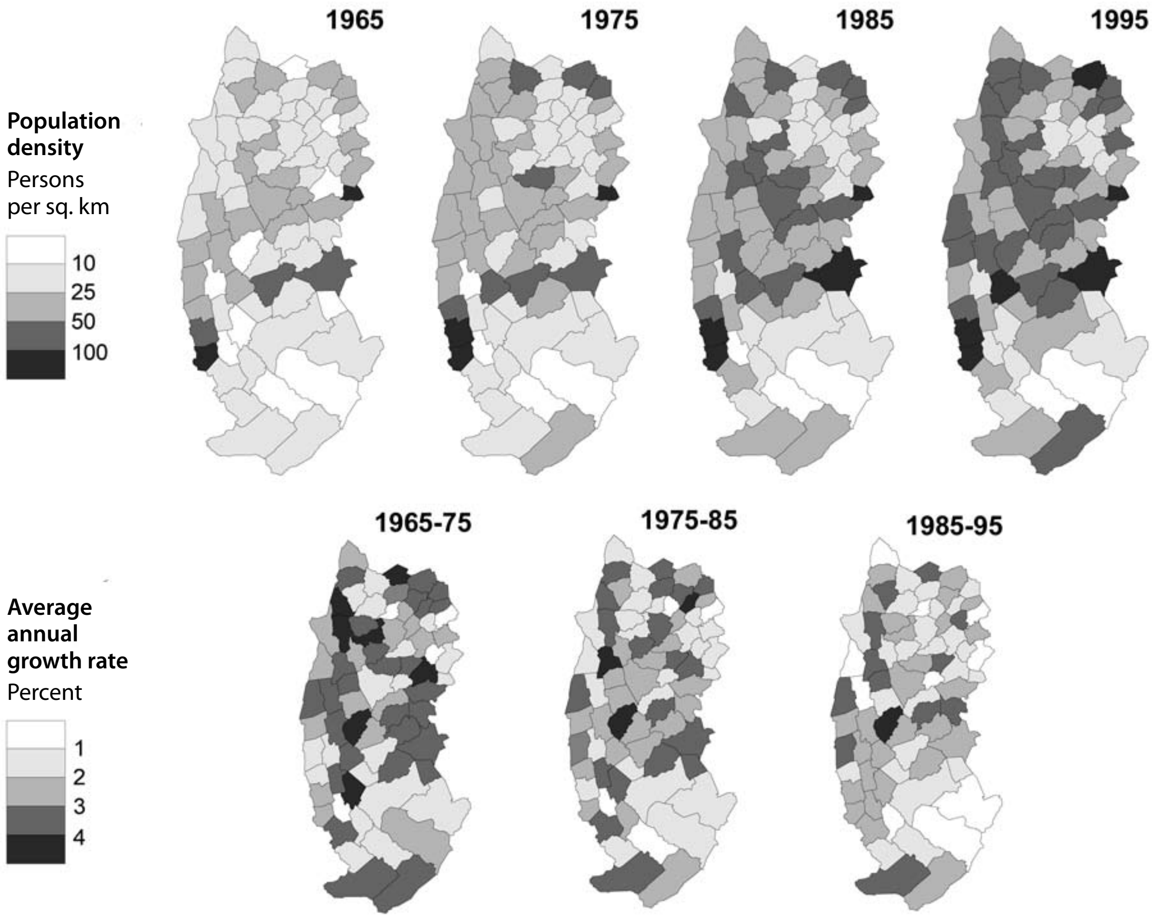
A5.120. Displays such as the one in figure A5.32 are termed "small multiples" (Bertin, 1973; Tufte, 1983). The same map design is repeated for each year or each population subgroup. Because the design is constant across all maps, they can be interpreted by the viewer quite easily. This allows the map designer to present a higher density of information than would otherwise be possible. Multivariate relationships are often clearer in designs that show several maps than in composite maps with a complicated legend design.

A5.121. Another example of a map that uses the concept of small multiples is shown in the United Nations publication *Geographical Information Systems for Population Statistics* (United Nations, 1997; see figure 4.8). It shows sex ratios for five-year age groups for 75 districts in Nepal. The graphic shows seventeen small maps with a diverging classification centred around a balanced sex ratio. The colour version of this map shows a surplus of females in varying shades of reds, and a surplus of males in blues. The black-and-white version uses solid gray shades for a surplus of females and a dot pattern of varying density for a surplus of males. Obviously colour enhances the message of these maps. Despite the large amount of information, the maps can be interpreted quite easily, since clusters of similar values are very obvious. Clearly a table of 1275 (17 times 75) values would be considerably more difficult to interpret than the same information presented geographically. In fact, the graphic for Nepal shows some clear trends which can be attributed to migration over the life cycle of men in a number of districts.

Figure A5.32

Small multiples—depicting changes over time

Population dynamics 1965-1995



Annex VI
Glossary

Accuracy—the degree to which a measurement or representation agrees with the true, real-world values. Determination of an acceptable accuracy requirement and development of an accuracy standard are some of the first steps in a GIS project. Accuracy is not to be confused with precision, which refers to the ability to distinguish between small quantities in measurement. For example, a point location might be measured precisely (e.g., with five significant decimal digits) but inaccurately (e.g., several metres off from it's true real-world position.

Address—a number or similar designation that is assigned to a housing unit, business or any other structure. Addresses mainly serve postal delivery but are also important for administrative purposes, for example in civil registration systems and in census-taking.

Address matching—the process of matching general attribute information to geographical locations on a street network using a street address. For example, a tabular address register can be matched to a comprehensive digital street map to produce a GIS point layer showing the location of each household. This is sometimes also called **geocoding**.

Administrative unit—a geographic area that serves administrative and governmental functions. It is usually defined and established by legal action.

Aerial photography —techniques for taking photographs from an aerial platform, usually a low-flying aircraft. Also sometimes called vertical photography or orthophotography. Air photos are used for photogrammetric mapping, allowing a high degree of accuracy.

Aerial survey—a cartographic survey by means of aerial photography or other remote sensing technology.

American Standard Code for Information Interchange (ASCII)—a computer code developed to facilitate the interchange of alphanumeric data and special characters between computers and across operating systems. Each character is assigned a one-byte code, i.e., a value between 0 and 255.

Annotation—text that is used to label features on a map. Annotation can be stored in a GIS and drawn onto maps for display or printing. In contrast to text information in an attribute table, annotation is only used for cartographic display and not for analysis.

Arc—see **line**.

Arc second—one second of latitude or longitude, or 1/3600th of a degree.

Area—a bounded, two-dimensional extent of the Earth's surface that is represented in a GIS as a polygon.

Areal interpolation—the transfer of an attribute from one set of reporting zones to another, incompatible, set of zones; for instance, the estimation of population totals for ecological regions based on a GIS data set of population by district.

Areal unit—a natural or artificial area that is often used to compile and report aggregate data. *Examples*: land cover zones or **enumeration areas (EAs)**.

Attribute—a characteristic of a geographic feature. For example, a numeric or text field that is stored in a relational database table which can be linked to the geographic objects in a GIS. Attributes of an enumeration area, for example, could be its unique identifier, the area in km^2, total population and number of households. A distinction is sometimes made between geographic and general attributes. The former are stored in a data table that is tightly linked to the geographic coordinate files and contains fields, such as the internal identifiers, feature codes, area etc. General attributes are typically stored in separate data tables that can be linked to the geographic attribute table.

Automated Mapping/Facilities Management (AM/FM)—GIS applications in the utility and public works sector that focus on engineering and maintenance issues.

Band—a layer of a multispectral remote sensing image that shows the signals measured in a defined range of the electromagnetic spectrum. See also **multispectral image**.

Bandwidth—the amount or volume of digital data that can be transferred through a communications connection.

Base data—see **framework data**.

Base map—a map that shows fundamental geographic features that can be used for locational reference. Sample features are roads, administrative boundaries and settlements. Base maps are used to compile new geographic data or for reference in the display of thematic map information.

Base station—a GPS receiver, whose location has been precisely and accurately determined, that broadcasts and/or collects differential correction information for mobile GPS receivers. See also **differential GPS (DGPS)**.

Beidou—an alternative to the GPS system proposed by China, which will consist of 35 satellites and service with an accuracy of up to 10m.

Binary—made up of or referring to two, as in binary variables (e.g., yes/no). Also, a form of computer encoding that is based on individual pieces of information called bits that can take on two values—i.e., 0 and 1.

Bit—a binary digit that can assume a value of 0 or 1.

Boundary—a line which defines the extent of an areal unit or the locations where two areas meet. A boundary is represented in a GIS as a line feature which may define a side of a polygon. The boundary may or may not be visible on the ground, i.e., it can follow real-world features, such as roads and rivers or it can be defined solely by geographic coordinates.

Bits per second (BPS)—a measure of transfer speed in digital communication networks.

Buffer—a zone or area of a specified distance around a geographical feature (points, lines or polygons). Buffer operations are one of the fundamental geospatial capabilities.

Byte—a group of eight binary digits or bits that can be processed as a unit by computer programs. A kilobyte consists of approximately one thousand bytes a megabyte of one million bytes and a gigabyte of one billion bytes.

Cadastral information—records that describe the past, present and future rights and interests in land ownership for legal and tax purposes. Cadastral maps show the geographic location and extent of land parcels. Cadastral surveys in many countries now use GIS to store this information. Also called land titling information.

Cartesian coordinate system—a system of lines that intersect at perpendicular angles in two-dimensional space. This system provides the framework to precisely reference locations as x/y coordinates.

Cartogram—a map that is constructed by scaling the reporting units according to the value of a variable recorded for them. Also called value-by-area mapping.

Cartographic generalization—the process of abstracting real world features through a reduction of detail for representation on a map. This involves selection, classification, simplification and symbolization.

Cartography—the art and science of creating a two-dimensional representation of some part of the Earth's surface. Features represented may be real objects (**topographic map**) or they may represent concepts and more abstract characteristics (**thematic map**).

Census geographic framework—the geographic collection and reporting units used by a census office in census enumeration and data tabulation. This includes the hierarchical structure of census and administrative units, their designations and codes and the relationships between different units.

Central meridian—the longitude that defines the origin of the x-coordinate of a cartographic projection.

Centroid—the mathematical centre of a polygon. For irregularly shaped polygons, this can be thought of as a "centre of gravity".

Chain—see **line**.

Channel—the part of a GPS receiver's electronics that captures the satellite's signal. Multi-channel receivers can capture and process signals from several satellites at the same time.

Chart—a map that is primarily designed for sea and air navigation, for example nautical or aeronautical charts.

Choropleth map—a statistical map in which values recorded for reporting units are first assigned to a number of discrete class ranges or categories. The reporting units are then shaded using symbols (colours or patterns) chosen for each category.

Classification—assigning objects into groups that share the same or similar characteristics. In cartography, the process of assigning symbols to map features that are similar or that have similar values. Classification is used to simplify a map in order to improve communication of the cartographer's message.

Clearinghouse—in the context of national spatial data infrastructures, a repository for accumulating and disseminating GIS data and metadata.

Client—a computer that uses data or software stored on another, often remote, computer (server).

Code—the alphanumeric characters used to identify geographic objects. Codes are also used to identify attribute categories, such as population density ranges, land use classes or industries. See also **geographic code**.

Colour model—a procedure for representing colours numerically in a computer. For example, in the RGB colour model, colours are represented as numeric levels of red, green and blue. Pure red, for instance, is defined as 255,0,0. Other examples of colour models are the hue, lightness, saturation (HLS) and cyan, magenta, yellow (CMY) models.

Colour separation—the process of dividing a graphical document into separate pages or files for each of four colours (cyan, magenta, yellow and black). Colour separation is the basis of most professional printing processes.

Column—in GIS, a group of cells or pixels in a grid or raster GIS database that are aligned vertically. In database management systems, a field or item in an attribute table.

Computer Graphics Metafile (CGM)—a standard file format for exchanging image or vector data.

Computer-Aided Design/Computer-Aided Design and Drafting (CAD/CADD)—a software system that provides the tools for drafting and design, specifically in engineering or architectural applications. CAD systems use graphical coordinate system and are therefore similar to geographic information systems.

Conformal projection—a cartographic projection in which all angles are preserved correctly at each point.

Connectivity—in topological GIS, when two or more lines are joined at a single point or node.

Contiguity—two or more geographical features that are neighbours or adjacent are said to be contiguous or to have "contiguity".

Continuous geographical phenomena—geographic variables that vary without clearly distinguishable breaks or interruptions, for example temperature or atmospheric pressure—as opposed to discrete geographical phenomena.

Contour—a line on a map that connects points of equal elevation. See also **isoline**.

Control—see **geodetic control**.

Control point—a point on a map, an aerial photo or in a digital database for which the x,y coordinates and possibly elevation are known. Used to geographically register map features.

Control segment—a global network of GPS monitoring and control stations that ensure the accuracy of the satellite signals.

Coordinate—two or three numbers that describe the position of a point in two or three dimensions (e.g., x/y or x/y/z, where z indicates height). A two-dimensional coordinate is sometimes called a coordinate pair, a three-dimensional coordinate a coordinate triplet. In GIS databases, coordinates represent corresponding locations on the Earth's surface relative to other locations.

Coordinate Geometry (COGO)—term used by land surveyors for dealing with precise measurements of locations.

Coordinate system—the reference system that is used to specify positions on a map or in a GIS database. A cartographic coordinate system is defined by a map projection, a reference ellipsoid, a central meridian, one or more standard parallels, and possible shifts of x and y coordinate values.

Coverage—in GIS, coverage sometimes refers to a vector GIS data set that contains geographic features belonging to a single theme, such as census units or roads.

Data capture—conversion of geographic coordinate data from hard-copy sources or by means of field measurements into a computer-readable format. Data capture usually involves digitizing or scanning paper maps or air photos.

Data conversion—the transfer of data from one format into another. Usually data conversion refers to the translation of paper map information into digital form. In a wider sense, geographic data conversion also includes the transfer of digital information from one GIS file format into another.

Data dictionary—a data catalogue that describes the contents of a database. Information is listed about each field in the attribute tables and about the format, definitions and structures of the attribute tables. A data dictionary is an essential component of metadata information.

Data format—usually refers to a specific, possibly proprietary set of data structures within a software system.

Data model—a user's conceptual design of a data set that describes database entities and their relations to each other.

Data sets—a logical collection of values or database objects relating to a single subject.

Data standardization—the process of reaching agreement on common data definitions, formats, representation and structures of all data layers and elements.

Data structure—implementation of a data model consisting of file structures used to represent various features.

Data type—the field characteristic of the columns in an attribute table. For example, character, floating point and integer.

Database—a logical collection of information that is interrelated and which is managed and stored as a unit, for example in the same computer file. The terms database and data set are often used interchangeably. A GIS database contains information about the location of real-world features and the characteristics of those features.

Database Management System (DBMS)—a software package designed for managing and manipulating tabular data. A DBMS is used for the input, storage, manipulation, retrieval and query of data. Most GISs use a relational DBMS to manage attribute data.

Datum—in cartography, a set of parameters that define a coordinate system. More specifically, a datum is a reference or basis for measurements or calculations. For example, a national cartographic datum establishes the reference framework for cartographic activities in a country.

Differential GPS (DGPS)—the set of techniques used to improve the accuracy of coordinates captured with a GPS by calculating the signal error (offset) for a second GPS receiver (the base station) at a location which has been precisely and accurately determined. The correction factor is applied to the coordinates captured by the mobile unit, either in real time or in post-processing mode (i.e., using a database of time referenced correction information). In some parts of the world, differential correction information is broadcast continuously from a set of permanent base stations.

Digital Elevation Model (DEM)—a digital representation of elevation information for a part of the Earth's surface. A DEM is usually a raster data set in which elevation values are stored for cells in a fine grid, but vector formats can also be used to store elevation. A DEM is s also called a digital terrain model (DTM).

Digital orthophoto—a digital image or aerial photograph, usually of very high resolution, which has been geometrically corrected. A digital orthophoto, also called orthoimage, combines the detail of an aerial photograph with the geometric accuracy of a topographic map.

Digital Terrain Model (DTM)—see Digital Elevation Model (DEM).

Digitizing table—a computer peripheral used to capture coordinate data from paper maps or similar cartographic materials. Also called a digitizer.

Digitizing—the process of translating geographic feature information on paper maps into digital coordinates. Digitizing usually refers to the manual process of tracing lines on a paper map attached to a digitizing table with a mouse-like cursor that captures coordinates and stores them in a GIS database.

Discrete geographical features—individual entities that can be easily distinguished, such as houses or roads—as opposed to continuous geographical phenomena.

Dissolve—a GIS function that deletes boundaries between adjacent polygons that have the same value for a specific attribute. For example, enumeration area polygons can be dissolved based on the code of their supervisory units to create supervisory maps.

Dot map—a map in which quantities or densities are represented by dots. Usually, each dot represents a defined number of discrete objects, such as people or cattle. The dots can be placed randomly in the reporting units or they can be placed to reflect the underlying true distribution of the variable.

Drawing Exchange Format (DXF)—an ASCII format for describing a graphic or drawing developed by Autodesk, Inc., of Sausalito, California. Initially developed for CAD applications, it has also become a standard for GIS data exchange.

Edgematch—a manual or automated editing technique in a GIS that matches shared features that were digitized from adjacent map sheets. Edgematching may be necessary, for instance, to connect roads or administrative unit boundaries after joining maps that were digitized separately.

Ellipsoid—in cartography, the three dimensional shape used to represent the Earth. The Earth ellipsoid is characterized by a smaller distance from the centre to the poles (semi-minor axis) than that from the centre to the equator (semi-major axis). Also called a spheroid.

Entity—a real-world phenomenon of a given type. In database management systems, the collection of objects (e.g., persons or places) that share the same attributes. Entities are defined during conceptual database design.

Entity-relationship model—a data model that defines entities and the relationships between them, for example, the relationships between enumeration areas and supervisory regions.

Enumeration area (EA)—usually the smallest geographic unit for which census information is aggregated, compiled and disseminated. An enumeration area is defined by boundaries described on a sketch map or in a GIS database. These boundaries may or may not be visible on the ground. Also called census block or census tract.

Equal area projection—a cartographic projection in which all regions are shown in correct proportion to their real-world areas.

Equator—in cartography, the reference parallel, i.e., latitude 0° north and south.

Equidistant projection—a cartographic projection which maintains the scale along one or more lines, or from one or two points to all other points on the map.

Feature—a geographic object displayed on a map or stored in a GIS database. Features can be natural or man-made real-world objects (a river or a settlement) or they can be conceptual or defined features (e.g., administrative boundaries).

Field—a column in a database table.

File Transfer Protocol (FTP)—a standard set of conventions for exchanging computer data files in digital communication systems, such as the Internet.

Flow map—a map in which movements, for example of goods or people, along a linear path are shown.

Foreign key—in relational database management systems, a field or item in a table that contains a value identifying rows in another table. It is used in joining two tables by defining the relationship between two elements of a relational database. A foreign key is the primary key in the other table.

Framework data—in the context of national GIS activities, a set of general purpose geographic themes or base data, such as administrative boundaries, elevation or transportation infrastructure. Framework or national spatial data infrastructure initiatives aim at coordinating the development and standardization of GIS data sets of framework data in a country.

Galileo—An alternative to the GPS system to be built by the European Union. It will consist of 30 satellites and two ground stations, and be compatible with the United States GPS system at the user level.

Gazetteer—a list of place names and their geographic location (usually latitude/longitude).

Generalization—see **cartographic generalization**.

Geocoding—(a) the process of assigning geographic codes to features in a digital database; (b) a GIS function that determines a point location based on a street address. See also **address matching**.

Geodetic control—a network of precisely and accurately measured control or reference markers that are used as the basis for obtaining new positional measurements. Also called benchmark points.

Geographic attributes file—a database table that is tightly linked to the spatial objects stored in a GIS coordinate file. The geographic attribute file or table contains specific information on each feature, such as its identifier, name and surface area. In some systems, this file is also called point, line or polygon attribute table. Data stored in external tables can be linked through a relational database operation.

Geographic code—a unique alphanumeric identifier that is assigned to a legal, administrative, statistical or reporting unit.

Geographic database—a logical collection of data pertaining to features that relate to locations on the Earth's surface.

Geographic hierarchy—in the context of census-mapping, a system of usually nested area units that are designed for administrative or data-collection purposes. For instance, a country is divided into provinces, which are divided into districts and so on to the lowest level, which may be the enumeration area. See also **census geographic framework**.

Geographic information system (GIS)—a collection of computer hardware, software, geographic data and personnel assembled to capture, store, retrieve, update, manipulate, analyse and display geographically referenced information.

Geographic object—a user-defined geographic feature or phenomenon that can be represented in a geographic database. Examples include streets, land parcels, wells and lakes.

Geographic reference file—a digital, tabular master file that lists the names, geographic codes and possibly attributes of all geographic entities that are relevant to census and survey data collection.

Geographically coincident—describes two or more geographic features that share the same location or boundary. For instance, some reporting or statistical units may also be administrative units.

Georeferencing—the process of determining the relationship between page coordinates and real-world coordinates. Georeferencing is necessary after digitizing, for example, to convert the page coordinates measured in digitizing units (e.g., centimetres or inches) into the real-world coordinate system that was used to draw the source map. See also **transformation**.

Geospatial—a term that is sometimes used to describe information of a geographic or spatial nature.

Geostationary satellite—an Earth satellite that remains in a fixed position above a point on the Earth's surface. Also called a geosynchronous orbit.

Geo-TIFF—see tagged image file format.

Global Navigation Satellite System (GLONASS)—the counterpart of the United States GPS system operated by the Ministry of Defence of the Russian Federation. The system is very similar to the GPS system but is not subject to selective availability. Some receivers that combine GPS and GLONASS signals to improve coordinate accuracy are available.

Global Positioning System (GPS)—a system of 24 satellites orbiting the Earth that broadcast signals which can be used to determine the exact geographic position on the Earth's surface. GPS is used extensively in field-mapping, surveying and navigation. GPS is maintained by the United States Department of Defense. See also **Differential GPS (DGPS), Beidou, Galileo and GLONASS**.

Governmental unit—see **administrative unit**.

Graduated symbols—in thematic cartography, the use of symbols (e.g., circles or squares) to represent the magnitude of a variable at a point or in a reporting unit. The size of the symbol is proportional to the value of the variable.

Graphic Interchange File (GIF)—a graphics image file format developed initially for transmission of images through electronic bulletin boards. The GIF format, which allows efficient compression of file size, is used for most graphics on web pages.

Graticule—in cartography, the grid of longitudes and latitudes drawn on a map.

Great circle—the circle that is formed by intersecting a plane through the centre of a sphere. For example, all meridians and the equator are great circles. On the sphere, the shortest path between two points is along the great circle that passes through both points.

Greenwich meridian—the longitude of reference, i.e., 0° east or west. It passes through the English town of Greenwich, a suburb of London.

Grid—a geographic data model that represents information as an array of uniform square cells. Each grid cell has a numeric value that refers to the actual value of a geographic phenomenon at that location (e.g., population density or temperature) or it indicates a class or category (e.g., the enumeration area identifier or soil type). See also **raster**.

Ground truth—information collected in a field survey to verify or calibrate information extracted from remote sensing data.

Heads-up digitizing—a digitizing technique that does not employ a digitizing table. Instead, features are traced with a mouse on-screen either from a scanned image displayed in the background or following features drawn on a clear medium (e.g., mylar) that is attached to the computer screen.

Hydrography—features pertaining to surface water, such as lakes, rivers, canals etc.

Hypsography—features pertaining to relief or elevation of terrain.

Image—a representation of a part of the Earth's surface. However, an image is usually produced using an optical or electronic sensing device. For instance, scanned aerial photographs or remote sensing data are usually referred to as images. In terms of data storage and processing, an image is very similar to a raster or grid.

Infrastructure—the system of public works in a country, state or region, including roads, utility lines and public buildings.

Integration—in GIS, the process of compiling a consistent set of spatial data from heterogeneous sources. Vertical integration refers to the ability of GIS to combine different data layers that are referenced in the same coordinate system.

Internet—a global system of linked computer networks that allows data-communication services, such as remote log-in, file transfer, electronic mail, bulletin boards and newsgroups. The Internet is also the foundation for the world wide web (www).

Internet Protocol (IP)—the most important set of codes and conventions which enable the transfer of digital data on the Internet.

Interpolation—the process of estimating a variable value at a location based on measured values at neighbouring locations. Used to produce a complete grid data set from point sample information, for instance, a precipitation surface from rainfall stations.

Intersecting—a GIS function that is used to topologically integrate or combine two spatial data layers so that only those features that are located within the area common to both are preserved.

Isoline—lines on a so-called isarithmic map that connect points of constant value. The best known example is an isohypse, which shows lines of equal elevation (also called an elevation contour map).

Java—a programming language that allows the creation of software packages that can run on multiple platforms (i.e., operating systems). Java programs, called applets, can be sent or retrieved through the Internet to be executed on a remote computer.

Join—in relational database management systems, the process of attaching values from a database table to another table, based on linking a foreign key to its primary instance in the external table.

Joint Photographic Experts Group (JPEG)—a graphics file format used primarily for photographic images that allows significant file-size compression.

Land Information System (LIS)—a term sometimes used for a GIS application that contains information about a specific region, including cadastral information, land use, land cover etc.

Latitude—the "y-coordinate" in a polar coordinate system on a sphere. Measured as the angular distance in degrees north or south of the equator. Also called parallel.

Layer—an individual GIS data set that contains features belonging to the same theme, such as roads or houses. The term layer refers to a GIS's ability to overlay and combine different thematic layers that are referenced in the same coordinate system. Also called **coverage**.

Legend—in cartography, the information on a map that explains which symbols are used for the features and variables that are represented on the map. This includes the symbol key required to interpret the map, for example, the shade colours and corresponding value ranges of a population density map.

Line—a one-dimensional object. A geographic data type consisting of a series of x,y coordinates, where the first and last coordinates are called nodes and the intermediate coordinates are termed vertices. Also referred to as an **arc** or a **chain**. The part of a line between two intersections with other lines is called a line or arc segment.

Line-in-polygon—a GIS operation in which line features are combined with polygon features to determine which lines fall into which polygons. Using this operation, polygon attributes can be added to each corresponding record in the line attribute table (e.g., the district into which the road falls), or line attributes can be summarized for each corresponding polygon (e.g., total road length in a district).

Local Area Network (LAN)—a computer network that connects computers over relatively short distances, for example within the same office building.

Logical accuracy—a term used for the degree by which relationships among geographic features on a map or in a GIS database are represented correctly (e.g., adjacent to, connected to). A GIS database can be logically accurate even if its positional accuracy is limited.

Longitude—the "x-coordinate" in a polar coordinate system on a sphere. Measured as the angular distance in degrees east or west of the Greenwich meridian.

Map—a representation of some part of the Earth's surface drawn on a flat surface (e.g., paper or a computer display).

Map compilation—the process of assembling, evaluating and interpreting cartographic measurements and materials in order to produce a new map.

Map composition—the arrangement of map elements to create a cartographic product that is visually appealing and correctly represents the phenomena that are represented.

Map elements—Components of a thematic or topographic map, such as title, legend, scale, north arrow, graticule, borders and neat-lines.

Map extent—the coordinates in map units that define the rectangle that encloses all features contained in a specific map display or a GIS database; i.e., the minimum and maximum x and y coordinates in a digital database or the part of a database shown in a map display.

Map projection—a mathematical procedure for converting locations on the Earth's surface into a planar coordinate system. Depending on the mathematical formulae employed, map projections have different properties. Some preserve the shape of regions on the globe, others preserve relative area, angles or distances.

Map units—the units of measurement in which coordinates in a GIS database are stored; e.g., centimetres, metres or degrees, minutes and seconds.

Meridian—a reference line that is defined by the corresponding longitude. For example, the Greenwich meridian.

Metadata—data about data. A collection of information that describes the content, quality, condition, format, lineage and any other relevant characteristic of a data set.

Minimum mapping unit—generally the size of the smallest feature that will be included on a map. Also, at a given map scale, this is the size or the dimension at which a small, compact polygon feature is represented as a point or a long narrow polygon feature is shown as a line. For example, a town is shown as a polygon if its size is larger than 3 mm on a page, but as a point if it is smaller.

Multipath—the error introduced to GPS readings as a result of reflection and scattering of GPS signals on neighbouring structures, such as houses or trees. Multipath error is a problem mostly in high-precision surveying.

Multispectral image—a remotely sensed data set that consists of a number of bands or layers. These are essentially separate images taken at the same time for the same area, each of which shows the signal of a different range of the electromagnetic spectrum.

Nadir—in aerial photography and remote sensing, the point on the Earth's surface that is located directly below a camera or sensor.

Network analysis—procedures to analyse relationships between points or addresses on a set of lines in a GIS database that may represent, for example, a street network. Network analysis is used for location decisions and routing, such as emergency management.

Node—the start-point or end-point of a line feature, or the point at which two or more lines connect.

Normalization—the conceptual procedure in database design that removes redundancy in a complex database by establishing dependencies and relationships between database entities. Normalization reduces storage requirements and avoids database inconsistencies.

Orthophoto—see **digital orthophoto**.

Overlay—the combination of two data layers that are in the same geographic reference system. Overlay can be done for cartographic display purposes, or the two layers can be physically combined to create a new GIS data set (e.g., polygon overlay, point-in-polygon, line-in-polygon).

Overshoot—in digitizing, a line that has been extended beyond the point where it should connect with another line. The resulting spurious line segment is sometimes called a dangle.

Panchromatic image—a remotely sensed image that records the signal in a broad range of the electromagnetic spectrum, similar to a black-and-white photograph.

Parcel—a single cadastral unit or land property.

Photogrammetry—the art and science of extracting measurements and other information from photographs. In the context of mapping, the procedures for gathering information about real-world features from aerial photographs or satellite images.

Pixel—from picture element. Similar to a cell in an image, grid or raster map.

Planar coordinate system—a system for determining location in which two groups of straight lines intersect at right angles and have as a point of origin a selected perpendicular intersection. See **Cartesian coordinate system**.

Planimetric map—a map that, in contrast to a topographic map, only shows the locations of features but not their elevation. A planimetric map may show the same features as a topographic map, with the exception of terrain or elevation contours, but will usually only show selected features chosen for a specific purpose.

Plotter—a computer peripheral that can draw a graphic file, similar to a printer, but usually for larger format output.

Point—a zero-dimensional object. An x,y coordinate that is used in a digital geographic database to represent features that are too small to be shown as lines or polygons. For example, households, wells or buildings are often shown as points.

Point-in-polygon—a GIS operation in which point features are combined with polygons to determine which points fall into which polygon. Using this operation, polygon attributes can be added to each corresponding record in the point attribute table (e.g., health service area information for a survey sample point) or point attributes can be summarized for each corresponding polygon (e.g., number of hospitals in each district).

Polygon—a two-dimensional object. An area feature that is represented in a vector GIS as a sequential series of x/y coordinates. These define the lines that enclose the area; i.e., the first and last coordinate of the polygon are identical.

Polygon overlay—a GIS operation in which two polygon data layers are combined to create a new data layer. The output layer consists of the areas of intersection of both sets of input polygons. The attribute table of the new data layer contains the attributes from both input data sets. Polygon overlay is one of the fundamental GIS operations that is often used to integrate information from heterogeneous sources, such as demographic and environmental data.

Positional accuracy—a term used for the degree by which positions on a map or in a GIS database are recorded correctly with respect to their true location on the Earth's surface. Logical accuracy, in contrast, only pertains to correct representation of the relationships among geographic features.

Postscript—a flexible, high-resolution page description language that is mostly used to send graphical information, such as GIS-produced maps, to printers. Encapsulated postscript format (EPS) includes a small bit-map representation of the graphic for previews.

Precision—the ability to distinguish between small differences in measurement. In GIS, coordinate precision is determined by the data type used to store the x and y coordinates (usually double precision, or 16 bytes for each number).

Primary key—One or more fields in an attribute table that uniquely identify a specific instance, row or record.

Protocol—a set of conventions that determine the treatment, exchange and formatting of data in an electronic communications system. Similar to a data standard but applied to procedures.

Quadrangle—a rectangular area that is bounded by pairs of meridians and parallels.

Quality control—the steps and procedures in a database development project or cartographic production system that ensure that the resulting data or output comply with specified standards of accuracy and usability.

Quantile—a statistical or cartographic classification method that assigns an equal number of objects into a fixed number of classes. Four-class systems are called quartiles, five-class systems are called quintiles, ten-class systems are called percentiles. For example, the first of the four quartiles of a data distribution would contain the 25 per cent of observations with the lowest values.

Radius—the distance from the centre of a circle to its outer edge.

Raster—a geographic data model that represents information as a regular array of rows and columns, similar to a grid or image. Raster cells are usually, but not always, square. Area or line features are represented as groups of adjacent raster cells with the same value.

Rectification—the process by which an image or grid is converted from image coordinates to real-world coordinates. This usually involves rotation and scaling of grid cells, and thus requires resampling or interpolation of grid values. Similar to **transformation** of vector data.

Registration—the process of matching features in two maps or GIS data layers so that corresponding objects are coincident. Registration is based on a series of ground control points, and is related to **transformation** and **rubber-sheeting**.

Reference map—in the context of census-mapping, a cartographic product (hard-copy or digital) that displays some portion of the census geographic framework, e.g., an data collection or statistical dissemination unit.

Relational Database Management System (RDBMS)—a database management system that allows the temporary or permanent joining of data tables based on a common field (a primary and foreign key). Each row, record or instance in a database has a fixed set of attributes or fields. Each table has a primary key that uniquely identifies each record. The table may also contain a foreign key which is identical to a primary key in an external table. A relational join is achieved by matching the values of the foreign key to the corresponding values in the primary key of the external table.

Remote sensing—the process of acquiring information about an object from the distance; i.e., without physical contact. Remote sensing usually refers to image acquisition by means of satellite sensors or aerial photography.

Resolution—a measure of the smallest detail that can be distinguished on a map or in a digital database. Resolution determines the accuracy at which the location and shape of a map feature can be accurately represented at the given map scale. In raster GIS and image data, resolution is sometimes used to refer to the cell or pixel size.

Row—in GIS, a group of cells or pixels in a grid or raster GIS database that are aligned horizontally. In database management systems, a record or instance in an attribute table.

Rubber-sheeting—a procedure in which the shape and location of objects in a GIS database are modified in a non-uniform manner. Rubber-sheeting is often used to bring a GIS data set in an unknown coordinate system into a known system. The adjustments are defined by specifying a large number of links from locations in the input data set to their corresponding correct reference or control points in the output coordinate system.

Run-length encoding—a compression technique for raster, grid or image data. Instead of storing each value of adjacent cells that have the same value, the system stores the value and the number of times the value is repeated. Compression will be significant when discrete objects are stored in a raster GIS.

Satellite image—a digital data set that has been recorded from an Earth-orbiting satellite, either photographically or by a scanner on-board the satellite. A satellite image in a GIS is similar to a raster or grid data set.

Scale—in cartography, the relationship between the distance on a map and the corresponding distance on the Earth's surface. Scale is reported as a ratio, for example, 1:100,000, which means that 1 centimetre on the map equals 100,000 centimetres on the Earth's surface. Since scale is a ratio, a "small scale" map shows a relatively large area, while a "large scale" map shows a small area. More generally, scale refers to the level of observation or enquiry, which may range from micro-scale to macro-scale phenomena.

Scanning—a data-capture technique in which information on hard-copy documents (e.g., paper or mylar) is captured and converted into a digital image by means of a light-sensitive optical device. For map data, scanning is an alternative to data input by digitizing. After scanning a map, the image data are usually converted to vector format, using raster-to-vector conversion software or on-screen tracing of line and point features.

Schematic map—see **sketch map**.

Selective availability—the degradation of accuracy of GPS satellite signals that was retained by the United States Department of Defense until 2000 but could be reinstated at a time of war.

Server—a computer that has been set up to provide certain services to other computers (clients), for instance, a web server is a central repository of data, software or content for the worldwide web.

Sketch map—a map (often hand-drawn) that shows main features of a given area but which may not have a high degree of positional accuracy and may thus not correctly represent distances and dimensions of objects. A sketch map may, however, have a high degree of logical accuracy, meaning that relationships between objects are correctly represented. Also called a schematic map or a cartoon map.

Source material—data and information of any type that is used to compile a map or a GIS database. This may include field observations, aerial and terrestrial photographs, satellite images, sketches, thematic, topographic, hydrographic, hypsographic maps, sketch maps and drawings, tabular information and written reports that relate to natural and human-made geographic features.

Space segment—the part of the GPS system that is located in space, i.e., the 24 GPS satellites.

Spatial analysis—the set of techniques for extracting useful information from geographically referenced data. Spatial analysis includes the integration of geographic data sets, qualitative and quantitative methods for evaluating the data, and modelling, interpretation and prediction. In GIS, spatial analysis often refers to the methods of GIS data integration, such as polygon overlay or neighbourhood analysis. In a wider sense, it includes, for instance, spatial process models (e.g., migration dynamics) and spatial statistics (e.g., regression models that account for the spatial arrangements and relationships among observations).

Spatial data—information about the location, dimensions and shape of, and the relationships among, geographic features. In GIS, spatial data are technically classified as points, lines, areas and raster grids.

Spatial data infrastructure—see **framework data**.

Spatial Data Transfer Standard (SDTS)—a data and metadata standard for the exchange of GIS data sets among data producers and users, and between software systems and file formats. Many national and international standards have been implemented or suggested.

Spatial index—a look-up table or structure within a geographic database that is used by a GIS or database management system to speed up queries, analytical operations and display of spatial features.

Spatial interaction—interdependence among geographic entities. It often refers to the flow of goods, services, information or people between geographic locations. Spatial interaction analysis is important in the study of human migration.

Sphere—a globular body similar to a ball. The Earth in its simplest approximation is a sphere, but in reality is more accurately represented as a spheroid (see **ellipsoid**).

Standard parallel—the latitude that defines the origin of the y-coordinate of a cartographic projection.

Standards—in computing, a set of rules or specifications established by some authority that define, for example, accuracy requirements, data-exchange formats, hardware or software systems.

Structured Query Language (SQL)—in relational database management systems, a standard syntax used to define, manipulate and extract data.

Surface—A term often used to describe GIS raster or image data that describe continuous, smoothly varying phenomena, such as elevation or temperature. Even population density is sometimes represented as a raster surface.

Symbols—in cartography, the design elements used to represent map features. Symbol types are points, lines and polygons of a certain shape. Symbolization involves the choice of graphic variables, such as shape, size, colour, pattern and texture.

Table—in database management systems, the set of data elements arranged in rows (records or instances) and columns (fields or items). The number of columns is usually fixed by the definition of the table structure, while the number of rows is flexible.

Tag Image File Format (TIFF)—a standard image or raster file format that can store black-and-white, gray-scale or colour images in compressed or uncompressed form. Scanners and other devices that create image data often provide output in TIFF format. In GIS, the Geo-TIFF format is defined as a standard TIFF image file that describes a remote-sensing image, digital orthophoto or raster GIS data set. It includes an associated file with a .tfw extension that contains information about the image's geographic reference information, cell size in real-world units and other relevant information.

Template—In cartography, a standardized design of peripheral map elements (borders, neat-lines, north arrows) that can be used for a standardized map series. In database management systems, an empty table created for multiple purposes for which only the fields or items have been defined.

Thematic layer—see **layer**.

Thematic map—a map that presents a specific concept, subject or topic. A thematic map can show quantitative or qualitative information.

Theme—in GIS, a set of geographic objects that usually belong to the same subject group (e.g., roads or settlements) and that are stored in the same GIS database.

Topologically Integrated Geographic Encoding and Referencing (TIGER)—a data format developed by the United States Census Bureau to support census programs and surveys. TIGER files are GIS data sets in an internal format that contain street address ranges along road network lines and census tracts, and census block boundaries. The TIGER system was one of the first efforts to create a complete digital census GIS database for a country.

Tile—in GIS, a term sometimes used to refer to adjacent digital map sheets that are stored in separate files. Tiles can be of regular shape (e.g., square or rectangular) or they can follow irregular boundaries, such as district or province borders. Storing all tiles in the same geographic reference system allows temporary or permanent joining of adjacent tiles.

Topographic map—a map of mostly real-world features, including elevation contours, rivers, roads, settlements, and landmarks. The standard map sheets created by a national mapping agencies at various scales are typically topographic maps.

Topology—in GIS, a term that refers to the spatial relationships among geographic features (e.g., points, lines, nodes and polygons). A topologically structured database stores not only individual features but also how those features relate to other features of the same or different feature class. For example, in addition to a set of lines representing a road network, the system will store the nodes that define road intersections, which allows the system to determine routes along several road segments. Or, instead of storing polygons as closed loops, where the boundaries between neighbouring polygons would be stored twice, a topologically structured GIS would store each line only once, together with information on which polygon is located to the left and the right of the line. This avoids redundancy and facilitates the implementation of many GIS and spatial analysis functions.

Transformation—the conversion of digital spatial data from one coordinate system to another through translation, rotation and scaling. Transformation is used to convert digitized digital map data from digitizer units (e.g., centimetres or inches) into the real-world units corresponding to the source map's map projection and coordinate system (e.g., metres or feet). See also **georeferencing**.

Transmission Control Protocol (TCP)—one of the protocols on which the Internet is based.

Undershoot—in digitizing, a line that has not been extended all the way to the point where it should connect with another line.

Universal Transverse Mercator (UTM)—a cylindridcal map projection that is often used for large-scale (i.e., local) mapping.

User segment—the portion of the GPS system that includes all types of receivers of GPS signals.

Vector data—a GIS data model in which the location and shape of objects is represented by points, lines and areas that are fundamentally made up of x,y coordinates.

Vector product format (VPF)—a vector GIS format developed by the United States National Map and Imagery Agency (formerly Defense Mapping Agency), intended to become a universally accepted vector data-exchange format.

Vertex—one of a series of x,y coordinate that defines a line. The first and last vertices of a line are usually called nodes.

Vertical integration—see **integration**.

Wide Area Network (WAN)—a computer network that connects computers over large distances by means of high-speed communications links or satellites.

World Wide Web (WWW)—originally developed by the European Laboratory for Particle Physics Consortium (CERN) in Switzerland as a system to distribute electronic documents that are comprised of, or point to, many different files of various formats that are located around the world. Documents are created in a standardized hypertext markup language (html) that can be interpreted by web browsers on a user's computer. The location of html documents are defined by links or addresses, called universal resource locators (URLs). The WWW has been rapidly growing and is becoming an important channel for distributing documents and data. Specialized GIS software allows an organization to serve maps on the WWW. For instance, a remote user can design and display a thematic map using GIS databases located on the distributing organization's web server.

Additional glossaries and dictionaries can be found in Padmanabhan and others (1992), ASCE (1994), McDonnel and Kemp (1995) and Dent (1998). Online resources include the following:

Canada Centre for Remote Sensing	http://www.ccrs.nrcan.gc.ca/
Geographer's Craft Project (University of Texas)	http://www.colourado.edu/geography/gcraft/contents.html
GPS World magazine	http://www.gpsworld.com/gpsworld/static/staticHtml.jsp?id=8000&searchString=glossary
Perry-Castañeda Library, University of Texas	www.lib.utexas.edu/Libs/PCL/Map_collection/glossary.html
United States Census Bureau	www.census.gov/dmd/www/glossary.html
United States Geological Survey	http://interactive2.usgs.gov/learningweb/explorer/geoglossary.htm

Annex VII
Useful addresses and URLs

GIS packages

Autodesk Inc.	San Rafael, CA	AutoCAD	www.autodesk.com
Bentley Systems Inc.	Huntsville, AL	MicroStation	www.bentley.com
ESRI, Inc.	Redlands, CA	ArcGIS, ArcInfo, ArcView, ArcExplorer, Atlas GIS	www.esri.com
Intergraph	Huntsville, AL	GeoMedia	www.intergraph.com
MapInfo Corp.	Troy, NY	MapInfo GIS	
Microsoft Corp.	Redmond, WA	MapPoint	www.microsoft.com
Oracle Corp.	Redwood Shores, CA	Oracle Spatial	www.oracle.com
UNSD Software Project	New York, NY	PopMap	www.un.org/Depts/unsd/softproj/index.htm
Siemens	Munich, Germany	SICAD Spatial Desktop	www.siemens.com
Smallworld Systems Inc.	Englewood, CO		
PCI Geomatics Group	Richmond Hill, Ontario, Canada	SPANS and PAMAP	www.pci.on.ca
ThinkSpace Inc.	London, Ontario, Canada	MFWorks	www.thinkspace.com
Vision* Solutions	Ottawa, Ontario, Canada	Vision*	

Specialty software

Blue Marble Geographics	Gardiner, ME	Coordinate management and GIS development tools	www.bluemarblegeo.com
Caliper Corp.	Newton MA	Maptitude, GIS+, TransCAD	www.caliper.com
Core Software Technology)	Pasadena, CA	TerraSoar (distributed geospatial databases), ImageNet (online geospatial data distribution)	www.coresw.com
Quantum GIS		Open source software	(http://qgis.org)
Thuban		Open source software	http://thuban.intevation.org
Open EV		Open source software	http://openev.sourceforge.net

Remote-sensing, image-processing systems

Leica GeoSystems/Erdas	Atlanta, GA	ERDAS Imagine	www.erdas.com
Earth Resource Mapping	San Diego, CA	ER Mapper	www.ermapper.com
Clark Labs	Worcester, MA	Idrisi GIS	www.clarklabs.org
MicroImages Inc.	Lincoln, NE	TNTmips	www.microimages.com
PCI Geomatics Group	Richmond Hill, Ontario, Canada	EASI/PACE, OrthoEngine	www.pci.on.ca
Research Systems Inc	Boulder, CO	ENVI visualization software	www.rsinc.com

High-resolution satellite imagery and digital orthophotography

GeoEye	Thornton, CO	Carterra and Ikonos satellites	www.spaceimaging.com
Digital Globe	Longmont, CO	QuickBird and EarlyBird satellites	www.digitalglobe.com
Orbital Imaging Corp.	Dulles, VA	Orbimage satellites	www.orbimage.com
EROS Data Center	Sioux Falls, SD		
Spot Image		Spot satellites	www.spot.com
Maps Geosystems	Munich, Germany	Aerial surveys (Africa, Middle East)	www.maps-geosystems.com
EarthSat	Rockville, MD	Satellite and mapping services	www.earthsat.com

Global Positioning Systems

Magellan Corp.	Santa Clara, CA		www.magellangps.com
Ashtech	Santa Clara, CA		www.ashtech.com
NovAtel Inc.	Calgary, Alberta, Canada		www.novatel.ca
Sokkia Corp.	Overland Park, KA		www.sokkia.com
Trimble Navigation Ltd.	Sunnyvale, CA		www.trimble.com
Garmin			

Journals

GeoWorld, GeoAsia, GeoEurope, GeoInformation Africa, Mapping Awareness, Business Geographics	GeoWorld, Fort Collins, CO		www.geoplace.com
GPS World			www.gpsworld.com
International Journal of Geographical Information Science	Taylor & Francis, London, United Kingdom		
GeoInfosystems	Advanstar Pub., Eugene, OR		
Journal of the Urban and Regional Information Systems Association	URISA, Park Ridge, IL		www.urisa.org/
ISPRS Journal of Photogrammetry & Remote Sensing			www.itc.nl/isprsjournal

Miscellaneous

National Center for Geographic Information and Analysis	Santa Barbara, CA	GIS research center	www.ncgia.ucsb.edu
International Institute for Aerospace Survey and Earth Sciences (ITC)	Enschede, Netherlands	GIS Training Courses	http://www.itc.nl/
European Umbrella Organization for Geographic Information (EUROGI)	Netherlands		www.eurogi.org
U.S. Federal Geographic Data Committee	Reston, VA		www.fgdc.gov
Permanent Committee on GIS Infrastructure for Asia & the Pacific			www.permcom.apgis.gov.au/