ST/ESA/STAT/SER.F/79

Department of Economic and Social Affairs
Statistics Division

Studies in Methods **Series F No. 79**

Handbook on geographic information systems and digital mapping

United Nations
New York, 2000

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 designation “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/79

UNITED NATIONS PUBLICATION
SALES No. 00.XVII.12
ISBN 92-1-161-426-0

Preface

The United Nations has, over the years, issued a series of handbooks and technical reports intended to assist countries in planning and carrying out improved and cost-effective population and housing censuses. These handbooks and reports have been reviewed from time to time to reflect new developments and emerging issues in census taking. The present handbook is part of a series of handbooks that have been developed to assist countries in their preparation for the 2000 and future rounds of censuses. The other handbooks in the series include:

(a) Handbook on Population and Housing Census Editing (ST/ESA/STAT/SER.F/82);
(b) Handbook on Census Management for Population and housing censuses (ST/ESA/STAT/SER.F/83).

The *Principles and Recommendations for Population and Housing Censuses – Revision 1* (United Nations, 1998) make reference to the emergence of new technologies for census operations. One of the new technologies is the application of geographic information systems (GIS) and digital mapping in censuses since technical developments in computer hardware and mapping software have already encouraged many statistical and census offices to move from traditional cartographic methods to digital mapping and geographic information systems.

The purpose of this publication is to assist countries by providing a reference document that focuses on digital mapping aspect when conducting population and housing censuses. Traditionally, the role of maps in the census has been to support enumeration and to present aggregate census results in cartographic form. In addition to enabling more efficient production of enumerator maps and thematic maps of census results, GIS now plays a key role in census data dissemination and in the analysis of population and household data.

In particular, the objectives of the publication are to provide guidance to countries on how to:

a) ensure consistency and facilitate census operations, particularly at the pre-enumeration phase;
b) support data collection and help monitor census activities during enumeration; and
c) facilitate presentation, analysis and dissemination of census results, during the post-enumeration phase.

The publication is divided into three chapters. The structure reflects as closely as possible the census cycle. The first chapter gives an introduction and overview of geographic information systems and digital mapping. The second chapter discusses, *inter alia*, cost-benefit analysis of an investment in digital cartography and GIS, plans for census cartographic process, digital map database development, quality assurance, database maintenance, and use of GIS during census enumeration. The last chapter describes the role of GIS and digital mapping in the post-censal phase and deals with tasks after the census and during the inter-censal period, such as database maintenance, dissemination of geographic census products, and geographic analysis of census data.

The handbook is as comprehensive as possible without overloading the reader with too much technical presentation, which is dealt with in the annexes. The annexes provide technical aspects such as an overview of GIS, coordinate systems and map projections, geographic data modelling, and thematic mapping.

During the revision process, the United Nations Secretariat consulted cartographic and GIS experts representing all regions of the world to review and finalize the handbook. The handbook also presents, some examples of country practices in the application of GIS and digital mapping used in censuses contributed by some of these experts. The present publication was drafted by Mr. Uwe Diechmann, a consultant for the United Nations Statistics Division.

Contents

Abbreviations and acronyms

ASCII	American Standard Code for Information Interchange
BMP	Bitmap
BPS	Bits per second
BUCEN	United States Bureau of the Census
CAD/CADD	Computer-aided design/Computer-Aided Design and Drafting
CCD	Charge-coupled device
CD-ROM	Compact disk-read only memory
CGM	Computer graphics metafile
CLA	Crew leader area
CMY	Cyan magenta yellow
CMYK	Cyan, magenta, yellow and black
CSDGM	Content Standards for Digital Geospatial Metadata
DEM	Digital elevation model
DGPS	Differential global positioning system
DHS	Demographic and Health Survey
DPI	Dots per inch
DVD	Digital video/versatile disk
DXF	Drawing exchange format
EA	Enumeration area
ED	Enumeration district
ESRI	Environmental Systems Research Institute
GB	Gigabyte
GIF	Graphics interchange file
GIS	Geographic information system
GLONASS	Global Navigation Satellite System
GPS	Global Positioning System
HLS	Hue lightness saturation
HPGL	Hewlett-Packard graphics language
HTML	Hypertext Markup Language
HVS	Hue value saturation
ISO	International Organization for Standardization
JPEG	Joint Photographic Experts Group
LAN	Local Area Network
MB	Megabyte
NSDI	National spatial data infrastructure
PDF	Portable document format
PES	Post enumeration survey
RDBMS	Relational database management system
RGB	Red, green and blue
SPOT	Satellite pour l'observation de la terre
SQL	Structured query language
TCP	Transmission Control Protocol
TIFF	Tagged image file format
UPS	Uninterruptable power supply
UTM	Universal Transverse Mercator
VPF	Vector product format
WMF	Windows metafile
WWW	World Wide Web

I. Introduction and Overview

A. The role of maps in the census

1.1. Many of the changes to the *Principles and Recommendations for Population and Housing Censuses* (United Nations, 1998) reflect the emergence of new technologies for census operations. Significant technical developments will undoubtedly benefit census data capture, processing and distribution. In the cartographic domain, advances in computer hardware and mapping software have already encouraged many statistical and census offices to move from traditional cartographic methods to digital mapping and geographic information systems (GIS) (see, e.g., Rhind, 1991; Ben-Moshe, 1997; and United Nations, 1997a).

1.2. Traditionally, the role of maps in the census has been to support enumeration and to present aggregate census results in cartographic form. Cartographic automation has greatly expanded this role. In addition to enabling more efficient production of enumerator maps and thematic maps of census results, GIS now plays a key role in census data dissemination and in the analysis of population and household data.

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

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

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

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

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

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

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

1.4. The remaining sections of the introduction will provide a brief overview of the objectives of the handbook. The following section summarizes the rapid developments in digital mapping that have also been the motivation for preparing the present handbook; the next section discusses why census offices are under increasing pressure to provide timely census data in geographically referenced form. Finally, the contents of the handbook are summarized in brief.

B. The mapping "revolution"

1.5. 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 tool for many aspects of professional and academic work.

1.6. Cartography has been affected by the information revolution somewhat later than other fields. Early computers were good at storing numbers and text. Maps, in contrast, are 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 mapping applications implemented on computers in the 1960s did not therefore find wide application beyond a few government and academic projects. It took until the 1980s for commercial geographic information systems to reach a level of capability that would lead to their rapid adoption, for example, in local and regional government, urban planning, environmental agencies, mineral exploration, the utility sectors and commercial marketing and real estate firms.

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

1.8. New information sources also shorten the time from project planning to operational database. The most important recent developments have been in navigation and remote sensing. The Global Positioning System (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 promise pictures of nearly any part of the earth's surface with enough detail to support numerous mapping applications. The cost of precision digital mapping will fall significantly as a result of the close integration of GPS techniques and digital cameras in aerial photography.

1.9. Similar advances are occurring in the areas of geographic data dissemination. All major GIS vendors now provide the tools to make geographical databases accessible via the Internet on the World Wide Web (www). 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 likely to replace printed maps and digital media as the most important means of data distribution.

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

1.11. Statistical offices were some of the early adopters of GIS. Population, social and economic statistics are the foundation of public planning and management. The spatial distribution of socio-economic indicators guides policy decisions on regional development, service provision and many other areas. Digital techniques allow better management, faster retrieval and improved presentation of such data. There has, therefore, always been a close linkage between geography and statistics—as reflected, for instance, by the fact that the national statistical and mapping agencies in many Latin American countries are housed under the same roof (see, also, EUROSTAT, 1996). 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.

C. Increasing demand for local area statistical data

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

- 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.
- 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.
- Utility service planning. Private and public water, gas, electricity and telecommunications utilities not only use GIS to manage their physical infrastructure, they 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 mobile phone systems around the world by helping to find the optimal location of transmission towers.

- 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 to transportation planning.
- Marketing analysis. Companies use small area census data to plan the location of new stores and warehouses, to manage customer service information and to target advertising. An entire branch of GIS—termed, variously, business geographics or geodemographics—has emerged. In fact, the strong demand for these types of analysis has been a major driving force for the development of inexpensive, easy-to-use desktop mapping packages.
- Voting district delineation. In representative democracies, parliamentary representation is based on the principle of equal weight for each vote. To guarantee this principle, small area population figures are used to delineate voting districts of approximately equal size. In fact, in the United States of America this is the main basis for the decennial census required by the constitution. GIS and census data are employed in the design of electoral districts.
- Emergency planning. The identification of highly populated areas that are difficult to evacuate in case of fires, earthquakes, volcano eruptions or tsunamis guides planning for emergency services and allows early removal of bottlenecks. Georeferenced census data, together with digital elevation and transportation maps, are essential tools in such analysis.
- 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.
- Flood plain modelling. Major flooding appears to be an increasing risk in many of the world's watersheds. Digital elevation and hydrological data, in combination with small area census statistics, allow planners to make detailed assessments to reduce the risk for populations in flood-prone areas and for emergency management planning. Insurance companies use the same tools to assess risk levels of homeowners, which leads to a fairer assessment of premiums.
- Agriculture. Geographic information on agro-ecological conditions, and production data together with small area data on the demand for food products, facilitate the analysis of food security issues. Famine early warning systems have been set up in many countries characterized by fragile ecosystems to prevent major food crises.

1.13. Common to all of these examples is that they rely on the availability of small area demographic and social data. The only reliable sources of such information are censuses or, where they exist, population registration systems. As the number of non-traditional uses of census data increases, so does the responsibility of the national statistical office as the main producer of such information. This implies that census offices need to widen their data distribution strategies from tabular reports of fairly aggregate data to detailed digital databases that link boundaries of reporting units to the rich small area population information collected during a census. This wider use of census data also has implications for institutional cooperation. To ensure the widest possible benefit of data collection, data development needs to be coordinated with other government departments, research institutions and private enterprises that are producing geographically referenced data. Statistical offices thus become one of the key participants in the development of a national spatial data infrastructure (NSDI).

D. Scope, purpose and outline of the handbook

1.14. The rapid recent developments in digital mapping technology and the increasing demand for georeferenced small area population data have been the main motivation for the present handbook. Any country embarking on a census project will need to evaluate available options to minimize the cost of and maximize the benefits from the required mapping activities. This handbook is aimed at providing technical and methodological background information to support the choice of a suitable set of tools and procedures for a given country.

1.15. Clearly, the choices will be different in each case given the multitude of available options and the differences in conditions and available resources among countries. The handbook is therefore not a step-by-step guidebook. Each country needs to evaluate how available mapping options fit into the context of its own census programme. Issues such as the already available digital map base in the country, existing technology resources and staff, available funds and the time-frame allocated to complete the census mapping program will determine the best mix of technology and approaches for each individual case.

1.16. The present handbook does not argue, however, that traditional mapping techniques that have been used successfully in many countries are completely obsolete. The main reference on the topic—*Mapping for Censuses and Surveys* (BUCEN, 1978)—continues to be an invaluable resource for beginners as well as experienced cartographers. In particular, the chapters on organization and control of a mapping program, enumeration area delineation and statistical areas continue to be relevant. As technology has progressed, however, there are now better ways of doing many of the census mapping tasks. The present handbook therefore aims to complement the United States Bureau of the Census (BUCEN) guidelines by providing information on recent technologies while avoiding reiteration of material that has already been well covered.

1.17. The main 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.

1.18. The main chapters of the handbook are divided into those topics relevant to the preparation of the enumeration and post-enumeration activities. Chapter II provides a discussion of the costs and benefits of a digital approach to census mapping. This will show that a switch to digital techniques—while inevitable in the long run—entails major upfront investments, while the main benefits may only be realized later. The remaining sections of the chapter discuss institutional issues in planning and setting up a digital census mapping program, GIS database development, and the development of output products for census activities. The chapter concludes with a brief description of GIS applications during enumeration activities such as monitoring the progress of census activities and updating of the map database.

1.19. Chapter III focuses on post-enumeration mapping tasks and on the use of digital mapping in presenting, analysing and distributing census data. These issues are important for all countries. Even if digital mapping has not been used for the actual enumeration, countries may still wish to develop GIS databases for analysis and distribution of census data. These digital base maps will also provide a basis for digital mapping in support of future census and survey application.

Pre-enumeration

A. Introduction

2.1. The decision to shift from traditional to digital census mapping techniques has a major impact on a census organization. The most immediate issue is the major investment required in converting existing analog map information into digital form and to compile new digital information. The associated cost of equipment and data purchase, staff training and operational expenses are significant. A discussion of the costs and benefits of this change in the census mapping approach is contained in section B, and it is argued that the investment will be worthwhile provided that the census office adopts a long-term strategy. Only if the initially developed digital database is maintained and updated after the census and if it is used for purposes beyond enumeration area mapping inside and outside the census organization will the benefits exceed the initial costs.

2.2. The remainder of the chapter is focused on practical, operational issues. Section C covers the initial planning stages, including institutional issues, the definition of the census geography and GIS database design. The issues discussed cover areas such as cooperation with other agencies, staffing requirements, geographic coding schemes and the selection of the scope of census activities. Section D discusses the technological options for data conversion from analog to digital maps and for fieldwork. Since the rapid technological developments in recent years have changed the nature of mapping, the topics covered here are the main reason for the development of the present handbook. Section E deals with quality assurance issues and the production of enumeration area maps for the census operation, and section F, finally, discusses the use of GIS during enumeration.

2.3. Although the sections of the present chapter present a logical sequence, the issues covered cannot be seen in isolation. Staffing, training and equipment purchase, for instance, are determined by the choice of the data conversion strategy. The production of enumeration area maps depends on the available digital data, which is determined by the scope of digital mapping activities. The material in this chapter should thus be seen as informative background material, not as a step-by-step manual.

B. Cost-benefit analysis of an investment in digital cartography/geographic information systems

2.4. The present section will discuss the costs involved and the potential benefits realized in using a digital cartographic or GIS approach in census mapping. The discussion is necessarily general, since there is no single approach to census mapping that will be best in each circumstance. Rather, there are a variety of options, ranging from a fully digital in-house mapping capability to using, for example, desktop mapping for presentation of results and dissemination only. In other words, there is no “one size fits all” solution to the introduction of digital mapping in the census process. In fact, GIS is sometimes criticized as a $500 solution to a $5 problem (e.g., quoted in Batty and others, 1995). This is certainly the case when a complex, high-end GIS is introduced where a simple desktop mapping package would have been sufficient. The appropriateness to the task is the overriding principle of any cost-benefit analysis.

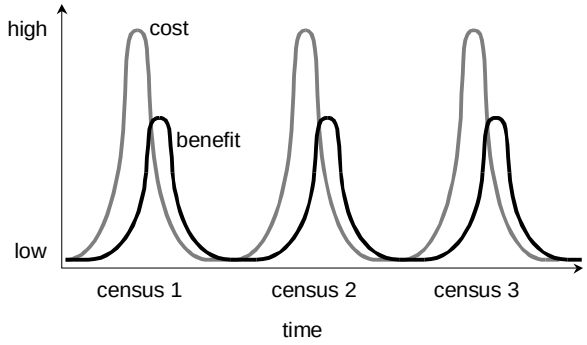
2.5. For various reasons, it is also difficult to assess the costs and benefits of using GIS quantitatively. For example, many of the benefits may not be realized by the agency paying for the GIS investment, but rather by outsiders who are gaining access to products of higher accuracy or lower cost, 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, in the short term, for producing census maps may be the traditional manual approach, especially in countries where labour costs are low. From a societal standpoint, however, it may be more cost-effective to initially invest more money in a digital approach because the digital output products will realize much larger long-term benefits inside and outside the census or statistical agency.

2.6. Investments in GIS are heavily front-loaded. That means that the major costs are incurred early on in a project while tangible benefits may only materialize long into the project cycle. This is illustrated in Figure II.1, which contrasts the costs and benefits of a traditional mapping approach with digital cartography. In the first case, maps are recreated manually for each census. The costs tend to be higher than the benefits, since the hard-copy maps are useful for census purposes only. In the second case, an initially large investment results in lower maintenance and updating costs and sustainable benefits in the long run. The long-term

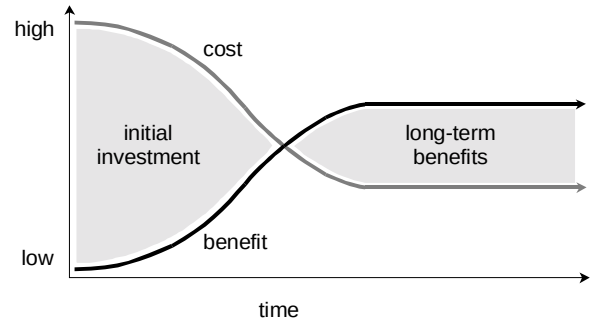
benefits are significantly higher because the process results in a multi-purpose digital database.

Figure II.1. Costs and benefits of census mapping options

(a) Traditional mapping approach



(b) Digital mapping approach



2.7. This also highlights the importance of a long-term strategy for census mapping. Often, census mapping is purely project-based. A few years before the census, a team is assembled to quickly produce census sketch maps by hand, which are only used for the enumeration. Several years later, the process, starts again for the next census. A better approach is to view census cartography as a continuous process, with regular maintenance of databases by a permanent core staff who receive frequent training.

1. Costs

2.8. The short-term investment and longer-term maintenance costs of GIS should not be underestimated. Like any new technology or organizational transformation (e.g., management information systems), the introduction of GIS involves a change in routine and significant expense, not only for software and hardware but also for data purchase, training, planning and organizational restructuring. In fact, the significant costs involved are the main reason why the sections on GIS in the revised *Principles and Recommendations for Population and Housing Censuses* (United Nations, 1998) are worded very carefully. The indirect costs, in particular, are often underestimated and may lead to the failure of a GIS project.

2.9. A list of tasks that may be involved in introducing GIS and that incur costs to the introducing agency are set out below (see Worrall, 1994; and Becker and others. 1996; see, also, Bond and Worrall, 1996; and Bond and others, 1994). These steps are discussed in much more detail later in the handbook. Many of the costs are obviously not unique to digital cartography. For example, the costs involved for coordination of decentralized data collection or data conversion are similar whether the maps are produced in digital form or manually. Also, not all steps will be required for every project. A census office that only wants to use desktop mapping to produce thematic maps for a census publication will not need to spend a lot of time or money on a detailed planning process. A comprehensive census mapping project, in contrast, may require a considerable investment, and its success or failure may be a direct function of how rigorously the project was designed.

Cost components

Systems design and planning, consulting services; managerial staff time	Overall planning of a GIS project or a GIS department within an agency will clarify the objectives and anticipate the costs and steps involved. Using outside expertise will often be useful, and a large commercial sector offering GIS consulting services has emerged in recent years. For census offices in developing countries, it is often useful to visit other offices with significant experience in GIS to learn from their experience (see, also, Coiner 1997). Also part of the overall system planning is an evaluation of available data and the development of a data conversion strategy, which is often the most resource-intensive part of a project.
Hardware acquisition or	Computers are becoming increasingly powerful, while prices continue to fall. Some of these gains, however, are offset by the increasing demands on processor speed and

integration

memory by new software products. If existing equipment is to be integrated, investments in upgrading memory or disk space may be required, and the issue of compatibility with newly purchased equipment needs to be addressed. Apart from computers with fast processors and plenty of storage space, GIS applications also require peripherals such as digitizers, scanners and large-format colour printers, which may not be standard equipment in the census office.

Evaluation and selection of GIS/mapping software

There are now dozens of suitable GIS and desktop mapping software packages on the market, ranging in price from a few hundred to tens of thousands of United States dollars. In the next few years, many analysts predict a further consolidation of the GIS software market, which should lead to lower costs for software since the remaining vendors will benefit from higher volumes.

For all practical purposes, the choices can be reduced to a few packages that have emerged as the standard among agencies and that have the capability of dealing with the large and complex databases of a census mapping project. These provide sufficient user support and the functions required to perform all tasks in a census project.

Compatibility with other government agencies is an important criterion if data are frequently exchanged or data production costs are shared among agencies. Also, a hierarchical approach may be appropriate where the main GIS unit may use a powerful software program, while regional field units or groups mainly involved in routine tasks rely on cheaper, less powerful software.

High-end software vendors often require or encourage the purchase of maintenance contracts, which need to be considered in operational budgets. Such services tend to be expensive, but are often critical to ensure uninterrupted operations.

Prototype development

Before embarking on a census mapping project, it is advisable to conduct a prototype or pilot project in a small area of the country. While this requires additional time and resources, the benefits are that problems in the methodology can be detected early. For large projects, potential software vendors should be asked to provide benchmark information, with a realistic application defined by the client organization. Thus, when evaluating a system, the census office should make sure that any demonstrations or benchmarks are carried out using a realistic data set that reflects the full complexity of census mapping. Demonstations always work well with the vendor's own packaged data set. But this may not reflect actual performance after deployment of the system for census mapping work.

Hardware/software system configuration/ customization

Since GIS data development is time-consuming and labour intensive, a distributed approach is usually advisable. This is greatly supported by a networked system where data can be exchanged easily either through a Local Area Network (LAN), a special network linking, for example, the national census office with regional offices, or, increasingly, through standard Internet connections.

For very large census applications, some customization may be required, for example, to develop an interface between the GIS package and a generic database management system that is already in use.

Human resources planning

The implementation of a new technology in an agency may require the addition of new staff. It may, for example, be necessary to bring in a person with experience in digital mapping or GIS to head a new section devoted to this area. Similarly, staff training needs or reassignments have to be established to ensure a smooth transition from the old to the new mapping system.

Training, skills development, re-training

Besides the cost of hardware, software and data conversion, training is the fourth major expense of any GIS activity. Estimates range from 5 to 10 per cent of the total project cost that should be budgeted for training. Training costs are high largely because of the lack of suitably skilled entry-level job applicants, the complexity of many GIS software packages and the limited background in geography and spatial analysis of most staff in statistical offices.

It is likely that these issues will become less of a problem in the future. Many

universities are now teaching GIS not only in geography departments, but also in computer science, natural resources, business and statistics programmes. Standard core curriculum development for universities and vocational schools (e.g., NCGIA, 1998) support this development. GIS software is becoming more user-friendly as the Windows platform emerges as a standard and as vendors consider the needs of an increasingly broad and non-specialized user community. For example, many low-cost desktop GIS systems now allow the display of remotely sensed images, from which features can be digitized on-screen. Such operations previously required specialized image processing software and training in remote sensing techniques.

Still, training requirements should not be underestimated and a continuous updating of staff skill levels is required given the rapidly changing hardware and software market. Ideally, staff training should not be limited to teaching the basic steps for carrying out routine tasks. In the long run, it will be beneficial to allow staff to learn about more general concepts such as accuracy of spatial data or the possibilities of geographic data analysis. A more informed, motivated and creative work-force will lead to better census geographic products.

Database design, data modelling, procedural manual development

Data modelling is the process of defining features to be included in the database, their attributes and relationships, and their internal representation in the database. Data modelling involves the development of conceptual, logical and physical models of the census geographic database. The outcomes include a comprehensive data dictionary that defines the content of the databases that are produced by the agency. In some instances, such dictionaries can be adopted or adapted from other agencies in the country, for example, where a national digital topographic database exists. In other cases, this data dictionary has to be developed from scratch. The resources required for this will depend on how comprehensive the database will be.

It may also be necessary to integrate existing database models that have been developed to manage tabular census information. This is necessary, for example, if data from previous censuses need to be integrated in the GIS databases.

In addition to the data dictionary, a procedural manual defines the steps necessary in the development and processing of digital spatial data. Such manuals are important for ensuring consistency in the products that are generated by different technicians or units that may, perhaps, be scattered across the country. They also define recurring analytical work such as the methods used to reconcile past census data with new boundaries after the administrative units have been changed.

Accuracy standards should also be defined as part of the overall database design process. While accuracy is often not a critical issue in census mapping – in fact, many countries rely on hand-drawn sketch maps for this purpose – it becomes an issue when the resulting census maps are used in combination with other, higher-accuracy data.

Transitional costs

Additional costs are incurred if old and new systems have to be operated in parallel during a transition period. This is necessary to ensure quality of service while problems in the new system are worked out. For a transitional period, keeping the old system as a back-up may be a good strategy if many users rely on the timely delivery of products.

Data acquisition, data purchase

Some of the information required for census mapping may be obtainable from commercial sources or other agencies that charge a fee for their use. Auxiliary geographic data sets describing road networks, hydrology or elevation are useful for census mapping since boundaries should ideally be designed to match features that are identifiable on the ground. Obtaining such data from outside vendors or other government agencies will save both time and money and will also increase the consistency of data products across different agencies.

Also, as described in section D below, aerial photography or satellite imagery can be used to support the production of census maps. These are obtained from outside vendors or, in the case of air photos, commissioned from a private company.

Data capture, conversion

Initial data development is probably the most expensive part of a GIS project. The share of this element in the overall project budget, together with data acquisition from outside vendors, is often estimated to be 60 to 70 per cent. This is significantly larger than the hardware and software costs.

Data capture includes cartographic fieldwork using traditional techniques or new methods that will be discussed in later sections. Data conversion or data automation, in contrast, refers to the process of creating digital GIS data layers from hard-copy maps. For this process, two options are available. Maps can be traced manually using a digitizing table, or the entire map can be scanned and a data set suitable for input to the GIS is generated through subsequent raster-to-vector conversion. Both approaches are discussed in section D.

Validation, quality assurance/quality control

No matter which data conversion strategy is chosen, data conversion is labour intensive and error prone. A rigorous procedure for checking the resulting data for both positional accuracy and logical consistency should therefore be part of the process. Similar procedures should be implemented to assure the quality of derived output products such as cross-tabulations or GIS overlays. If high accuracy is the objective, it is not conservative to allocate resources equal to those budgeted for data conversion to the final editing and quality control stage.

Quality control also relates to the development of metadata standards. One major problem of digital data is that documentation is often detached from the actual data and may thus be easily lost. Rigorous procedures are required to avoid loss of accuracy and data quality owing to a lack of information about each specific data set. Metadata should include all information relevant to the data set, including source map reference, date, projection and scale, processing steps performed on the digital data set, data lineage and accuracy standards. Metadata formats for digital spatial data have been developed by many national mapping agencies and can be adapted to the requirements of a statistical office.

System maintenance

System maintenance involves software and hardware upgrades, as well as any training that may be required as a result of such upgrades. This component is often estimated to consume about 10 per cent of the initial investment per year, although this figure will vary depending on the scale and scope of the project.

Post-implementation review

Even after a detailed planning process and pilot study, further improvements will often be possible after full-scale implementation. It may therefore be useful to have an in-house or external review of the system to identify weaknesses and to improve productivity.

However, the development of a GIS capability within an agency should not be seen as a linear process, with a clearly defined ending date, but rather as a continuing process of improving operational procedures. A periodic review of the GIS group's work should thus be part of the regular activities.

Development of data distribution strategies

While anyone can make use of printed census publications, and most users of digital, tabular census data will have access to spreadsheet or similar office software, users may not have easy access to mapping or GIS software for using digital census maps. To achieve maximum use of such data, the census office should design a strategy to help users obtain access to such software.

This will not be a large problem in the more developed countries, where users will be able to purchase the required software. In less developed countries, several options are available to increase the use of digital spatial data. These include cooperative agreements with a software vendor to reduce the purchase price or subsidizing software purchase with public funds, in-house development of a data viewer, and the use of free or public-domain software such as PopMap (United Nations, 1997b; and Vu, 1996).

2. Benefits

2.10. Following Worrall (1994), we can distinguish between *efficiency* benefits and *effectiveness* benefits. The first implies that after a transition period, more or better output can be obtained with the same amount of input, or that the same output can be produced with fewer inputs. Such efficiency effects include cost savings or productivity gains and are mostly realized by the census organization itself, which may be able to produce maps faster or with fewer resources than before. Effectiveness, in contrast, refers to the impact of policies or programmes that benefit from improved information. These benefits are mostly realized by the users of statistical data derived from a population and housing census. For example, the availability of digital population maps that can be used in combination with environmental information may result in better decision-making within the environmental protection agency of a country. Both efficiency and effectiveness benefits are discussed below.

(a) Efficiency benefits

2.11. Efficiency effects will largely be realized through cost savings, cost avoidance and productivity gains through reduction in the time required to produce output products. Such benefits are usually measurable, although they may not be realized until well into a GIS project. However, benefits also accrue if higher-quality or entirely new products can be generated. For example, if a digital map is produced at a higher accuracy compared with a manually drafted map, no time or cost savings may be involved, but an overall benefit is still realized. The following list contains a mix of measurable "hard" benefits and "soft" indirect effects.

Productivity gain and time savings	After the initial investment in creating the digital database, updates can be produced faster, and more and better output products can be generated with the same number of staff. Digital data also allow a much wider range of applications within the national statistical office such as sampling frame development or combination with other data layers such as land use information to create new statistical indicators. Copies of updated maps can be printed immediately, without tedious manual redrafting. This also enables the national statistical office to respond more quickly to changing demands and needs by data users.
Cost saving/cost avoidance	The replacement of a technician who is manually redrafting map data sets with a computer operator can – after a transitional training period – result in lower staff requirements and associated cost savings. Similarly, digital census maps can be adapted more easily to other purposes such as agricultural or economic censuses or special-purpose sample surveys. Digital mapping using remote sensing can be cheaper than extensive fieldwork, especially in areas of rapid change, where timely maps are difficult to obtain, or in remote areas that are difficult to access. Likewise, production of output products—especially of low-volume special products—will be less expensive using a digital census database than using manual techniques. Digital map data provide a more secure archiving system compared to paper maps since multiple back-ups that can be stored off-site are cheap and easy to produce. Such back-ups will also require less storage space than a large collection of paper maps.
Greater credibility and authority of map products	Apart from productivity gains and general cost savings, digital mapping will help census operations in several other ways. For instance, digital techniques enable the production of professional-looking enumerator maps in small numbers. These carry greater authority with the large number of temporary census employees than hand-drawn sketch maps.
Better service	Digital data result in faster turnaround times for standard census products. For example, if the enumeration area maps were already created digitally, tabulated census data can immediately be linked to produce thematic maps. Similarly, special-purpose products such as tailor-made maps or custom aggregation of census data can be created quickly. Special mapping products cannot be produced cost-effectively in small volumes using manual techniques. With digital techniques, even one or two copies of a map requested by a census office customer can be produced quickly and cheaply.

Increased accuracy — Compared to sketch maps, the digital approach encourages higher accuracy which results in better products and supports a wider range of applications. Some digital techniques such as digital orthophotos provide a high degree of "built-in" accuracy. For census mapping, improved accuracy in mapping results in a more exact definition in the definition of enumeration areas. This will reduce census errors such as undercounts or double counts owing due to imprecise boundary delineation.

Improved consistency — Similarly, a digital database is likely to result in a seamless database for the entire country. This will ensure a high degree of consistency-which is important, for example, if census units are rearranged.

It is also easier to incorporate metadata in a digital database. For instance, systems have been developed that track operations on digital GIS databases, so that the end product of such operations will be accompanied by a full description of data lineage and GIS procedures that have been used. Alternatively, one could introduce a system where the operator needs to fill in a pre-designed metadata form whenever changes have been made to a data set or when a new data set is added to the archive. Enforcing such procedures in a traditional, manual system is more difficult.

A complete digital census mapping operation will also ensure that there is complete agreement between boundaries used for data collection and those used for the production of output products, since both come from the same digital master database.

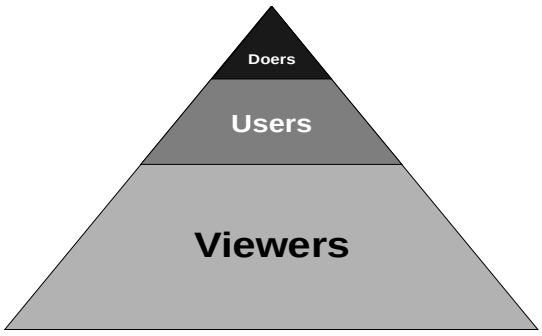
Income generation — Since digital map data allow for a much wider range of applications, a market for such products has developed in many countries around the world. Data users in the private sector include marketing firms, banks, real estate companies, health-care providers, environmental organizations and academic institutions. Moderate pricing of such products increases their widespread application, leads to larger volumes combined with lower production costs, and will support a thriving secondary market in associated mapping services.

Attempts to achieve full cost recovery through high prices and strict copyright enforcement, in contrast, prices casual and non-profit users out of the market and limits the accessibility of such data to a relatively few well-off commercial data users. As the experience of various countries shows, the conflict between increasing pressures to generate maximum revenue versus the overall societal benefit of inexpensive, widely accessible data has so far not been resolved.

(b) Effectiveness benefits

2.12. Effectiveness benefits reflect the impact of digital GIS data in the work of other government institutions, academic or non-profit organizations and the private sector. User needs vary. Rajani (1996), for example, reviews two market segmentation models used by GIS vendors. In the first, the market is divided into the degree of sophistication of the data user. "Doers" are the people who input, maintain and create digital spatial data, do advanced analysis and modelling and will typically use high-end GIS software on powerful computers. "Users" are in the middle category and will perform basic analysis such as combining several map layers to create cross-tabulations. Finally, "viewers" will use spatial data for basic tasks such as making thematic maps and querying an existing database. It is estimated that the number of "viewers" is larger than the number of "users", which outnumber the "doers", each by an order of magnitude. An alternative market segmentation model is based on the software's cost and capabilities, both of which increase steadily from basic consumer maps to desktop mapping, desktop GIS (which allows data creation and simple analysis) and professional, fully functional GIS.

Figure II.2. GIS market segmentation
(after Rajani,1996)



2.13. Many users may not realize the full potential of census data. Since, traditionally, census information has mostly been available in very aggregate form in printed publications, many users who might benefit from detailed, small-area statistics in digital form do not have the background to envision how these data can help in their work. Outreach seminars and publications produced or contracted by the census office that focus on the use of the data may help increase the user base and, consequently, the indirect benefits of census taking. The *Census Users' Handbook* (Openshaw, 1995), which covers the 1991 census of the United Kingdom of Great Britain and Northern Ireland a good example of such a publication.

2.14. Below is a list of effectiveness benefits that may be realized by data users and, to some extent, also by the census agency. Only some of these benefits can be quantified in terms of time or money savings or increased productivity (see, also, Nordisk Kvantif, 1987 and 1990). Mostly, however, the benefits are more indirect. For example, improved visualization or analysis may not necessarily save time or money, but will lead to better insights and understanding and, consequently, to better decision-making.

Improved analysis

The purpose of statistical data compiled and published by a census office is to support planning and decision-making in the country. Thematic maps based on census derived statistics provide an analytical basis for a wide range of public policy applications. Combined with tables and statistical graphics, maps provide an added dimension to data analysis which brings us one step closer to visualizing the complex patterns and relationships that characterize real-world planning and policy problems.

For example, a clustering of high childhood mortality in a number of enumeration areas may point to some environmental condition that has caused this pattern. Higher fertility rates in another set of regions may point to a cultural preference for large families. This information could be used to adapt family planning outreach programs. Visualization of spatial patterns also supports change analysis which is important in monitoring of social indicators. This in turn should result in improved needs assessment. In short, the availability of statistical and other information in spatially referenced form and the functions provided by a GIS can allow analyses that were previously too expensive or impossible to perform.

Improved policy-making

Improved analysis should, in turn, improve policy-making. For example, statistical GIS databases are useful in site selection for public services such as hospitals, fire stations or schools, or in evaluating different planning scenarios. Auxiliary GIS data layers, in combination with small-area statistical data, can be used for targeting of interventions to alleviate poverty or to reduce economic imbalances in a country.

In combination with statistical or simulation models, GIS can also be used to perform "what-if" scenarios and to support resource allocation decisions. For example, after estimating an econometric relationship between some indicator of interest and a number of explanatory variables that can be affected by policies, we can estimate the impact of a number of different policies (such as a given increase of per capita spending on education) on the villages or enumeration areas. GIS allows us to put the results in a spatial perspective and to determine where the impact will be greatest. This clearly encourages a disaggregated approach to policy analysis. Instead of looking only at overall impacts, the focus is on zooming in on areas of greatest need.

Improved data sharing

Converting data into digital form should lead to improved coordination and data sharing among government agencies (Batty, 1992). Data sharing should also result in improved consistency of derived products that are created by other organizations. To realize these benefits, clear collaborative agreements between partner agencies within the Government have to be developed. Such agreements should cover any cost accounting that may be required and should also cover issues of data formats, accuracy standards and content definitions.

Improved outreach

Another benefit that should not be underestimated is the fact that graphical representations of data are usually more appealing and generate more interest than tables of numbers alone. One of the main reasons for the success of GIS is undoubtedly the power of pretty maps. This can also help to make the work of a statistical office more accessible, improve outreach and raise awareness of the benefits of census taking.

3. Critical success factors

2.15. Besides the obvious costs that can be quantified for a given GIS project, there are a number of stumbling blocks that may cause a project to fail or to fall short of realizing its full potential. For the most part, such problems are connected to a lack of planning, the choice of inappropriate hardware and software, and various organizational mistakes. Surveys of real-world GIS projects can reveal a set of characteristics shared by successful GIS implementations. The absence of these factors, in turn, also points to possible reasons for the failure of such projects. The following list of critical success factors is adapted and expanded from Johnson (1997):

1. A key person to promote GIS development within the organization.
2. High-level management support.
3. Decision to invest in GIS is need-based and problem-driven rather than technology-driven.
4. Detailed strategic, operational and managerial planning based on a realistic assessment of costs and effort involved.
5. Clear goals and objectives defined for the GIS department.
6. GIS education and training for affected employees *and* management.
7. Staff continuity—the ability to retain skilled staff members.
8. GIS treated not as an independent add-on, but as an integral part of the overall information management strategy.
9. Completion of a user needs assessment and a priori definition of output products.
10. Development of cooperative agreements with other interested parties.
11. Clear implementation schedule.
12. Defined long-term funding plan, including cost-recovery and data pricing strategies.
13. Accurate estimates of maintenance and associated costs.
14. Explicit operational procedures guiding the use of GIS facilities.
15. Well-established quality control/ quality assurance procedures.
16. Clear specifications, requirements and benchmarks to deal effectively with vendors and contractors.
17. Well-defined written contracts with vendors, consultants, partners and clients within and outside the Government.
18. Completion of a prototype pilot project to test appropriateness of equipment, software and procedures.
19. Frequent milestones and delivery of output products to encourage adherence to pre-set time-frames.
20. Outreach and marketing, including published successes.

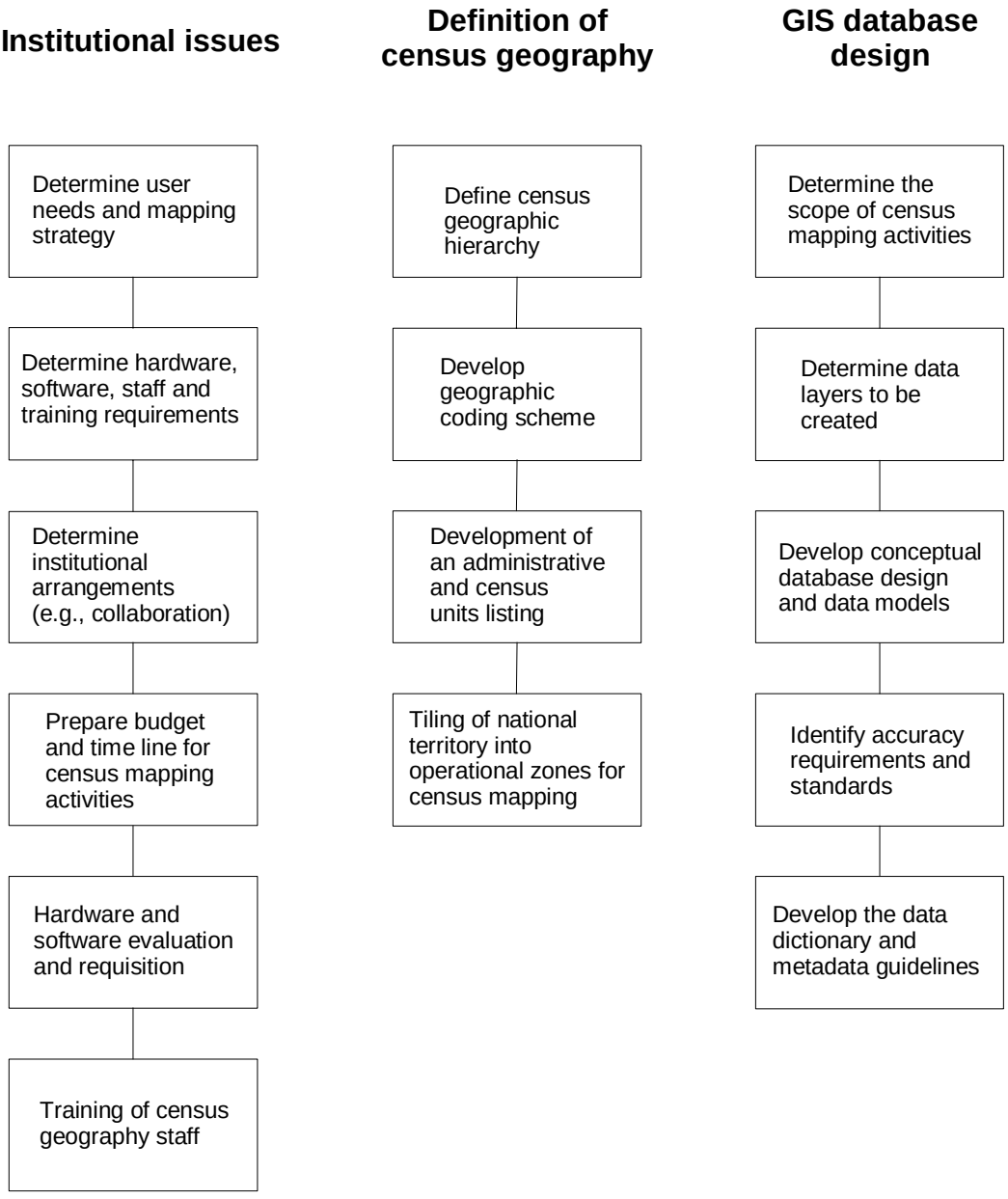
C. Planning the census cartographic process

1. Overview

2.16. The present section deals with preliminary organizational tasks in a census mapping project and with critical design issues that determine the nature of the resulting databases and, thus, the range of applications that it will support. The success of the actual data conversion process depends on a well-designed institutional environment and a well-planned operational strategy. The planning steps are divided here into institutional issues such as staffing and cooperation with other agencies, the definition of the census geography and the design of the GIS database. As

illustrated in Figure II.3, these stages can be carried out more or less simultaneously, and many of the choices depend also on the chosen data conversion strategy.

Figure II.3. Stages in planning census cartographic work



2. Needs assessment and determination of mapping options

(a) User needs assessment

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

2.18. A successful census planning process requires extensive consultations with the main users of the

information that will be produced in the census. This process should be embedded in the general consultation programme for the census (see United Nations, 1998; paras.1.73–1.76 discuss these issues in detail). As the demand for spatially referenced census data increases, consultations concerning mapping products will receive a more prominent role in this process. Institutions that use statistical maps should therefore be included in the advisory panels that provide input in the census planning process.

2.19. Following United Nations (1998), the census office must consult with three main groups in the planning stages:

a) Census map product users. These will mainly come from other government departments, the academic research community and the private sector;

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

c) The general public. However, with access to computers and Internet mapping options, private users will also become 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 data that did not exist yesterday, may not be apparent today, but may be commonplace tomorrow.

(b) Determination of output products

2.20. 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 census mapping agency, which are discussed in more detail in Chapter III, may include:

- A set of digital enumeration area maps that are designed to enable the production of all output products that will be disseminated to government departments and the public;
- Geographic boundary files for all statistical reporting units for which census indicators will be tabulated;
- Listings of all statistical and administrative reporting units, including towns and villages;
- 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;
- 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 settlements and other important geographic features in the country.

2.21. User requirements are the most important determinants of a census mapping design. However, these must be weighed against available resources. Various other factors determine the choice of the mapping strategy. Among these are

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

(c) Mapping options

2.22. Different countries will start their census mapping efforts from a different basis of existing information, budgets, technical capabilities and available time-frames. *There exists, therefore, a multiplicity of paths towards a fully digital map database for census collection and dissemination purposes*. A partial list of available options—in increasing order of complexity— follows:

- Production of rudimentary digital maps created on the basis of existing sketch maps;
- Georeferenced enumeration area maps that can be properly integrated with other digital geographic databases;
- Inclusion of geographic reference layers, showing, for instance, roads, rivers, and other features; these can be included as simple images from scanned maps or designed as a structured vector 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.

2.23. The above list is for illustrative purposes. All of these issues are discussed in detail in the remainder of the handbook. The best census mapping strategy for a country will consist of a tailor-made approach that considers the country's needs and resources. While a step-by-step type approach is thus not feasible, the present 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.

3. Institutional issues in setting up a digital mapping program

(a) Staffing, responsibilities and training requirements

2.24. Motivated and well-trained staff are a key factor that will determine the success or failure of a digital census mapping project. The goals of a census mapping project are similar whether the maps are produced by hand or by computer. But the use of computers requires a number of new skills from census cartographic staff since similar products are created using different techniques (see Broome and others 1995). Furthermore, a digital GIS database is useful for many more purposes. A census office is thus likely to fulfil additional demands for products and services that were not available before. Every member of the census cartographic staff should therefore have some degree of computer literacy.

2.25. Much of the expertise required in the traditional, manual census mapping approach is relevant also for a digital mapping project. Rather than completely replacing existing skills, the digital mapping approach requires additional expertise in computer methods. Thus, only relatively little of the expertise of cartographers and geographers on the staff is obsolete, but the demands on their job skills have increased. For instance, traditionally trained cartographers will no longer need some techniques of manual 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.

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

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

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

2.29. *GIS data conversion.* Responsible for the actual implementation of the conversion of map information to digital database format, data conversion specialists have training in relevant GIS techniques such as digitizing, scanning and editing of GIS databases, and attribute database development using relational database management systems. Data conversion specialists must determine the most efficient way to develop the digital database and supervise technical staff.

2.30. *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, and cartographic communication, and training in GIS and digital mapping. Classically trained

cartographers will have most of the required skills, but should receive sufficient training in computer methods.

2.31. *Fieldwork*. The requirements for census mapping fieldwork have changed with the techniques used for digital map production. As global positioning systems have become an essential tool for field data collection, field staff must now be trained in the operation of these systems and possibly also in the use of laptop 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.

2.32. *Map digitizing*. Digitizing is a very repetitive task. 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 requires good concentration, attention to details and a good understanding of the structure of digital geographic databases. The best performing digitizing staff should also receive training in quality control/quality assurance approaches.

2.33. *Systems administration*. Timely completion of a digital census mapping 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 back ups). Even if they are not directly involved in census mapping activities, systems administrators are vital members of the cartographic team, since almost every aspect of the work depends on a well-functioning computer system. Administration of computer systems for the geography branch of the census office can, in some instances, be covered by general computer support staff in the agency.

2.34. *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. Such skills are helpful in the development of the attribute databases and in any customization of software systems.

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

- Short seminars to raise the awareness of the digital census mapping program should be conducted for all staff of the census office, including staff from other sections. This will foster the integration of the digital mapping project into the overall census process. Better utilization of census mapping products by other census office branches will be another benefit of broad dissemination of information. Such seminars can be conducted by the project leader or specialist census cartographic staff.
- Training for repetitive tasks such as 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.
- The core geographic staff involved in census mapping should receive additional training in GIS and digital mapping techniques. Since training is 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.
- Applications of specialized techniques such as digital image processing or advanced computer database applications usually require a professional degree or equivalent practical 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. Several universities and training centres around the world now specialize in professional one- or two-year degree courses in GIS, remote sensing and related techniques.

(b) Institutional cooperation

i. Ensuring compatibility with other government departments

2.36. In many countries, several 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 GIS to manage the information they collect and use for analysis and planning. Additionally, private sector companies, for example in the utilities, telecommunications and mining sectors, have realized the advantages of managing their information needs in digital geographic form.

2.37. Numerous users inside and outside government agencies require access to these basic geographic databases. Many of these users need access to several databases or use a standard geographic data layer as a template for their own spatial data collection. Such standard data layers, which provide the basis for many mapping and data collection activities, are termed *framework data* (FGDC, 1997a; and Rhind, 1997). In the United States, for example, the core data layers that form the national spatial data framework are:

- *Geodetic control* – a system of precisely determined geographic control points that serve as the reference for all mapping activities in a country; frequently also referred to as benchmarks;
- *Ortho-imagery* – air photos or high-resolution satellite images that have been processed to have the same geometric accuracy as a topographic map,
- *Elevation*,
- *Transportation* – infrastructure used to move people or goods,
- *Hydrography* – surface water features; these can be natural such as rivers and lakes or artificial such as canals,
- *Governmental units;*
- *Cadastral information* – an official register of rights and interest in land property.

2.38. Other basic data layers such as soil types, vegetation zones and planning information could be added to this list. Most relevant for the census office are the governmental units, since enumeration areas need to be consistent with the boundaries that form the administrative hierarchy in the country. 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. Conversely, enumeration area boundaries with census information are an important data source for other government and private organizations. Health sector analysis, for example, requires detailed information about at-risk populations. Transportation sector planning needs 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.

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

- The census office has a responsibility to contribute to the national spatial data infrastructure a consistent set of data reporting units 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.
- 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.

ii. Standards

2.40. To facilitate data exchange between data users it is clearly necessary to coordinate development of geographic databases. In several countries, national geographic data committees have been formed for this purpose, which bring together the key persons in charge of spatial data development. In addition, supranational organizations such as the European Umbrella Organization for Geographic Information (EUROGI), the Permanent Committee on GIS Infrastructure for Asia and the Pacific, the European Commission and the International Organization for Standardization (ISO) are active in defining geographic data standards (see Open GIS Consortium, 1996; Heine, 1997; Moellering and Hogan, 1997; and Rhind 1997).

2.41. Unfortunately, this multitude of players has given rise to a confusing array of definitions and standards. For an individual agency, it is therefore difficult to determine the most appropriate guidelines in the choice of geographic feature definitions, data formats, metadata, and software platforms. These issues are discussed in more detail in sections (iv) below.

iii. Collaboration

2.42. In the process of digital census mapping, the census organization may have the option of collaborating with other government agencies or with the private sector. Both options have been used successfully in different countries. Among government agencies, the national mapping agency is the most natural first point of contact. But other 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. It should be noted, however, that collaboration with other agencies is desirable but not mandatory. Since the census mapping agency must produce a map base for the census at a given time, it needs to avoid complete dependency on an outside supplier of map information.

2.43. 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 (adapted from FGDC, 1997a):

- Formalization. Is a loose collaboration sufficient or do the arrangements need to be highly formalized? A more formalized agreement will take considerable time to put in place, but can avoid later disagreements about rights and responsibilities concerning the development and use of data products. In most instances, therefore, a formal, legally binding letter of understanding between the census office and the cooperating agency should be put in place that covers all relevant aspects of the partnership. Such formal contractual arrangements are mandatory when dealing with private sector suppliers of data or services.
- Scope of partnership. Collaborative agreements may cover simply the use of another agency's data, or they may involve the development of a large, comprehensive spatial database from scratch.
- Responsibilities. Who will perform which tasks and functions? Issues that need to be addressed include data development, maintenance, data access, project supervision and resource use.
- Benefits. Clearly, the arrangement must be of benefit to all participants, unless one agency simply purchases the services of another. It is useful to clarify how the different partners will gain from the arrangements in order to fairly divide tasks and responsibilities.
- Resource requirements. Resources include staff, computing environment, materials and communication. Resources required for management and project supervision must also be considered.
- Cost sharing. Any direct and indirect costs connected to the activities of the partnership must be divided fairly. Accounting may not be straightforward since contributions can be in cash, data, labour, equipment use, or some other way.
- Cost recovery. If any revenues are generated from the distribution of the final products, they need to be shared, with consideration of the costs that are incurred by managing and operating data distribution. This also involves a clear determination of agreed-upon uses and the copyright of the output products.
- Conflict resolution. In the event of disagreements during the course of the project, it is useful to a course of conflict resolution in advance.

Twinning strategies

2.44. Collaboration is not limited to the sharing of products or services among agencies within a country. Some census offices have implemented collaborative mechanisms with census offices in other countries. Such twinning arrangements can be set up between countries that have a similar level of resources, technology use and statistical systems, or they can be arranged as a technical assistance strategy between countries that currently employ a different level of census mapping technology. Depending on available resources, a collaborative agreement can cover the exchange of ideas through regular visits and workshops, collective research projects or even the joint procurement or sharing of resources such as equipment that is not required on a continuous basis, or special expertise.

Box II.1.: Interagency collaboration among mapping agencies in Australia[*]

2.45. Australia provides a good example of how collaboration between various government agencies can have a positive effect on the availability of digital geographic data even beyond the needs of the census agency. For the 1996 census in Australia, the national mapping agency was unable to supply a comprehensive national digital map base because it had no mandate in larger-scale mapping in urban areas. The Australian Bureau of Statistics, therefore, facilitated the formation of a consortium of state, territory and federal mapping agencies for the development of a national digital base map. This consortium, the *Public Sector Mapping Agencies*, is updating the digital map base for the 2001 census and is widely praised as one of the most positive recent developments in the mapping field in Australia.

Source: Frank Blanchfield, Australian Bureau of Statistics, personal communication; Rhind, 1997, chap. 13.

(c) Equipment and software for census mapping applications

2.46. The selection of suitable computer equipment can only be made after all other aspects of the census mapping project have been thoroughly planned. Computer hardware and software technology develops very fast, with new and improved products coming to market constantly. Purchases should therefore not be made too early, since there is a danger that the equipment or software will be out of date by the time it will be put to use. Most of the equipment required for census mapping is standard for other computer applications as well. Computers, monitors or printers procured for census cartographic purposes can therefore be used for data entry or processing at a later stage in the census.

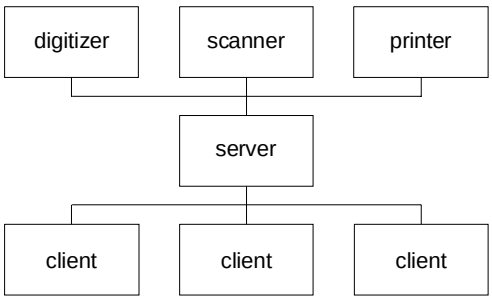
2.47. Additional hardware—scanners, digitizers, and large-format plotters—are specific to GIS and other graphical applications. Two notable characteristics of census mapping applications are the large volume of materials produced and the importance of the timely completion of a map database creation since the entire census operation depends on these map products. In purchasing equipment and software, the census agency therefore needs to ensure that the chosen products support high-volume applications and have shown a high degree of reliability and dependability in similarly demanding applications. Another desirable characteristic is ease of use and maintenance, since the large number of staff involved in a mapping program will include many novice computer users. Equipment components are discussed in the following paragraphs.

i. Computers and networks

2.48. Recent years have seen a move from high-end Unix-based GIS workstations to mapping software run on personal computers. Complex and demanding GIS applications, which required powerful workstations only a few years ago, can now run on standard, off-the-shelf PCs. This has significantly lowered the cost of GIS installations. It has also made adoption of GIS easier since the software can be run on familiar interfaces in the Windows environment. While Unix based GIS will remain popular in high-end applications or as servers for large networks of PC workstations, the needs of most census GIS applications can be met by standard PCs.

2.49. In order to facilitate data exchange and sharing of peripherals—printers, plotters, scanners and so on.—computers need to be connected in a Local Area Network. In peer-to-peer networking, computers simply allow access to local files from other computers. This type of networking and file sharing is supported by standard operating systems. The more prominent networking model is the client-server architecture see (Figure II.4), where a powerful computer serves as a central repository of file archives and software and as a link to peripherals. For example, printers are accessed from other computers through the server. Central storage of important files and software makes maintenance—such as software updates and back-ups—easier. The client computers are standard PCs, possibly with quite powerful processors and large local file storage. Standard software packages, for example business packages, may be installed locally on each PC.

Figure II.4: The client-server model

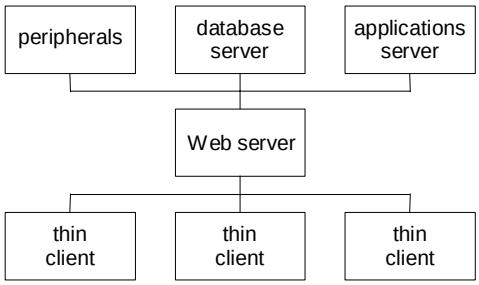


2.50. The client-server model enables the census mapping organization to design a heterogeneous computing environment. Older or cheaper computers can be used as digitizing workstations, which do not require larger processing power. GIS database development and analysis require faster computers, and for output production a number of machines can be optimized for fast printing. The advantages and possible cost savings from a heterogeneous computing environment must, of course, be weighed against the more complicated maintenance and support required when networking different types of computers.

ii. Current developments in computer networking

2.51. The current client-server network model will serve the computing needs of most census offices for some time to come. But the model may well have seen its prime. Internet and intranet technology provides a platform-independent, standardized networking environment. Network servers may, in the future, be replaced by Web servers that maintain files and software and manage network traffic locally and in communication with the outside world. The Web server acts as a database and applications server to any number of so-called *thin clients* or *network computers* see (Figure II.5). These are simple computer workstations, with limited processing power and storage space. Users simply download software from the server as required—for instance, a digitizing module or a cartographic design program—together with the necessary components of the database. Elements of this network computing model are available at the time of writing, although a fully operational Web-based GIS, with input, manipulation and output functions, has not yet been developed.

Figure II.5. The network computing model



2.52. Internet connectivity is already present in many census offices. Some census offices use the Internet as the main communication and data exchange mechanism between the central office and local offices that are distributed around the country. In a distributed census mapping project, where much of the data collection and basic mapping is done locally, the Internet also provides the means for exchanging digital maps and other relevant data, as well as reports, guidelines and documentation.

2.53. The next step from network computing is considered “pervasive computing”, in which Web servers support not only thin clients, which are essentially scaled-down PCs, but also other types of devices. Smart mobile phones, digital set-top boxes, networked vehicles, personal digital assistants and notebook computers will all be able to interact with Web servers, using standard Internet protocols. It is too early to speculate on the implications for census mapping. But it is not inconceivable that in the future, census enumerators will collect information with digital personal assistants that automatically collect the geographic coordinates of the household, using a built-in global positioning system. The information is immediately transmitted to a central Web server in the census office, using wireless data transfer technology, allowing real-time monitoring of data collection activities. While the technology for such a process is currently available, the cost of deploying it at the scale required for census data collection is far beyond typically available resources.

2.54. Computer networking technology is constantly changing. The planning of the computing environment for a census mapping project must therefore include a careful review of the current state of the art in computer networking. Of course, this needs to be coordinated with the general computing environment of other parts of the census operations in order to maximize the use of investments in all stages of census taking.

iii. Storage devices

2.55. GIS applications are often characterized by a large data volume. Digital maps can consist of hundreds of thousands or even millions of coordinates. Derived products such as plot files and data tables also require a large amount of storage space. Systems planners should therefore budget for sufficient storage space—internal on the computer’s hard drives, as well as on external storage devices. Hard-drive capacity is increasing constantly, while prices continue to decline. While allowing storage of large data amounts locally, hard drives have the disadvantage of not being portable and of limited expandability. External storage devices include the following:

- Magnetic tapes are still popular for back-ups. They are inexpensive and have a high storage capacity. Modern tape drives use small tapes that nevertheless can hold many gigabytes of data. The disadvantage of tapes is that data access is slow.

- CD-ROM (compact disk-read only memory) is a popular distribution medium for software and data. Most PCs are now equipped with CD-ROM reader and external readers are inexpensive and easy to connect to a computer. CD writers have also dropped significantly in price and have become more reliable in recent years. Write-once CDs are inexpensive and provide a convenient way of distributing low-volume, tailor-made data sets or to produce a master copy for external high-volume production. A CD-ROM can store about 630 megabytes of data.
- DVD (digital video/versatile disk) will likely replace CD-ROM as the dominant data and software distribution medium in the future since it has higher capacity (several gigabytes). Fairly inexpensive DVD writers are available, although at the time of writing there is still some uncertainty as to which rewritable DVD standard will become universally accepted. Writable DVDs have a capacity of more than 5 gigabytes. DVD drives should become as widespread as CD drives over the next few years.
- High-capacity diskettes are likely to replace the 3 ½ inch floppy disk as a flexible storage device for data exchange. Current models use disks that are not much larger than a floppy disk, but hold between 100 megabytes and 1 gigabyte.

iv. Input devices

2.56. A census mapping project is a large data conversion effort. Cartographic and attribute information from many sources are compiled and converted into digital form in a consistent GIS database. Data input devices are therefore a key component of the computing environment in the census mapping office. Input devices are listed below with only brief explanations. More detailed discussion is given in section 0.

- Keyboard. Manual keyboard entry of coordinates is rare. Sometimes, however, it may be faster to type in coordinates from a gazetteer or similar source than to digitize point locations from a map sheet.
- Mouse. The standard pointing device for graphical user interfaces becomes a coordinate entry device when features are digitized on the screen, using a scanned map or satellite photo as a backdrop.
- Digitizer. High-quality data conversion from maps printed on paper or drafted on stable materials such as mylar are the domain of digitizing tables. Digitizing tables come in many sizes—the larger the digitizing area, the larger the maps that can be converted in one piece. Even if the data conversion project is largely based on scanning technology, a digitizing board can be useful for specific applications. Software drivers for standard digitizing tables are included with most GIS or desktop mapping packages.
- Scanner. Large-format scanners can significantly reduce the time and cost required for data conversion. Features are extracted from the scanned maps either by subsequent on-screen digitizing or through automated raster to vector conversion software. For applications requiring high quality, similar features contained in complex maps are often redrafted onto separate sheets of paper or mylar before scanning. Raster to vector conversion routines are included in some GIS packages, but for large-scale applications a specialized software package is preferable.
- Global positioning systems. GPS are handheld devices for field data capture. The receiver is activated by an operator and determines its location with a fairly high degree of accuracy. The coordinates can be stored in the GPS and downloaded directly into the GIS. GPS has quickly become the most important field data collection device for geographic applications.

v. Output devices

2.57. Computer mapping is a graphical application, and high-quality output devices are crucial for working with digital maps and to present the results of data compilation and analysis. Large, high-resolution monitors are relatively expensive, but make working with digital maps significantly easier. Monitors with 17 inch or larger screens should be purchased for the graphics workstations, while smaller screens are sufficient for non-graphical work such as data entry or processing. A good video card and a large amount of dedicated video memory can increase drawing speed significantly.

2.58. Hard-copy output is produced using printers and plotters. A mapping project will need both-large-format plotters for printing test plots for quality control and supervisory maps, and small-format printers to produce a large number of enumerator maps in a cost-effective way. Among large plotters, colour ink-jet technology has replaced pen plotters as the standard for GIS applications. For small-format printing, black and white laser printers are fast and reliable. Although small-format colour printers based on ink-jet technology are inexpensive, they do not provide the throughput required for printing a large volume of maps. Supplies such as ink and toner can also be quite expensive, leading to higher prices per page printed and, consequently, higher overall costs compared to the initially more expensive laser printers. Printers and plotters are discussed in more detail in chapter III.

vi. Systems safety and maintenance: uninterruptable power supply and back-up strategies

2.59. A safe computing environment requires a reliable power supply. Inadequate supply of electricity can cause data loss, computer crashes, systems damage and loss of productivity owing to down times. Electricity problems can occur in various forms:

- Power outage (blackout or power failure), a complete loss of power owing to, for example, unreliable utility generation, insufficient electricity distribution infrastructure, lightning or heavy rains;
- Surge voltage (spikes, transient or impulse), a short-duration overvoltage;
- Sag (brownout or undervoltage) often owing to undercapacity of electricity generation.
- Swell (overvoltage), which in contrast to surge voltage delivers higher than normal voltage for a longer period.

2.60. In areas where power supply is unreliable, an uninterruptable power supply (UPS) is a mandatory component of a computer installation. These systems compensate for over- and undervoltage and provide power for a time that is sufficient for a safe shutdown of systems in case of blackouts. Uninterruptable power supply are useful in any setting, but in countries where power supply problems are frequent, they have to meet higher than normal demands. If blackouts are frequent, the batteries in the UPS will be discharged and recharged more often. If power outages last fairly long, the systems will be discharged deeper and often may not be completely recharged before the next power outage. The available back-up time of the system and the life expectancy of the batteries in the UPS can thus be drastically reduced. For these reasons, a UPS operating in an environment of frequent power failures needs to be of higher quality and must undergo more frequent maintenance than a system in a more stable setting.

2.61. The required size of a UPS can be determined based on the power demands of the systems that will be connected to it. Computer equipment have a voltage (V) and current (ampere(A)) rating that is specified on the machine and in the user's manual. UPS capacity is measured in voltage amperes (VA). The VA required is calculated by multiplying the voltage and amperes for each piece of equipment and summing the results. For instance, a small set-up of a computer with 120V and 2A, plus a monitor at 120V and 1A, would require a 360VA or higher rated UPS. This UPS would provide back-up for at least eight minutes. If longer back-up is required, a larger UPS needs to be selected.

2.62. Computer equipment require a controlled environment for long-term reliable operation. In addition to a dependable source of electricity, the equipment should be protected from large temperature fluctuations such as extreme heat or cold. Ideally, computers should be operated in an air-conditioned environment that also offers protection from dust. These requirements are no different from those concerning the computer equipment used in census data entry and processing.

2.63. On an operational level, a comprehensive back-up strategy must be followed throughout the course of data development and maintenance. Inexpensive back-up systems are available that can produce copies of data distributed on a network. Back-ups are time-consuming and are best run overnight. Usually, an incremental back-up is made everyday, adding only those files that are newly created or changed on that day. A weekly full back-up produces a copy of the entire file system. Back-ups do not have to be made of software files and data sets for which the original media are available. However, software parameter files that store customized information should be backed up regularly. It is a good idea to store a weekly or monthly back-up off-site in a safe place. This can prevent complete data loss if a fire or other disaster destroys the computers and locally stored back-ups.

2.64. Finally, a last systems safety issue pertains to unauthorized access to files produced by the census agency. Maps by themselves are usually not sensitive information. However, census microdata are typically subject to privacy regulations that prevent the release of information about individual persons or households. While Internet connectivity in a census organization facilitates data exchange and access to outside information, faults in a system can also enable outsiders to access internal file systems. The network system should therefore be designed to secure the internal computing environment, for example, by means of a fire wall, and only allow outside access to a separate system that may contain aggregate maps and data tables that are released for general distribution.

vii. Software

2.65. The growth of applications areas of GIS has also led to rapid developments in the field of GIS and desktop mapping software. The field is dominated by a small number of market leaders, with numerous companies providing add-ons and more specialized applications software. Packages can be crudely divided into high-end and low-end systems. High-end systems include software that provides hundreds of functions for both vector and raster data, integration of remote sensing products, support for surveying and other specialized applications, and

virtually unlimited options for customizing the installation. Such systems require considerable training as they are often characterized by non-intuitive user interfaces. Until recently, these packages only ran on powerful workstations under the Unix operating system. Most of these have now been ported to the Windows NT operating system and will also be available for the forthcoming Windows 2000.

2.66. Low-end systems include so-called desktop mapping packages that emphasize thematic cartography and ease of use, and are often sold with pre-packaged generic data sets. Some desktop mapping packages can be customized using add-ons written in a software-specific macro language or in Visual Basic. Commercial GIS and desktop mapping packages have become so functional and adaptable that in-house development of mapping software, which had until recently been quite common in large mapping projects, has all but disappeared.

2.67. Census mapping requires, first of all, data conversion and database development. These are basic functions that do not require a high-end GIS package. The bulk of the demands of a census mapping program can therefore be satisfied with relatively inexpensive software for data input—digitizing and editing—and map output. Yet, for more advanced applications, such as spatial analysis or complex topology building, one or a few licences for a higher-end GIS package are useful. Choosing a suite of appropriate software packages for a large-scale mapping package requires a clear definition of the tasks to be completed and the number of operators involved at each step.

2.68. Questions to ask when choosing software include the following:

- Does the software provide all functions necessary for the census mapping project?
- Does the software handle all types of data that will be used in the census project (vector GIS, raster GIS, air photos, satellite images and text data)?
- Does the software provide an interface to the database management software used by the census office?
- Is expensive customization required?
- Does the software support hardware that is already available in the agency?
- Will it import/export data from/to other packages used in the organization or by collaborating agencies?
- Does it support existing standards used by other agencies involved in mapping in the country?
- Does the vendor have a good maintenance and customer support policy? Is a knowledgeable local representative available?
- Does the vendor provide good conditions for site licences that allow the office to run the software on several machines at the same time?
- Are training materials available, or does the vendor provide training seminars locally?
- Can the software be easily adapted or expanded if requirements change during the project or at a later stage? Is it possible to migrate to a more powerful system later with minimum cost for data translation and adaptation of custom-designed functions and interfaces?

2.69. The GIS software market changes quickly and developments are difficult to predict, even for the next few years. Fortunately, the growth of the market also means that many sources of information about software and hardware trends are available.

2.70. One recent trend in GIS software development is the move towards incorporation of geographic data types in generic relational database management systems (RDBMS). Most GIS software today use system specific geographic data formats, while storing the attribute information in a generic database management software format. An alternative is to store the geometric descriptions of spatial features such as an enumeration area or a road as a special data type (wide fields, abstract data types, or binary large objects) in a database engine. This technology, which is offered by vendors of RDBMS in cooperation with GIS software firms, has great potential for storing large geographic databases and for adding a spatial component to existing tabular census databases.

viii. Importance of a long-term view

2.71. Just as it is important to maximize compatibility with previous censuses it is also desirable to consider future census activities during the planning of an enumeration. Especially in cases where no permanent cartographic unit exists in a census organization, a strong emphasis must be put on documentation, metadata and archiving. New staff who are hired for the next round of censuses must be able to reconstruct what has been done in previous enumeration activities in order to make full use of existing digital cartographic materials.

2.72. Similarly, in the choice of hardware and software, the census organization should attempt to keep all systems as open as possible. Often, an organization can become trapped in a particular technology that may be outdated by the time the next census comes around. Migrating to new platforms is then very expensive. It is, of course, difficult

to predict changes in the dynamic field of computing over a time span of 10 years or more. Even so, there are some general rules that can help to ensure continuity in census mapping operations:

- Today's market leaders are more likely to be around in the future than are start-ups or small companies. These firms also have a strong incentive to provide backward compatibility in terms of data formats to allow users to migrate to their newer systems with little effort.
- The census office should maximize the use of data formats that can be imported by many systems. If a proprietary data format is used, it is advisable to export all data into a generic, widely used format at the end of the census process.
- It is usually not advisable to create software in-house. In the past, census agencies often developed their own systems for census mapping because suitable software was often not available at reasonable cost. Especially for data dissemination, the lack of inexpensive, easy-to-use software limited a wide distribution of census data. In the long run, however, maintenance of in-house developed systems becomes expensive and a census organization usually does not have the resources to keep up with the fast-changing computer industry. Private companies now offer a wide range of mapping software packages at reasonable cost. Some mapping packages that are suitable for census data dissemination are even available free of charge.

(d) Decentralization of census mapping activities

2.73. In a relatively small country, census mapping can be carried out by a centralized cartography unit in the national office of the census organization. For a larger country, in contrast, it is beneficial to decentralize mapping activities. The basic structure of a decentralized mapping effort should be based on the system of national and regional census offices that is set up for the actual enumeration and for other statistical data collection efforts.

2.74. The decentralization of mapping activities has several advantages. Local staff are more knowledgeable of the geography, administrative structure and recent changes in their assigned region. It is easier for a local office to maintain a continuous working relationship with local authorities. Field-work will be less expensive, since travel distances are smaller. Especially in cases where problems encountered in map preparation requires a return to a field site, resources may be saved if local staff are conducting the cartographic work locally. Finally, a major benefit is that local expertise in an important new technology is created, which will generally benefit the region, even if census cartographic staff move on to other positions after the census operation is concluded.

2.75. Decentralization of mapping activities also has some potential drawbacks. Training and supervision of census cartographic activities need to be coordinated among several, possibly far apart places. While it is possible to train core staff centrally before the census work begins, supervisory functions must be conducted throughout the census in each regional office. A decentralized approach also requires a clear definition of the respective tasks of the regional and central offices. The flow of materials and products must be carefully monitored to ensure consistency across the entire country. Finally, a digital mapping infrastructure must be replicated at several points in the country. This is not a major problem for relatively inexpensive equipment such as computers and digitizing tables. Some functions will therefore be centralized since they require specialized, expensive equipment or highly specialized expertise. Examples are map scanning on expensive, high-volume scanners or remote sensing and air photo interpretation.

Table II.1. Possible division of tasks between central and regional census mapping offices

Central geography office	Regional or local office
• Overall coordination and training, including inter-agency collaboration • Specialized functions (e.g., high-volume scanning, remote sensing) • Overall data integration • Quality assurance • EA map production	• Fieldwork • Basic digital data development (digitizing, editing) • Liaison with local authorities • Specific quality control tasks • EA map production

(e) Timing of census mapping activities

2.76. A critical part of the planning process in a digital mapping project is to define each required task in detail and to estimate the time required to complete each project component. The required tasks and the timing of mapping activities will be quite similar whether the maps are produced digitally or manually. The detailed descriptions and time lines developed in BUCEN (1978) will therefore serve as an excellent starting point in the planning of a digital mapping project. Just as with manual techniques, the time required for digital mapping depends on many conditions and the choices made. Among the factors

determining the timing of digital census mapping activities are:

- The area and population of the country, and whether all areas of the country are easily accessible;
- How much fieldwork is required;
- The resources available to hire and train staff, purchase equipment and procure outside services;
- The types of resource materials available, such as topographic map series, satellite image coverage for the country, high-quality sketch maps from a previous census and so on;
- Whether digital base data are available from collaborating institutions in a format that can be easily adapted to the needs of the census organization.
- The techniques for data conversion chosen and the types of base maps available (for instance, much time can be saved if colour separations of topographic base maps can be scanned rather than the full colour maps themselves).

2.77. No attempt will be made here to suggest the time required for each step, since conditions will vary widely among countries. Table II.2 shows a list of tasks required. The list has been adapted from BUCEN (1978) to reflect the requirements of a digital mapping strategy. The original table also suggests the sequence and duration of individual tasks.

2.78. Some of the tasks listed in the table will require significantly less time if a digital mapping strategy is chosen. For instance, printing a map for any new or non-standard area in the country will not require redrafting or manual pasting of hard-copy map sheets once the digital database is complete. On the other hand, significant time savings are unlikely to be realized during the initial census map database development, because conversion from paper to digital maps is very time-consuming. Time savings will only be realized in subsequent activities and census enumerations. Consequently, digital census mapping activities tend to be more front-loaded than traditional approaches. That means that the largest effort is required in the early stages of database creation, while later stages—for example, output production and revision cycles—will require comparably less effort.

2.79. A further issue that needs to be considered in the scheduling of census mapping tasks is risk avoidance. Owing to the dependencies of subsequent steps on output from earlier stages, contingency plans must be made for every critical task in the census mapping project. The planning staff should go through a series of "what if?" scenarios to determine back-up options in case a key activity cannot be completed in time.

Table II.2. Tasks in a digital census mapping project

Planning, administration, training

1. Determine scope of mapping program.
2. Determine mapping needs and specifications.
3. Identify statistical areas.
4. Prepare geocoding scheme.
5. Prepare detailed calendar of activities.
6. Design control procedures.
7. Estimate personnel, training, hardware and software needs.
8. Prepare budget.
9. Hire and train additional personnel.
10. Order supplies, equipment and software.
11. Install and test all new equipment.
12. Prepare instructions and training materials on the use of maps in enumeration.
13. Train field staff in the use of maps in enumeration.
14. Receive and file enumeration maps returned after the census and post-enumeration survey (PES).

Preparation of base maps

15. List and code areas for which maps are needed.
16. Make and maintain inventory of existing resource materials.
17. Prepare priority list of areas for compiling maps.
18. Prepare and verify map compilation packages.
19. Digital data conversion (digitizing, scanning, editing, integration of GPS-derived field maps).
20. Review and verify digital base maps—print large-format maps for quality control.

Preparation of enumeration area maps

21. Delineate and code enumeration areas (EA) and crew leader areas (CLA) for the census on the digital base maps.
22. Review and verify delineation and coding.
23. Print EA, CLA, and field office maps for enumeration.
24. Review and verify census enumeration maps.
25. Delineate PES sample segments.
26. Print PES maps.
27. Prepare PES maps for enumeration.

Fieldwork

28. Contact local officials and other agencies for places to be added to list of areas.
29. Acquire resource materials—available digital maps, paper maps, satellite images, air photos, sketches.
30. Update map information (boundaries, names, location of features, etc.)
31. Make quick count of housing units for census EA delineation.
32. Divide PES sample areas into sectors and make quick

counts for sampling.

Distribution of enumeration maps

33. Distribute maps for census enumeration.
34. Distribute maps for PES sectoring and quick count.
35. Distribute maps for PES enumeration.

Preparation of publication maps and charts

36. Design maps and charts using desktop mapping and desktop publishing software.
37. Review and verify census maps and charts.
38. Print and publish census maps and charts in hard-copy format, on CD-ROM or on the World Wide Web.
39. Design and implement a plan for disseminating geographic census databases.

Source: Adopted from BUCEN (1978), exhibit 2-4.

(f) Process control

2.80. The calendar of activities will also allow the project supervisors to monitor progress and to determine whether the products will be finished by the anticipated target dates. Adherence to the timetable must be strictly enforced because of the strong dependencies of each census mapping and enumeration stage on products completed in previous stages.

2.81. To be able to determine at any time the status of each task in the mapping process, a system of process control must be implemented. Process control involves keeping track of the location and status of data sets, materials and products. This will ensure the timeliness of product completion and allow project managers to react to any delays, bottlenecks or necessary adjustments to the process. Should any delays in one activity become apparent, efforts must be increased since subsequent steps in the mapping process depend on outputs from previous activities. Similarly, other tasks in census preparation depend on the work of the cartographic section, and the timing must therefore be coordinated with the overall census planning staff.

2.82. Mapping can be of use in monitoring also. For instance, weekly or monthly status reports on the completion of tasks can contain an overview map for each important step (fieldwork, map automation, EA delineation, etc.), in which the operational areas in the country are shaded according to their status. Alternatively, areas can be shaded by per cent completion. This example illustrates the utility of GIS as a management tool.

2.83. Process control also serves important documentation functions. If problems become apparent with one of the digital map outputs, the process control forms provide a (possibly digital) "paper trail" that allows the census staff to trace and correct the problem.

2.84. In general, then, process control will be based on forms that accompany each output product from the first to the last step of the data conversion and map production process. BUCEN (1978) discusses the purpose, design and use of process control forms in great detail. The control forms trace the path through all processing steps, where materials or products were sent and to whom they were sent, who performed which task, when the tasks were begun and completed, data sources and other pertinent information.

2.85. In a computerized environment, process control can, of course, be automated as well. Commercial project management software exists and can be adapted to suit the needs of a census mapping project. The main advantages are a high degree of consistency, tight control, data security and the ease in which information can be queried and summarized at any given time. Several options are possible for automated process control:

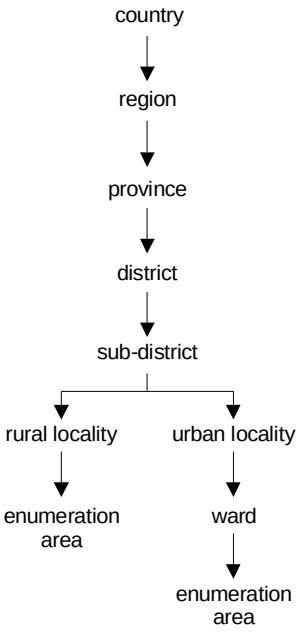
- A central database containing all the forms as data entry interfaces. This can be set up as a stand-alone application implemented in a standard database management system or it can be set up on a password protected Internet or Intranet Web site so that external offices can access it also.
- Separate files that are stored with the digital data products and essentially become part of the lineage information stored with the digital maps in the metadata description (see section 0 below.
- A mixed system, in which some control materials are paper based but the master database is digital. The disadvantage of a purely digital system is that it may separate hard-copy products such as topographic paper maps, printed air photos or printed maps that are sent to local administrators for verification from the process control documentation. While saving paper is a good idea, a mixed approach in which some paper forms accompany physical materials may be preferable. The information from the paper forms can be entered into a central system periodically to allow integration with all other process control information.

4. Definition of the national census geography

(a) Administrative hierarchy

2.86. One of the earliest decisions in census planning pertains to the administrative areas for which census data will be reported. Census preparation involves creating a list of all administrative and statistical reporting units in the country, and the definition of relationships between all

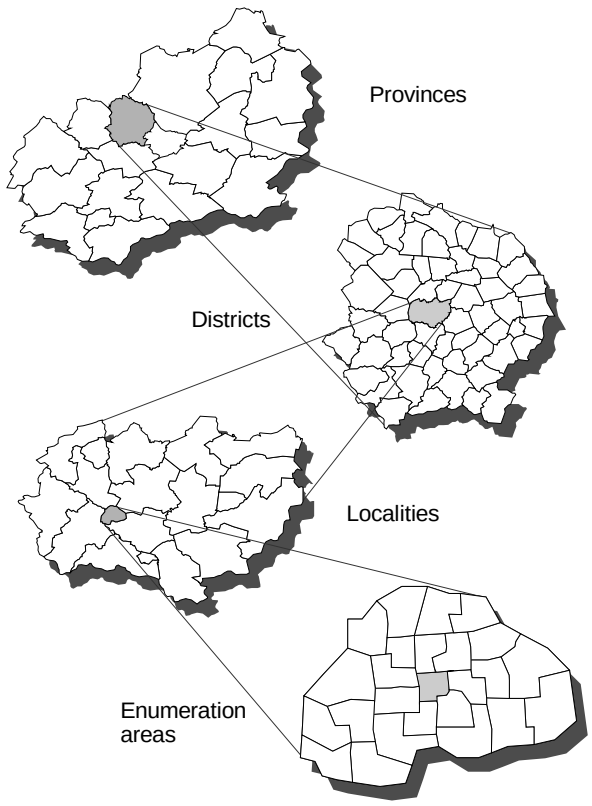
types of administrative and reporting unit boundaries. Every country has its own specific administrative hierarchy, that is, a system by which the country and each lower-level set of administrative units (except the lowest) are subdivided to form the next lower level. For example, for the purposes of the census, a country may have been divided into seven hierarchical levels in urban areas and six in rural areas:



2.87. Only some of these may have actual administrative roles, for example, the province, district and locality levels may have capitals, with local government offices that are responsible for those regions. Figure II.6 illustrates the nesting of administrative and census units using a simple example with only four hierarchical levels. In some instances, however, administrative units may not be completely nested. Especially when considering both administrative and other statistical reporting units, the census office may need to deal with a very complex system of geographic regions.

2.88. Not all levels are 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.

Figure II.6. Illustration of a nested administrative hierarchy



(b) Relationship between administrative and other statistical reporting or management units

2.89. 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. 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).

2.90. Many of these areas will 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.

2.91. 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.

2.92. 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 with agricultural or economic data. A high level of agreement between the geographic units used to compile these types of data will greatly increase their utility in public and non-governmental applications.

(c) Delineation of enumeration areas

2.93. The delineation of enumeration areas is discussed at length in the BUCEN manual on census mapping (BUCEN, 1978 in particular chaps. 2 and 7). The concepts and guidelines pertaining to enumeration area definition are similar whether manual or digital cartographic techniques are used. Therefore, only the major relevant issues will be briefly summarized.

2.94. The design of enumeration areas should take the following criteria into account:

- They should be mutually exclusive (non-overlapping) and exhaustive (cover the entire country);
- They should have boundaries that are easily identifiable on the ground;
- They should address the needs of government departments and other data users;
- They should be consistent with the administrative hierarchy;
- They should be useful, also, for other types of censuses and data collection activities;
- They should be compact without pockets or disjoint sections;
- They should be of approximately equal population size;
- They should be small enough and accessible to be covered by an enumerator within the census period;
- They should be small and flexible enough to allow the widest range of tabulations for different statistical reporting units;
- They should be large enough to guarantee data privacy.

2.95. Among these criteria are some that facilitate census data collection, while others pertain to the usefulness of EAs in producing output products—that is 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 thus has precedence over convenience of census enumeration.

2.96. The size of enumeration areas can be defined in two ways: by area or by population. For census mapping, population size is the more important criterion, but surface area and accessibility also have to be taken into account to ensure that an enumerator can service an EA within the time allotted. The chosen population size varies from country to country and is determined based on 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.

2.97. 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 or, in some cases, by means of aerial photographs, as discussed in a later section.

2.98. Enumeration area boundaries need to be clearly observable on the ground. All enumerators, even if they do not have considerable geographical training, need to be able to find the boundaries of the area for which they are

responsible. Thus, population sizes between enumeration areas may be varied in order to produce an easily identifiable delineation. Natural features that can be used for this purpose are roads, railroads, creeks and rivers, lakes, fences or any other feature that defines a sharp boundary. Features with more gradual edges, such as brushes, forests or elevation contours such as ridges, are less ideal. In some instances, it is unavoidable to use EA boundaries that are not clearly visible on the ground. In this case, an exact verbal 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.

2.99. Specific issues related to EA delineation will be encountered in many countries. For instance, while villages may be assigned to specific administrative units, the actual boundary delineating the village area may not be defined. Also, special populations, such as transient, nomad or military personnel, need to be assigned a geographic reference. For instance, naval personnel are often assigned to their home ports. These issues are discussed in detail in the revised *Principles and Recommendations for Population and Housing Censuses* (United Nations, 1998).

(d) Delineation of supervisory (crew leader) areas

2.100. 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 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.

2.101. 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.

(e) Consistency with past censuses

2.102. 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. Although statistical or other techniques exist that reconcile information for incompatible area units, such shortcuts introduce errors in any subsequent analysis. Most census data users also lack the expertise and tools to do such interpolations. The problem of changing the geographic base between censuses is no less serious than changes of definitions of items on the census questionnaire.

2.103. In designing the census geography, the census office should therefore attempt, inasmuch as possible, to preserve boundaries from the previous census. Owing to increases in population size, new enumeration areas may have to be defined. In these cases, it is always preferable to subdivide an existing enumeration area rather than to change the 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.

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

(f) Coding scheme

2.105. 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. The numeric code also provides the link between the aggregated census data and the digital EA boundary database stored in a GIS. Geographic coding is discussed in BUCEN (1978) (chaps. 2, 6 and 7). The ideal 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 with, the geography and data-processing staff. The most important principles when designing a coding scheme are flexibility, intuitiveness and compatibility with other coding schemes in use in the country. The statistical office is often the custodian of coding schemes in the country and should also be the focal point for the design of the census mapping codes.

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

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

Province	2 digits
District	3 digits
Locality	4 digits
EA	4 digits

2.108. An EA code of 12 035 0175 0023 means that enumeration area number 23 is located in province 12, district 35 and locality 175. The unique code is stored in the database as a long integer or as a 13-character string variable. The variable type needs to be the same in the census database and in the geographic database. Storage as an integer variable has the advantage that subsets of records can be selected easily, using standard database query commands in any database management system or GIS package. For example, the following query will find all enumeration areas within locality number 175 in the database or on the digital map:

```
SELECT ID > 1203501750000 AND
       ID < 1203501760000
```

2.109. 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 rather than a sequential number.

2.110. In cases where administrative and reporting units are not hierarchical, special coding conventions need to be developed. In any case, it is important to be completely consistent in defining and using the administrative unit identifiers, since they are the link between the GIS boundaries and the tabular census data. The census office should therefore maintain a master list of EA and administrative units and their respective codes, and commit any changes made to the master list to the GIS and census databases.

5. Geographic information system database design

(a) Scope of mapping activities

2.111. Once a census organization has decided that the benefits of a digital mapping program outweigh its costs, the next step is to define the scope of mapping activities. Clearly, there is no uniform approach that is appropriate for all countries. Depending on available time and resources, a country may choose to launch a comprehensive mapping program leading to a complete database of enumeration area boundaries, or the goal may be to produce a digital map of more aggregate units such as districts for post-census mapping only. The following paragraphs describe a set of available options.

i. Full census mapping program (complete enumeration area database)

2.112. An ambitious strategy for census mapping is to produce a complete digital database of enumeration areas. The resulting database will be properly georeferenced, allowing aggregation, combination with other digital map layers and dissemination for users who require detailed, spatially referenced population data. Again, there are several options for producing this database. Digital boundaries may be extracted from available hard-copy sketch maps that may have been produced for the previous census. Techniques for converting the line work on the paper maps into digital boundaries are discussed in section 0 below. An alternative approach is to create new enumeration boundaries from topographic maps, which makes it easier to attach consistent geographic coordinates, or coordinates that define EAs can be collected during fieldwork, using global positioning systems.

2.113. Regardless of which approach is chosen, the development of a complete enumeration area database is a challenging task. For most countries, this will likely require several years of work and a large commitment of staff and computer resources. Nevertheless, this should be the long-term goal of any census project.

Box II.2. Mexico's experience in census mapping

2.114. To obtain a sense of the effort involved in a comprehensive census mapping project, the experience of the Mexican National Institute for Statistics, Geography and Informatics (INEGI) in producing digital census cartography for the 1990 population census is instructive. [a/] In 1987, INEGI decided to produce digital maps for the 32 States, 2,403 municipalities, 24,131 AGEBs (statistical units consisting of 25 to 50 blocks) and 905,576 census blocks. The resulting database is known as the Automated Geostatistical Information System (SAIG).

2.115. Sources for the boundary data were a variety of standard cartographic products—urban line maps and topographic maps—as well as digital orthophotos and etensive fieldwork conducted by the operational personnel in the various regional offices of the census organization. The entire process took approximately two years, with 123 staff persons. The time required for digitizing, plotting and quality control in the production of a digital AGEB map dropped from 4.5 hours at the beginning of the project to about 45 minutes in the final stages.

2.116. INEGI initially relied on standard software: AUTOCAD, which is a computer aided design (CAD) package that supports digitizing and basic editing, and Arc/Info, which is a comprehensive GIS package. Later, INEGI developed in-house mapping software that is also used for dissemination (SCINE—System for Consulting Census Data). The system has subsequently been adapted to support the agricultural and economic censuses, and a number of special derived products such as crime maps and databases on disability have been produced. INEGI has also provided advisory services on the development of census GIS databases in several Latin American countries.

2.117. Among the factors that have facilitated the success of the Mexican census mapping project, the following stand out. INEGI has had a strong institutional commitment to the project. Within this large and well-funded statistical organization, an early decision was made to make use of emerging new technologies in the mapping field, not only for census mapping but also for other mapping applications. The close institutional ties between the statistical and mapping offices, which are housed within the same agency, has led to synergies that benefited both branches: the census offices obtained access to technology and advice from the mapping agency, which in turn can integrate socio-economic information with its own products. Furthermore, INEGI follows a long-term strategy of census mapping that is characterized by a continuous program of updating and refining the digital database. After the huge initial effort of digitizing hundreds of thousands of polygons and integrating these with tabular census data, continuous maintenance is relatively inexpensive and ensures a high level of quality for frequent census and other statistical operations. Finally, the availability of census data products in spatially referenced form has created new markets for census data in the country among private companies, educational institutions and researchers.

[a/] The present summary is based on Espejo (1996) and on personal communication with INEGI staff.

ii. Reference information in vector format

2.118. Enumeration area boundaries by themselves are not sufficient for the purposes of enumeration. Enumerator maps must show additional base map features that allow the enumerators to find their assigned region and to navigate within the census unit. Candidate features include the street network, railway lines, hydrological features such as rivers and lakes, major landmarks such as places of worship, schools, factories, or airports, settlements at small cartographic scales and individual buildings at large scales, and terrain information. Some of these features define natural boundaries for enumeration areas. For instance, defining EAs as city blocks that are delineated by streets makes it easier for the enumerator to identify the assigned census unit.

2.119. Which features should be captured digitally for enumerator map preparation depends on available resources. Some of the information might be available from other government agencies or from the private sector. Otherwise, the census organization must carefully weigh the advantages of having additional information available in digital form against the time and staff resources required to produce another geographic data layer.

2.120. During the conceptual database design stage, decisions must also be made on how features are represented in the database. For instance, streets can be digitized as double lines or as centre lines only. Houses can be represented by polygons that reflect their actual shape or by standardized symbols.

2.121. Resource needs also depend on the number and complexity of attributes that are collected for geographical features. For instance, a street network database for a city might consist simply of a set of lines, without any further characteristics. Storing information on street names, surface type, number of lanes, direction (one-way or two-way streets) or address ranges for each street segment

significantly increases the time required to complete a database. But this additional investment will also make the data set more useful for census purposes and for many other applications. The trade-offs must be weighed for each case separately.

iii. Reference information in scanned raster format

2.122. An alternative to the time-consuming development of a vector database of geographic reference features is to scan and georeference existing topographic maps onto which the EA boundaries can subsequently be printed. This has the disadvantage that changes in the background information—such as new roads—cannot be incorporated easily. Also, there is no attribute information recorded for background features that could be used to select or symbolize individual features. On the other hand, scanning topographic maps is considerably faster and less expensive than digitizing, and the cartographic design of topographic maps allows a clear representation of dense geographic information which is difficult to achieve with commercial GIS software.

iv. Recording enumerator area centroids only

2.123. Computer technology has enabled average users to utilize large, detailed geographic databases of enumeration areas only relatively recently. Before cheap computing power became widely available, some statistical offices used a simpler method for representing census information spatially. Instead of representing an enumeration area as a complete polygon, each EA was summarized by a representative point location—usually the centroid. An example is the population-weighted representative points of enumeration districts (EDs) (the smallest zone for which data are released) for the 1991 census of the United Kingdom (Openshaw, 1995). The point locations were determined by eye during census geography design at the Office of Population Censuses and Surveys.

2.124. The advantage of this approach is its simplicity, since a single point coordinate is used to represent each enumeration area. This results in small file sizes and fast display. The coordinates can be determined from available maps or collected in fieldwork, using a global positioning system. Census data can be linked to the centroids in the GIS database and displayed as point symbols. The disadvantage is that an EA centroid database is of no use to enumerators during census taking. It would therefore only be an added component in a traditional sketch map approach. Also, especially in rural areas where the size of enumeration areas varies greatly, a single coordinate does not provide sufficient information about the actual extent of each zone. Cartographic displays using point symbols can therefore be misleading.

v. Post-census mapping only

2.125. Few countries used a fully digital census mapping strategy in the 1990 census round. The number is likely to increase for the 2000 round, but only during following census rounds can we expect that the majority of countries will use GIS techniques from beginning to end of the census process. A large number of countries have, however, used digital mapping in their post-census activities to present results of the 1990 round of data collections and for data dissemination. For countries that have not used GIS in pre-census activities, post-census mapping at more aggregate levels (e.g., district, arrondisement or municipio) provides an opportunity to gain familiarity with the techniques, support presentation of census data and widen the user base of statistical information.

2.126. Post-census mapping requires far smaller resources than a complete enumeration area mapping, since usually only a few hundred administrative units need to be digitized. This activity can be performed centrally at the census office, which can then distribute the information to regional planning and administrative authorities.

vi. Mixed approach

2.127. In the light of the time it takes to develop a complete mapping program, national census offices may decide to select a gradual approach towards digital census mapping. A country may decide, for example, to use GIS to map enumeration areas in the largest cities only. For the rest of the country, traditional, manual techniques could be used. In future censuses, GIS will then also be used in these areas.

2.128. In some situations, it may be beneficial to use new technologies such as digital air photos or satellite imagery in remote areas where up-to-date maps are unavailable or fieldwork is difficult. In rapidly growing urban areas, remote sensing techniques also allow a census office to update city maps. New technologies can thus be useful to fill gaps that would be difficult to cover by traditional approaches. Another mixed approach could be to determine coordinates that define the outlines of EAs using global positioning systems and to add buildings, roads and other features useful for orientation in the EA by hand.

vii. Georeferenced address registry

2.129. Some countries go one step further than digitizing enumeration area or census block boundaries. Instead of producing maps of small reporting zones, they are

developing databases in which each building address is represented by a coordinate in a proper geographic reference system. An example is the *Address-Point* system developed and commercialized by the Ordnance Survey in the United Kingdom (Ordnance Survey 1993), which provides geographic precision—though not necessarily accuracy—at the submetre level. There are two ways of creating an address point database.

2.130. The first option is to collect a coordinate for each building in the country, either by digitizing from available small-scale topographic and city maps or by collecting the coordinates using field techniques. Statistics Canada, for example, has initiated pilot testing for activities in the 2001 census, in which census enumerators will collect a coordinate for each housing structure in the country, using a global positioning system (Li, 1997). The resulting georeferenced dwelling frame will be the most detailed reference system for census information in the country. The points that represent households for which census data are available can then be aggregated to any desirable statistical reporting zones, using simple GIS operations.

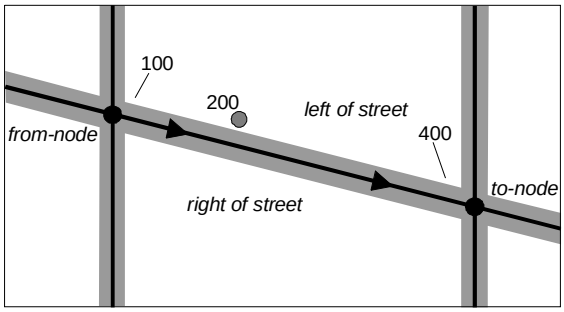
2.131. A second approach is possible where a comprehensive street network database and a master file of addresses of the population exist. A street network database consists of street or road segments. A street segment is a stretch of a street between two intersections, where an intersection is defined as the place where three or more street segments meet or where the street changes its name. In GIS, the streets are defined by lines representing the centre of the street, and the intersections are defined as nodes (see Figure II.7). In the internal geographic attribute table, a from-node and a to-node is listed for each line segment. Which is the from-node and which is the to-node is determined by the direction in which the line was digitized. Given the direction, it can be determined of which side of the street segment is left and which is right.

2.132. For each line segment in the street database, the range of address numbers needs to be recorded for both sides of the street. The attribute table thus has at least five fields, with a record for each street segment:

- The street name;
- The first address on the right side of the street;
- The last address on the right side;
- The first address on the left side of the street;
- The last address on the left side.

2.133. In most countries with street address systems, the numbers on one side of the street are even-numbered, and those on the other side are odd-numbered. With this information, GIS can locate any given address on the street network in a process known as *address matching* (sometimes also referred to as *geocoding*). Each street address from a list is evaluated and matched to a location on the corresponding street segment. The location is chosen based on the address number in proportion to the range of addresses on the street segment. For example, if the addresses on the street range from 100 to 400, an address value of 200 would be placed at a location one third of the length of the street segment Figure II.7). Determining the address location requires interpolation. It is thus not exact, but usually close enough for most purposes.

Figure II.7.: Address matching (geocoding)



2.134. The address geocoding system has a number of obvious advantages. Since the location of each street address, and therefore of each household, is known, a census office can reaggregate census information spatially to any new set of reporting zones—for example., postal codes, health districts or administrative units.

2.135. Development of a geocoded address database, however, requires a considerably larger investment than databases representing reporting zones. It is therefore usually only employed in countries where other authorities, such as the postal office, also have an interest in creating such a database. In some countries, it is the private sector that is creating street address databases for commercial—usually marketing—applications.

2.136. Automated address matching requires a comprehensive street database, with information about address ranges attached. This method will not work where address numbers are not assigned sequentially. In some countries, houses are numbered according to their building date, not according to their sequence along a street. Address matching that relies on interpolation of street addresses will therefore not be appropriate. In such situations, an explicit recording of each household's location, which results in a master address file, with the address and coordinate of each living quarter, is a better strategy, provided that the resources to compile such data

are available. While useful in an advanced setting, address geocoding is unlikely to be adopted for census cartographic work in many developing countries.

2.137. Georeferenced address registries or dwelling unit databases require special consideration of data confidentiality issues. Since every household can be identified by its coordinate, even if textual address information is unavailable, the census office must keep full control over the master database and release information only in aggregated anonymized form.

viii. Capturing related information

2.138. Census mapping requires considerable fieldwork, which usually implies that census cartographic staff visit all places in the country. This data collection process provides a unique opportunity to collect additional information, with only marginal extra effort. One useful output is a complete list of coordinates and names of all the villages and other settlements in the country. National mapping agencies produce such gazetteers, but these are usually updated irregularly and are therefore often out of date. A census office could therefore collaborate with the mapping agency to bring the gazetteer up to date and, at the same time, to check all coordinates, using traditional field techniques or global positioning systems.

2.139. In addition, with somewhat more effort, detailed inventories of the location and characteristics of service facilities can be created. Many government agencies require such information to study and plan the population's access to public services such as hospitals, schools or government offices. A geographic database of the location of such service centres in combination with spatially referenced census data, greatly expands the options available for analysis and policy planning.

(b) Implementation choices

2.140. After the scope of a digital census mapping project has been defined, some additional decisions need to be made concerning the implementation of the selected strategy.

i. Georeferenced versus non-georeferenced

2.141. What distinguishes GIS from computer graphics or CAD systems is the ability of spatial information systems to support consistent geographic referencing. That means that every geographic entity, such as an administrative unit, a village or a facility location, is defined by real-world geographic coordinates. Georeferencing allows the combination of spatial map data sets that come from different sources (e.g., districts and ecological regions) within one consistent framework, and also enables the user to merge individual subsets of a larger data set. For example, district boundaries for several provinces that had been created separately can be appended to produce a national-level data set.

2.142. In principle, georeferencing of census enumerator maps is not necessary for the purposes of conducting a census. Traditionally, hand-drawn sketch maps have been used, which provide sufficient information for each individual enumerator to perform his or her duties. These sketch maps are not usually combined to produce maps that cover a larger region, so that it is not important that the boundaries of adjacent enumeration areas that are drawn on separate sketch maps match perfectly. Sketch maps can, of course, also be drawn with a computer-based graphics package. In that case, each sketch map is referenced in its own relative coordinate system, which is measured in centimetres or inches from an origin in the lower left corner of each map page. Producing sketch maps using graphics packages will facilitate corrections and updates and makes it easier to produce multiple copies. Compared to GIS software, graphics packages are also usually cheaper, easier to learn and require less powerful computers.

2.143. Using a GIS or desktop mapping package will increase the costs of producing census cartography. Most importantly, georeferencing requires some expertise in the handling of geographic coordinates (see annex II). This increases training requirements, as well as the time required to complete the census cartography. Also, not all desktop mapping packages and few graphics or CAD systems provide the functions required for georeferencing. These functions are used to define and change the cartographic projection of a map and to remove the distortions present in hand-drawn sketch maps.

ii. New delineation versus conversion of existing delineation

2.144. The national census organization must also decide whether to rely on existing census cartographic products, such as sketch maps from a previous census, or whether a complete new delineation of enumeration areas will be created. In practice, in most cases, often a mixture of existing map data sources and fieldwork for updating and cross-checking will be used. Section D below discusses techniques for converting existing hard-copy data sources, as well as modern field techniques.

iii. In-house development versus outsourcing of cartographic work

2.145. A statistical office can reduce or shift the costs involved in setting up a GIS program in several ways. The most effective means is cost sharing with other agencies in

the country. The example of statistical offices in Latin America, which are often located in the same umbrella organization as the national geographic or mapping agency has shown the synergy effects that can be realized from close collaboration. The statistical office benefits from the mapping and GIS capabilities in the geographic division, while the mapping division can integrate spatially referenced census information in its cartographic work and products. Even where such institutionalized links do not exist, close collaboration among agencies with similar data needs will be beneficial. For example, a statistical office could coordinate data collection efforts or data purchases with the planning, education or natural resources departments. This can significantly reduce the cost of acquiring, for instance, remotely sensed data.

2.146. An alternative is to assign the entire cartographic process to another government agency or to a private sector company. For example, for its 1991 census the Australian Bureau of Statistics worked together with a private company that produced digital enumeration area maps for the entire country. Agreements between the company and the statistical office guide the use and further commercialization of the data.

2.147. Outsourcing raises many of the issues discussed in section 3 (b) on institutional cooperation. The advantage for the statistical office is the reduced investment in training and equipment, and the time savings in getting immediate access to extensive GIS expertise. The disadvantages are the loss of control over the cartographic process, the fact that no in-house expertise is developed and, possibly, higher costs in the long run as the agency becomes dependent on outside suppliers. In some countries, it may also be difficult to find a domestic company that is able to provide services at the scale necessary to perform a large census mapping project. In practice, a mix of in-house activities and outside consulting services will usually be the most appropriate approach.

iv. The importance of risk avoidance

2.148. In choosing a suitable census mapping approach, a proper risk management strategy must also be adopted. Since the entire census depends on the timely completion of the enumeration area mapping program, there must be a certain level of redundancy and back-up plans in census mapping operations. At its most cautious, this may require a parallel digital and manual mapping approach. This dual strategy can be followed until there is complete assurance that production of digital maps will be on schedule.

2.149. Other risk minimizing approaches are to apply digital mapping in one region first before expanding the program to all regions—this would be similar to an extended pilot phase—or to limit digital approaches to only selected aspects of the census mapping process. For instance, field mapping could be done manually, with subsequent digitizing of sketch maps, rather than relying on digital field mapping techniques from the beginning. While the issue of risk avoidance should not prevent the adoption of innovative mapping techniques, the importance of being able to complete cartographic work on schedule is the most important consideration in selecting an implementation strategy.

(c) Definition of the geographic information system database structure

i. Relational databases

2.150. Before discussing specific structures of the census GIS database, the concepts of relational databases, which are used by most GIS packages, will be reviewed. 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.

2.151. In a geographic context, an *entity* can be administrative or census units, or any other spatial feature for which characteristics will be compiled. For example, an entity might represent the feature “enumeration area” (see Figure II.8). 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 that 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 area (CLA) that the EA is assigned to and so on. It should be noted 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 II.8. Example of an entity table – enumeration area

Entity: enumeration areas

Type (attributes)

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
…	…	…	…

Instances

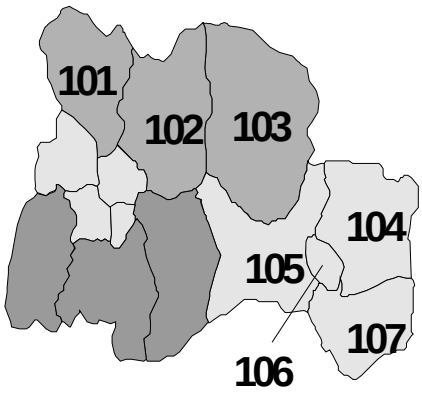
Primary key

2.152. 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 CLA table.

2.153. The process of designing a relational database structure through a series of steps is called *normalization*. The outcome is a database with minimum redundancy. In other words, the data are organized in a number of tables so that values that are repeated many times are avoided. This reduces storage space and avoids errors that might be introduced in standard database operations such as insertion, deletion or updates.

Figure II.9. 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
…	…	…

2.154. Figure II.9 illustrates the difference between a simple data table and its normalized form, using an example of a district database. In the first instance, the information for the provinces is repeated for each district in the province. This not only wastes storage space, it also makes it more difficult to update or change information for provinces. The values would need to be replaced for each individual district. In the normalized database structure, the name of the province has been replaced by a more compact numeric code, which provides the link to a second table. Here, the province code becomes the primary key for the province information that includes the province name, population and total fertility rate. After joining the two databases temporarily by means of the province code, province information can be accessed for each instance in the district table.

2.155. Defining a clean database structure is not a trivial task. Some database management programs provide normalization functions that automatically create a relational database structure. However, this is usually not a good substitute for a comprehensive design of the overall database. The entity-relationship model is described in more detail in Hohl (1998) in the context of GIS data conversion. Batini and others (1992) provide a more generic and comprehensive introduction.

ii. Components of a census database

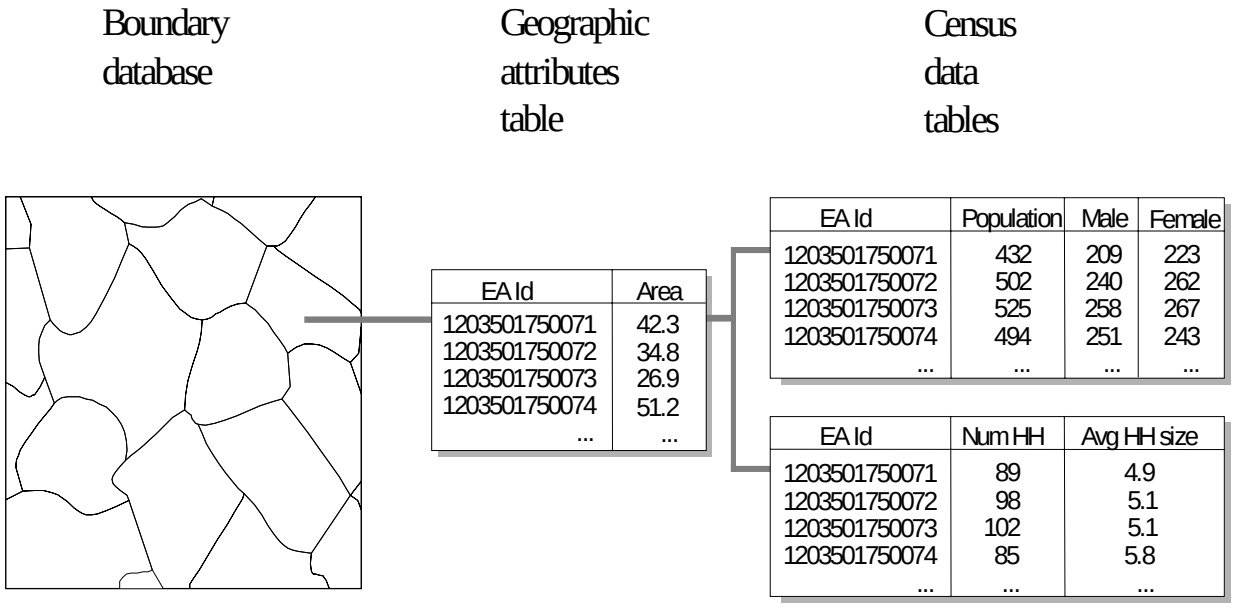
2.156. A comprehensive census GIS database consists of a digital map of census enumeration areas and, in most instances, a series of base map layers that provide the context and orientation in the final enumerator maps. Base data layers might be roads, rivers, buildings or settlements. Each of these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

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

2.158. The first step is to think about what the final products will look like. The complete digital enumeration area database, for example, will likely consist of the following components (see Figure II.10):

- The spatial *boundary database*, consisting of area features (polygons) that represent the census units;
- The *geographic attributes table*. A database file which is linked internally to the spatial database and contains one record for each polygon. This table contains the unique identifier for each census unit and possibly some additional static, that is., unchanging, variables such as the unit's area in square kilometres;
- The *census data tables*, containing non-spatial attributes—that is, the census indicators for the spatial census units. Each of these files must contain the unique identifier of the census unit, which provides the link to the corresponding polygon attribute table records. There will be one record for each census unit.

Figure II.10. Components of a digital spatial census database



2.159. The boundary database and geographic attributes table are tightly linked – essentially, they represent one data set. During census planning, some basic-census related information such as housing unit or population estimates and documentation information will be compiled for each enumeration area. This external information about the census units will be stored in separate data tables in a generic database management system. From there, it can be linked as needed to the boundary data through the common identifier—the EA code—in the geographic attributes table. Similarly, after census completion, the census information is stored separately in a database management system. To create thematic maps of census results, the boundary and census data are then linked via the unique identifiers in the polygon attribute table. Clearly, to ensure that the census databases that are the product of the data entry and tabulation program will match the geographic boundary files, close cooperation between the census cartographic and data-processing sections is required.

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

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

2.162. At certain stages in between and during census cycles, benchmark data sets should be created. For instance, there should be a unique version of the country's census map database that matches each data collection effort or related statistical application. Separate aggregated boundary data sets can be produced for each statistical reporting unit for which data are required. These benchmark data sets should be permanently archived. Thus, benchmark data sets created from the same master database may exist for a census in 1995, for a large survey in 1997 and for an election in 1998.

iii. Definition of database content (data modelling)

2.163. 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.

2.164. 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.

2.165. As discussed earlier, many national and international agencies have already been active in developing generic data models for spatial information as part of a national geographic data infrastructure (sometimes also known as a geomatics infrastructure). Often, a census office will be able to simply adapt a national spatial data 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.

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

(d) Metadata development

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

2.168. 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 tasks:

- 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.

2.169. Obviously, what different data producers consider essential metadata can differ widely. Many countries have therefore started the development of general geographic metadata standards. These aim at unifying the conventions for documenting spatial information. They therefore support the development of a national spatial data infrastructure by facilitating spatial data exchange and integration. At the international level, several organizations attempt to coordinate the development of spatial metadata standards among groups of countries. Among these are the ISO 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).

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

2.171. An example of a well-developed and widely used metadata standard is the Content Standards for Digital Geospatial Metadata (CSDGM), developed by the National Geographic Data Committee in the United States (www.fdgc.gov). It serves as an illustration of the types of information contained in a metadata database. The complete standard is 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 Bureau of the Census

(www.census.gov/geo/www/
standards/scdd; see FGDC, 1997b). Only discuss the main components of the metadata definition are discussed in the paragraph below.

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

2.173. The main components of the standard are:

- *Identification information,* including the data set title, area covered, keywords, purpose, abstract, and access and use restrictions;
- *Data quality information,* such as horizontal and vertical accuracy assessment, logical consistency, semantic accuracy, temporal information, data set completeness and lineage. Lineage includes data sources used to produce the data set, as well as processing steps and intermediate products;
- *Spatial data organization information, which* 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,* which contains detailed definitions of the attributes of the data set including the attribute data types, allowable values and definitions. This is largely the same information that is contained in a data dictionary as described earlier;
- *Distribution Information,* including the data distributor, file format of data, off-line media types, on-line link to data, fees and order process;
- *Metadata reference information, which* provides information about the metadata itself, most importantly, who created the metadata and when.

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

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

2.175. 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 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.

2.176. The 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 simply 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.

(e) Data quality issues

i. Accuracy requirements

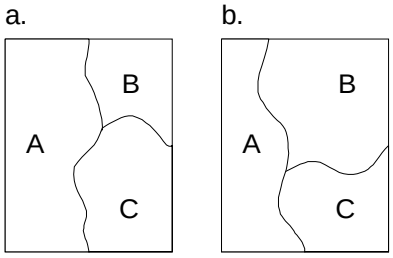
2.177. The development of acceptable data accuracy standards is perhaps one of the most important tasks in planning a digital database development project. In many fields such as utilities and facilities management, terrain or hydrological mapping, accuracy database standards exist that can be adopted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

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

2.179. Geographical data accuracy can be divided into *logical* and *positional* accuracy. Positional accuracy is sometimes also called absolute accuracy. Logical accuracy refers to the integrity of relationships among geographic features. For instance, a road in one GIS database layer must connect to a bridge in another layer. A river stored in a hydrological database that defines the boundary between two administrative units should coincide with the boundary between those units. And, a town represented as a point in one GIS database should fall into its corresponding administrative unit in another GIS layer. The same logical relationships can be represented correctly in different maps that have very different appearance. For instance, in Figure II.11, the two maps correctly represent the neighbourhood relationships between three administrative units.

Figure II.11. Logical accuracy

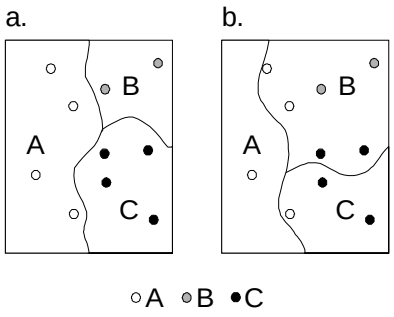


2.180. Positional accuracy, in contrast, maintains that the coordinates of features in the GIS database are correct relative to their true positions on the earth's surface. This means that cartographic measurements must be conducted with a sufficient degree of precision, using accurate measurement devices 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.

2.181. 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.

2.182. Figure II.12, 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 II.12-b, in contrast, while logically accurate, has a low degree of positional accuracy. Some of the accurately measured GPS points, therefore, fall into the wrong administrative units. This will lead to incorrect results when survey responses are aggregated by administrative unit.

Figure II.12: Problems if positional accuracy is not maintained



2.183. 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).

2.184. 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.

2.185. Although a high degree of positional accuracy is desirable, accuracy standards that are too limited will lead to increased costs, exaggerated user expectations and possibly frustrations among cartographic staff who may not be able to attain goals that have been set too high. Accuracy standards that are too low may lead to products that are of insufficient quality. Users may either reject the product if they are aware of its limitations, or they may use it with an overstated level of confidence 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.

2.186. 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.

ii. Quality control

2.187. 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, 1998) stress the importance of quality control and contain an overview of these issues in the census process. These general concepts also apply to census mapping.

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

2.189. 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 monotonic 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.

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

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

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

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

2.193. A consistent set of quality control procedures should result in an end product of acceptable accuracy. However, in most projects, a final step known as quality assurance is usually added, which consists of another round of checking and a last process of problem resolution. Quality assurance is discussed in section 0.

(f) Tiling of national territory into operational zones

2.194. A complete digital enumeration area database will consist of thousands of units. For larger countries, it is not usually practical to store all EA polygons in the same physical data layer. Instead, the national territory can be divided into operational zones. In a decentralized census administrative structure, different regional offices and different operators within each regional office can thus work on separate parts of the database simultaneously. Provided that consistency between the boundaries of the subsections of the national database has been enforced, the separate pieces can be combined at a later stage to produce district, province or national-level maps. This process will, however, require some edge-matching, which involves the manual linking of connected features that cross two or more tiles.

2.195. 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.

(g) The digital administrative base map

2.196. If a decentralized approach is chosen, the national census office should first create a national boundary template for the major administrative levels in the country. For example, the census office should create, obtain or commission a set of digital spatial boundaries of provinces, districts and, ideally, also subdistricts. 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.

2.197. Such boundaries may have already been produced in digital form by the national mapping agency. In that case, they will represent the officially recognized digital administrative base map for the country (see the discussion about national spatial data infrastructures in section 3 0 above. Codes used in the administrative base should correspond to the codes used in the census database.

2.198. 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 have been 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.

(h) Dealing with disjoint area units

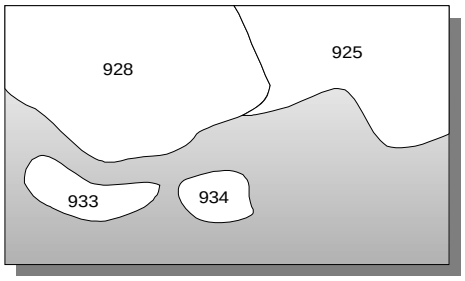
2.199. Administrative units are frequently split into separate, distinct spatial units or polygons. For example, a district may consist of an area on the mainland and a number of islands. For census data processing, this is not a problem since there will be only one record in each census data table that applies to the district. In the geographic attributes 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 by the way the geographic attributes table. In a relational database

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

2.200. 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 attributes 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.

2.201. 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 attributes table. This field will assume the value of one for the largest polygon belonging to the district and zero for the smaller ones. Before summing or averaging any attribute value, the user can first select only the polygons with a value of one in this field. An additional field could be added that contains the number of polygons belonging to the same unit. This information can be generated quickly, using the frequency or cross-tabulation feature of the GIS package.

Figure II.13. Dealing with administrative units consisting of several polygon



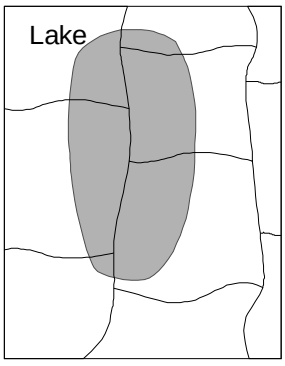
Internal identifier	Admin. identifier	Admin. name	Flag
925	02015	Kalana	1
928	02012	Bagor	1
933	02012	Bagor	0
934	02012	Bagor	0

(i) Computing areas

2.202. 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.

2.203. 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—that is., 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 illustration in Figure II.14). 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.

Figure II.14. A lake covering a large area in several administrative units



2.204. 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*— that is the total area minus the area

covered by water bodies and possibly other uninhabited areas such as protected conservation areas. Some countries also report the area of agricultural land. This allows the users to compute agricultural population densities or vice versa, the number of hectares of agricultural land available per inhabitant in the district. These area figures can be computed quite easily in a GIS, using appropriate geographic data layers, subject to the caveats relating to map generalization mentioned above. In any case, it is important that the definitions of the net areas are well documented.

2.205. 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.

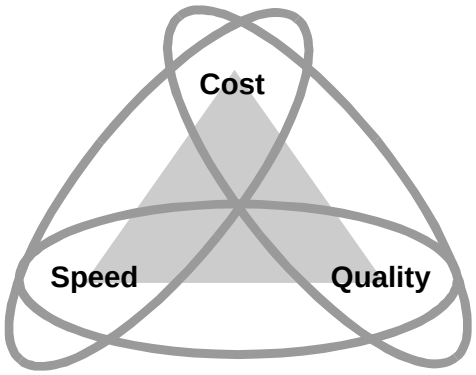
D. Digital map database development

1. Overview

2.206. The development of the digital census database will be based on two data sources: the conversion and integration of existing map products, which may be in hard-copy or digital form, and the collection of additional data, using fieldwork, air photos or satellite images. Collectively, the term *data conversion* is used to refer to these steps (see Montgomery and Schuch, 1994: Hohl, 1998).

2.207. The best strategy for data conversion depends on many factors, including data availability and time and resource constraints. There will always be a trade-off between the cost of a project, the amount of time required to complete data conversion and the quality of the final product (Figure II.15). It is usually only possible to 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. Good data can be produced cheaply, but this will take a long time. Or, a database can be developed quickly and cheaply, but the quality of the resulting product will be low.

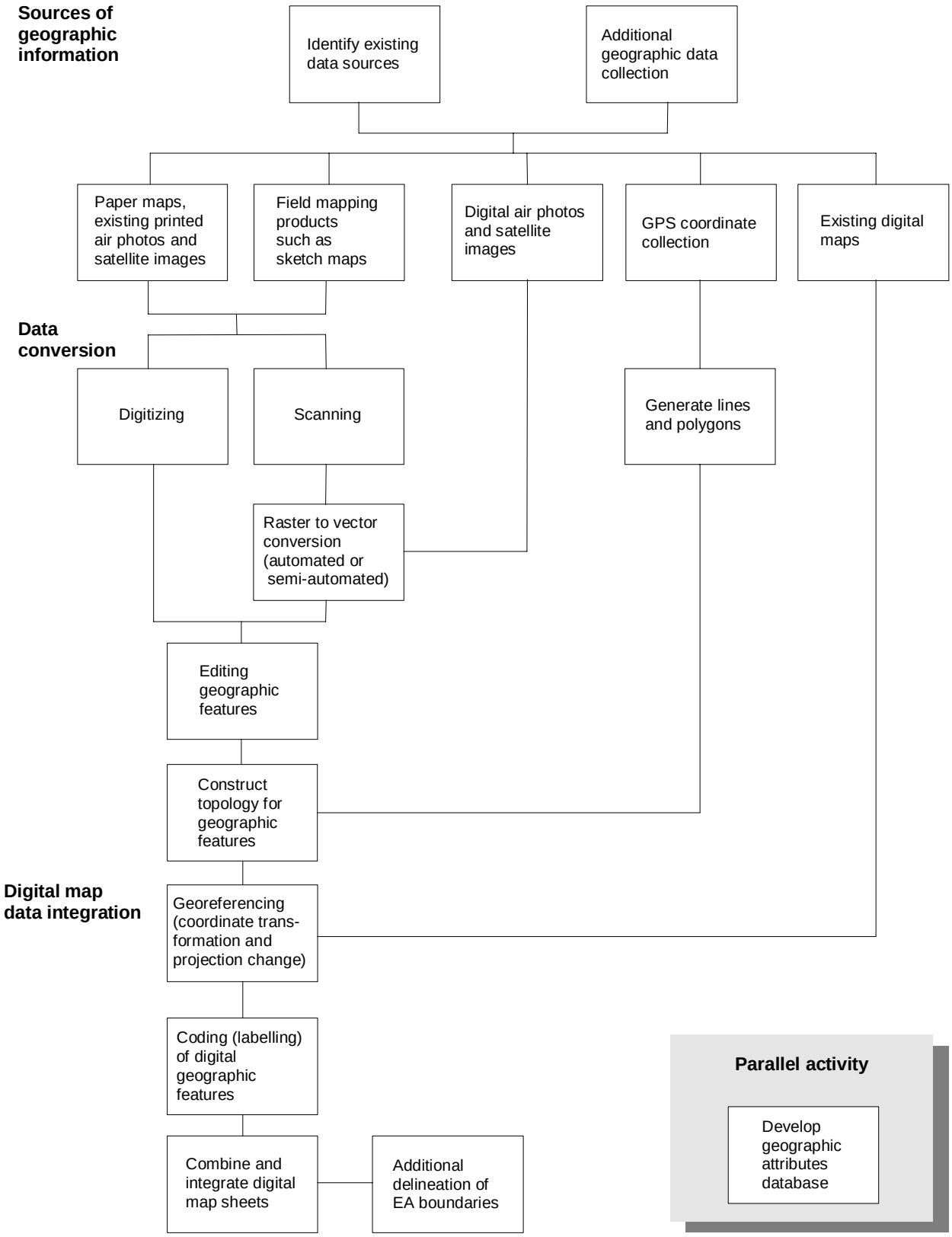
Figure II.15. Trade-offs in the data conversion process
(after Hohl, 1998)



2.208. Figure II.16 outlines the basic steps in the data conversion process that leads 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.

2.209. 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 digitizing—tracing the features with a mouse-like cursor—or by scanning with subsequent image to vector conversion. Although digitizing and scanning 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 II.16. Stages in the census GIS database development



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

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

2.212. The flow chart in Figure II.16 shows only one of many possible sequences in the data conversion. EA boundaries, in particular, can be delineated at several points during the process. For instance, scanned and properly georeferenced air photos 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.

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

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

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

2. Cartographic data sources for enumeration area mapping (secondary data acquisition)

(a) Types of maps required

2.216. In nearly all cases, a census cartographic program 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 the following types of maps (see, also, BUCEN, 1978, chap. 2):

- National overview maps, usually at scales between 1:250,000 and 1:5,000,000, depending on the size of the country. These maps should show major civil divisions, the location of urban areas, and major physical features such as important roads, rivers, lakes, elevation, and special points of reference. These maps are used for planning purposes;
- Topographic maps at large and medium cartographic scales. The availability of maps at these scales will vary by country (see Boehme, 1991: and Larsgaard, 1993). While some countries have complete coverage at 1:25,000 or 1:50,000, the largest complete map series in others is only 1:100,000 or 1:250,000 scale;
- Town and city maps at large cartographic scales, showing roads, city blocks, parks and so on;
- Maps of administrative units at all levels of civil division;
- Thematic maps showing population distribution for previous census dates, or any features that may be useful for census mapping.

2.217. For incorporation in a GIS database, ideally, these maps should all have comprehensive

documentation. This includes the geographic referencing information, including the map scale, projection and geographic datum, the map compilation date, compiling agency and complete legend. However, even maps that are not properly georeferenced are useful if they show information relevant to census mapping. 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.

(b) Inventory of existing sources

2.218. All maps that have been obtained should be well documented and organized according to the organization of the census mapping program—that is, by census region or district. BUCEN (1978, chap. 6) presents a discussion of map inventories and the development of a map library.

2.219. 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 is absent, it is often not possible to determine the correct projection information, and data quality is difficult to evaluate. The following agencies and institutions should be contacted to see whether they can contribute useful hard-copy or digital maps:

- National geographic institute/mapping agency. This is the lead agency in the country concerned with mapping. However, in some countries, the mapping agency is lacking the resources necessary to produce topographic maps at large cartographic scales or to convert maps into digital databases;
- Military mapping services. In some countries, the main mapping organization is part of the military. Military mapping organizations are often strong in aerial photography and in the interpretation of remotely sensed data;
- Province, district and municipal governments. Local government organizations increasingly use GIS to manage information about transportation, social services, utility services and planning relevant information;
- Various government or private organizations dealing with spatial data:
    - Geological or hydrological survey authority;
    - Environmental protection authority;
    - Transport authority;
    - Utility and communication sector companies;
    - Land titling agencies;
    - Donor activities. Project-level activities by multinational or bilateral aid organizations sometimes include mapping components. Such projects often have the means to purchase and analyse remotely sensed data or aerial photographs, which can be of great use to the mapping agency.

(c) Importing existing digital data

2.220. Direct import of digital data is in most cases the easiest form of digital spatial data conversion. Unfortunately, to date no universally supported spatial data transfer standard has emerged. Data transfer therefore relies on the exchange of data in mostly proprietary file formats, using the import/export functions of commercial GIS packages.

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

2.222. Problems can be reduced if the census cartographic agency employs a widely used, comprehensive GIS package. High-end systems are more likely to provide import functions for a large number of exchange formats. 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.

2.223. Apart from problems in converting the data files from one format to another, the most often encountered difficulty in using existing digital data is insufficient or absent metadata. Without such information, it is difficult to assess the quality of the digital information. Even worse, missing information about the geographic reference framework might make it impossible to convert data from the external data set's coordinate system to the one used by the census organization. Similarly, a missing code book or data

dictionary will make it difficult to interpret the geographic and data attributes included in the GIS data set's attribute tables. When data are procured from external sources, the census office should therefore always insist that extensive documentation is provided.

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

3. Additional geographic data collection (primary data acquisition)

(a) Field techniques overview

2.225. Despite the advent of technologies such as global positioning systems, traditional field data collection skills continue to be useful for census cartographic work. Often, maps will need to be updated by trained census cartographic staff in the field. This may involve the preparation of sketch maps, which are later georeferenced using information derived by GPS. Traditional field techniques for census applications are described extensively in BUCEN (1978, chap. 5) and will therefore not be discussed in the present handbook.

(b) Global positioning systems

2.226. GPS technology has revolutionized field mapping in recent years. As the prices of GPS receivers have dropped, GPS methods have been integrated in many applications areas. The largest user groups are in the fields of utilities management, surveying and navigation. But GPS has also contributed to improved field research in areas such as biology, forestry and geology, and also finds increasing application in epidemiology and population studies. GPS is also becoming a major tool in census cartographic applications.

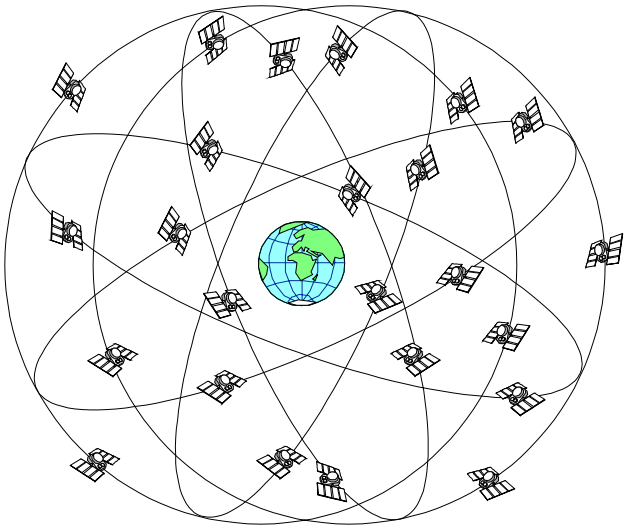
2.227. Most of the discussion refers to the United States system commonly referred to as GPS. This is the system that is most widely used and for which a large commercial market of receiver manufacturers and surveying services has developed. A second satellite positioning system, the Russian system known as GLONASS, is discussed below.

i. How global positioning systems work

2.228. GPS receivers collect the signals transmitted from a system of 24 satellites—21 active satellites and three spares (see Figure II.17; see Leick, 1995: French, 1996: Schmidt, 1996: Kennedy, 1996: and Dana, 1997). The system, which is known as 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, five to eight GPS satellites are within the “field of view” of a user on the earth’s surface.

2.229. The position on the earth’s surface is determined by measuring the distance from several satellites. The GPS satellite and the receiver each produce a precisely synchronized signal (a so-called pseudo-random code). Synchronization is made possible by very precise clocks on the satellite and in the receiver. The receiver can measure the lag between the internal signal and the signal received from the satellite. That 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, the lag time simply needs to be multiplied by the speed of light to obtain the distance.

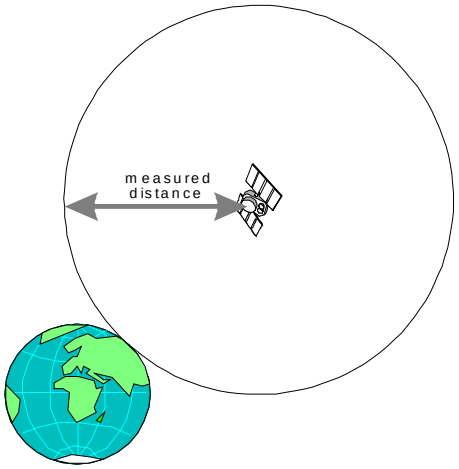
Figure II.17. The global positioning system



2.230. Once the distance from several satellites is known, position can be determined by trilateration. Since it is difficult to show this graphically in three dimensions, the following figures show the principle in simplified form in two dimensions. In the first figure (Figure II.18a), we have only one satellite above the earth’s surface. The circle around the satellite has a radius corresponding to the measured distance between the GPS user and the satellite. Of course, at this point we do not know exactly where on the circle we are located.

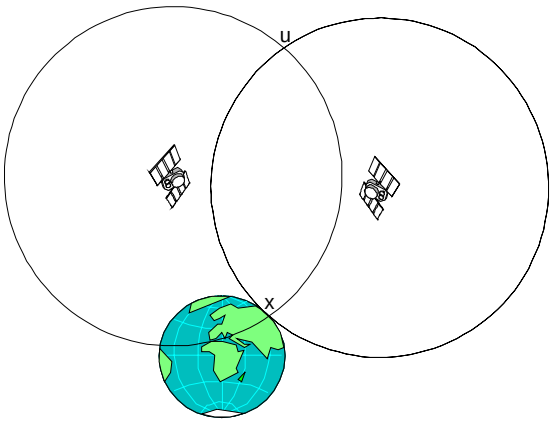
Figure II.18. How GPS determines a location's coordinates

a.



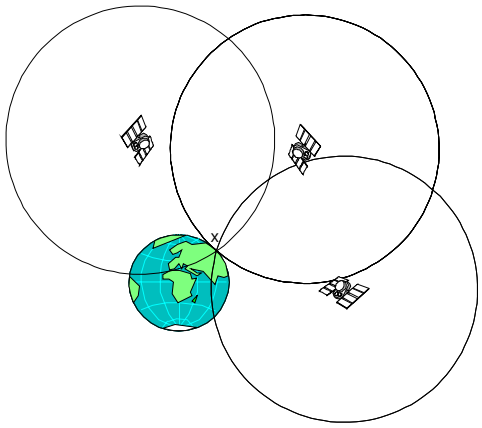
2.231. If we obtain a distance measurement from a second satellite, we can narrow our position down to the two points, x and u, at which the two circles intersect. Normally, these two measurements should be enough, because one of the candidate positions is likely to be very unrealistic.

b.



2.232. However, to confirm our exact position, we should determine the distance from a third satellite. The distance circles around all three satellites intersect at only one point, which is our true position.

c.



2.233. 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, in contrast, is very precise.

ii. Global positioning system accuracy

2.234. 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 GPS, the recorded position is accurate within 15 to 100 metres for civilian applications. According to the more precise United States Department of Defense specifications, there is a 95 per cent chance that the recorded position is accurate within 100 metres and a 50 per cent chance that it is accurate within 40 metres. In practice, an accuracy of 30 to 50 metres can typically be achieved. Altitude information is somewhat less reliable than latitude and longitude. Here, a rule of thumb is an accuracy of about 80 metres.

2.235. 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. Fieldworkers can identify the optimal periods for data collection by

consulting an almanac that provides a detailed schedule for all GPS satellites. Additional sources of error are atmospheric disturbances that modify the signal as it travels through the atmosphere, and so-called multi-path error caused by scattering of the signals from buildings or other solid objects. Such errors represent more or less random noise—random, short-term fluctuation of the position (Lang, 1997).

2.236. However, these errors add up to only about a fourth of the total error for standard GPS receivers. By far the greatest source of error is the so-called selective availability. To prevent hostile countries from using high-precision GPS, the United States Department of Defense deliberately introduces noise to the signal. Only the military has access to the correction information. Selective availability, is scheduled to be phased out in the next few years, since, as described below, its purpose is defeated by various ways of improving GPS signal accuracy. Yet, even without selective availability, the accuracy of GPS coordinates will not be perfect.

2.237. Repeated readings of GPS coordinates will not necessarily improve the coordinate estimate. This is because the error introduced by selective availability is not randomly scattered around the true position and because most systems use some form of averaging of repeated measurements, which reduces the variance of measured positions. Turning GPS off and on after each measurement will provide a better indication of available accuracy (see Lange, 1997). To obtain more accurate positions, one would need to average coordinate readings over a long time period – that is, more than 24 hours. In practice, there are better options for improving GPS coordinates.

iii. Differential global positioning systems

2.238. For applications requiring higher accuracy, differential global positioning systems (DGPS) use correction information transmitted from a base station with precisely known coordinates to correct the satellite signals. The signals received by the DGPS base station and the mobile GPS unit are subject to the same errors. The DGPS base station can therefore compare the difference between the computed position and its known correct position and send this information to the mobile unit (see Fig II.19). The accuracy that can be achieved with DGPS depends on the system and coordinate collection procedure. Accuracy of about 3 to 10 m can be achieved with quite affordable hardware and shorter observation times. More expensive systems and longer data collection for each coordinate reading can yield sub metre accuracy.

2.239. 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 that cost several thousand dollars can be converted into DGPS base stations that broadcast correction information. 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 by way of geostationary satellites, for example for aircraft navigation. In the future, these DGPS options are likely to be available to the average user, so that DGPS correction information will be available everywhere at any time.

2.240. Post-processing of GPS coordinates is often a less complicated and less expensive option. Here, the user collects coordinates with a standard GPS receiver. For each coordinate the time and satellites used are recorded in the receivers memory. Back in the office, the user can download correction information for that time period, and apply the correction factors to all collected coordinates. Correction data files are available from a number of commercial or public sources in many countries. 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, several base stations may have to be set up.

iv. Global Orbiting Navigation Satellite System

2.241. The Russian counterpart of GPS is GLONASS, which is operated by the Ministry of Defence of the Russian Federation. This system is also based on a constellation of 24 active satellites that circle the earth in three orbital planes (as opposed to six for GPS). The characteristics of both systems are very similar. One difference is that GLONASS does not employ selective availability for civilian users, which means that GLONASS provides higher accuracy positions than GPS in an autonomous (i.e., non-differential) mode. Although the GLONASS project started in 1982, the full constellation was only completed in early 1996. Since then, the launch schedule for new satellites has been delayed and, owing to the break-down of several satellites, the number of usable satellites has been between 11 and 16.

2.242. Dedicated GLONASS receivers are not in widespread use. However, several academic research institutes and private companies have developed positioning systems that combine the signals of both

GPS and GLONASS. The use of both systems means that, at any given time and place, there will be more satellites within the user's field of view than if only one of the systems is used. This is especially important in areas where part of the horizon is obstructed such as urban canyons, mountainous areas or under trees. Table II.3 shows that the combination of GPS and GLONASS improves the accuracy of position measurements considerably compared to GPS with selective availability. Although the figures suggest only marginal improvement compared to GLONASS by itself, the fact that only a reduced number of GLONASS satellites is currently available means that the combined system will produce more reliable results.

Figure II. 19. Differential global positioning systems

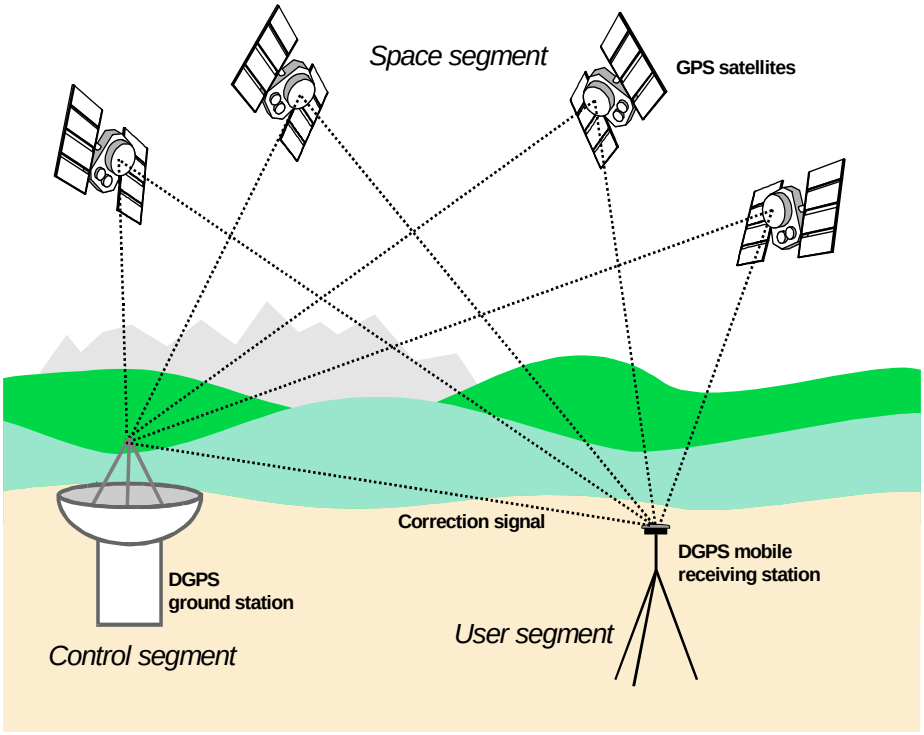


Table II.3.: Accuracy of GPS and GLONASS positioning

	Horizontal error (metres)		Vertical Error (metres)
	(50%)	(95%)	(95%)
GPS (SA off)	6	18	34
GPS (SA on)	25	72	135
GLONASS	7-10	26	45
GPS and GLONASS	9	20	38

(Source: Misra, 1993, and Hall and others 1997)

v. Selecting a global positioning systems unit

2.243. 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 plot simple 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, cigarette lighter adapters for cars are useful), coordinate storage capacity, and the ease of transferring stored coordinates to a laptop or desktop computer.

2.244. Most vendors offer integrated products that combine a GPS receiver with a palmtop or notebook computer so that the captured coordinates can be plotted on the screen immediately, either in isolation or on a digital base map. For census applications, the equipment required for a large number of fieldworkers would likely be beyond the resources of a census project. Storing the coordinates in the system and possible manual recording on data sheets as a back-up provide a lower-cost alternative.

vi. Global positioning systems in census mapping applications

2.245. GPS technology has obvious application in any kind of mapping activity, including the preparation of enumerator maps for census activities (e.g., Tripathi 1995). With DGPS accurate geographical positions of enumeration area boundaries can be determined with GPS, 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 digital mapping system or GIS, and can be combined with existing, georeferenced information.

2.246. The exact way in which GPS coordinates are used in census mapping will vary depending on the chosen census mapping strategy. GPS can be used in point mode to collect a coordinate, for example, for each building in a village or for each intersection in the street network of a town. Available maps or sketch maps drawn during data collection will help to interpret the coordinate information back in the office. A second possibility is to collect GPS coordinates in stream-mode, where the system records coordinates at regular intervals. This way, line features can be recorded automatically by walking along a road or travelling in a vehicle or on a bicycle. This is a cost-effective way of creating a street or road network database (see Box II.3: Eritrea example), although it will depend on the chosen data quality standards whether the accuracy of the resulting lines is sufficient.

2.247. In the application of GPS, problems can, of course, arise. In dense urban settings, the possible error of standard GPS (up to 100 metres) is not sufficient to define adjacent EAs accurately. In those cases, DGPS must be used or GPS readings have to be cross-checked with additional data sources such as published maps, aerial photographs or even sketch maps produced during fieldwork. Some cities, for example, Doha, have developed a system of GPS base stations that support very high accuracy mapping using DGPS. But in many developing countries, such networks do not yet exist. 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.

vii. Integrated field mapping systems

2.248. Field data collection in the utility sector and other mapping applications now relies heavily on GIS. In many of these applications, GPS is integrated with a portable computer or personal digital assistant. Coordinates are captured and immediately displayed on the portable computer screen. If a digital base map is available, the coordinates can be displayed on top. Field staff can add any required attribute information and store these data in a GIS database. This GIS information can then be incorporated in a GIS at the home office. Given that notebook computers and other portable computing devices are constantly becoming cheaper, integrated field mapping systems may soon become a viable option for field data collection for census purposes. Likewise, GPS receivers continue to shrink in size and cost. At the time of writing, the first GPS receiver to be integrated in a wristwatch has been announced. Widespread integration of GPS in cars and electronic equipment is likely.

viii. Summary: advantages and disadvantages of global positioning systems

2.249. The advantages of GPS include the following:

- Fairly inexpensive, easy-to-use field data collection. Modern units require little training for proper use.
- Collected data can be read directly into GIS databases, making intermediate data entry or data conversion steps unnecessary;
- Worldwide availability;
- Sufficient accuracy for many census mapping applications—high accuracy achievable with differential correction.

2.250. The disadvantages are as follows:

- The signal may be obstructed in dense urban or wooded areas;
- Standard GPS accuracy may be insufficient in urban areas and for capturing linear features, making differential techniques necessary;
- DGPS is more expensive, requires more time in field data collection and more complex post-processing to obtain more accurate information;
- A large number of GPS units may be required for only a short period of data collection.

Box II.3. Census mapping in Eritrea

2.251. GPS is being used extensively in the preparation of enumerator maps for the 2000 round census in Eritrea (see Eritrea National Statistical Office, 1996). In collaboration with experts from Statistics Canada, the national statistical office decided on a digital approach for census mapping. Basic features such as transportation hydrography, spot heights and ridge lines were digitized manually from available 1:100,000 scale maps. Since reliable village and town maps were not available, GPS was used to record a coordinate for each village that consisted of only one EA (fewer than 100 households).

2.252. For larger villages and towns, census field staff walked along the centreline of all pathways and streets in the settlement while recording GPS coordinates in stream-mode, where the instrument records positions at regular time intervals automatically. At the same time, basic sketch maps were drawn by hand, which helped to link dwelling counts to mapped village blocks.

2.253. To facilitate later EA map interpretation by enumerators, the location of landmarks in the village (places of worship, schools, and so on) were also recorded. This was done by collecting coordinate information on the street network nearest to the landmark and noting the offset and direction from that point to the landmark. In bigger towns, city blocks were recorded, using vehicle-based GPS in stream-mode.

2.254. The cartographic field staff collected coordinate data using standard, inexpensive GPS receivers. These data were backed up on inexpensive laptop computers and on diskettes after each day of work. An operational problem was the recharging of the GPS batteries in remote areas. All GPS readings were post-corrected with information collected by a base station located on the roof of the National Statistical Office. Locating the base station at the census office's headquarters had several advantages. It ensured a fixed, permanent and verified source of correction data. Continuous operation was ensured by having a reliable source of electricity, and the installation on the roof of the building provided a controlled and secure operating environment. The location of the base station a few hundred kilometres away from some of the fieldwork locations did not introduce serious inaccuracies for census purposes, although several base stations may be required in larger countries.

2.255. Interestingly, the GPS coordinates for some villages pointed to inconsistencies in the administrative structure of the country. Some villages turned out to be located outside the boundaries of the administrative units to which they had been assigned. Such problems highlight the importance of close cooperation between the census organization and local administrative structures. This cooperation also extended to a review of all maps produced by the census office by local administration officials before final EA map production.

———————

Source: Larry Li, Statistics Canada, personal communication.

(c) Aerial photography

vi. Air photo overview

2.256. Aerial photography is the method of choice for mapping applications that require high accuracy and a fast completion of the tasks (Falkner, 1994). Photogrammetry—the science of obtaining measurements from photographic images—is used to create and update topographic base maps, and to carry out agricultural and soil surveys, and for many aspects of urban and regional planning. Census projects have also frequently made use of air photo surveys to quickly create maps for areas for which up-to-date maps are not available or that 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.

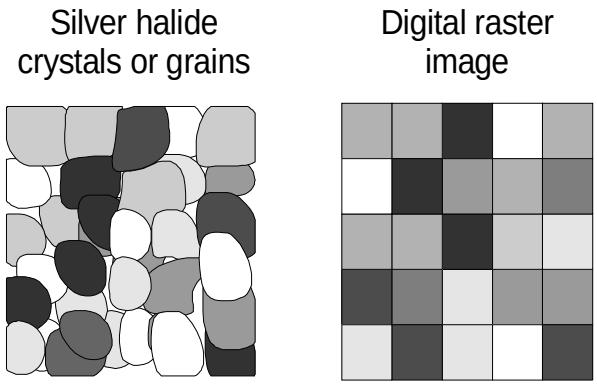
2.257. Air photos have been used for mapping since shortly after the invention of airplanes. Early applications made use of standard cameras. Very soon, however, specially constructed camera systems that minimize geometric distortion were mounted on specially adapted airplanes that allow the camera system 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, interpretation of stereo pairs of images became the dominant method for producing maps of elevation contours. A detailed review of traditional air photo interpretation techniques is given in the BUCEN, manual on census mapping (BUCEN 1978). The following paragraphs are therefore limited to a description of some recent innovations in computer-supported aerial photography methods.

2.258. Aerial photography is obtained using specialized cameras on-board low-flying planes (Michael, 1997). The camera captures the image on photographic film. In comparison to digital sensor systems, film currently still provides a far superior resolution (i.e., the ability to distinguish small details). Given the rapid developments in the area of digital imaging this may, of course, change in the near future. 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 GIS databases.

2.259. Figure II.20. Black and white photographic film, for example, consists of a layer of gelatine in which tiny, light sensitive silver halide crystals are embedded. These crystals or grains are irregular in shape and size. The scanned image, in contrast, is a regular array of pixels (picture elements).

Figure II.20. Photographic film versus the scanned image

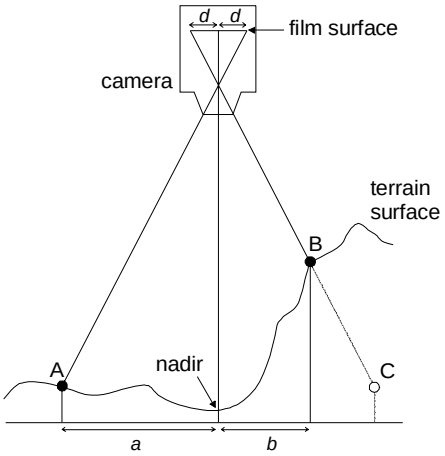


2.260. Aerial photographs are similar to maps as they provide a top-down view of features on the earth's surface. They are different from maps in that they only show features that are actually visible on the ground. Artificial boundaries, thematic information and annotation are, of course, absent. Without further processing, air photos also do not provide the geometrical accuracy of a map. Camera angle and terrain variation distort the view of an air photo. Additional processing is therefore required to produce so-called orthophoto maps, which combine the geometrical accuracy of a topographic map with the large detail of a photograph (see II.4).

Box II.4. Development of digital orthophoto maps

2.261. 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 II.21 (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 centre of the film's surface as point A.

Figure II.21. Distortion owing to terrain



2.262. 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 that are displaced by a small distance. Analytical stereoplotters allow the operator to correctly co-register the stereopair 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 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.

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

ii. Implementation and institutional issues

2.264. 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 internationally operating mapping companies that provide the airplane, camera and processing equipment.

2.265. These services are not cheap, however. Fortunately, air photos are useful for many different applications, including planning of service provisions, updating of town maps and land-titling projects (see, e.g., Ahmed, 1996: and Clarke, 1997). 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 owing to resource constraints, they can still be produced for specific areas. An example is the use of aerial photography by the statistical office of Hong Kong to estimate the number of people living on boats (NIDI, 1996). This illustrates the use of these techniques for counting populations that are hard to enumerate. Other examples are nomadic or refugee populations, rapidly growing urban areas, or regions that are seasonally inaccessible.

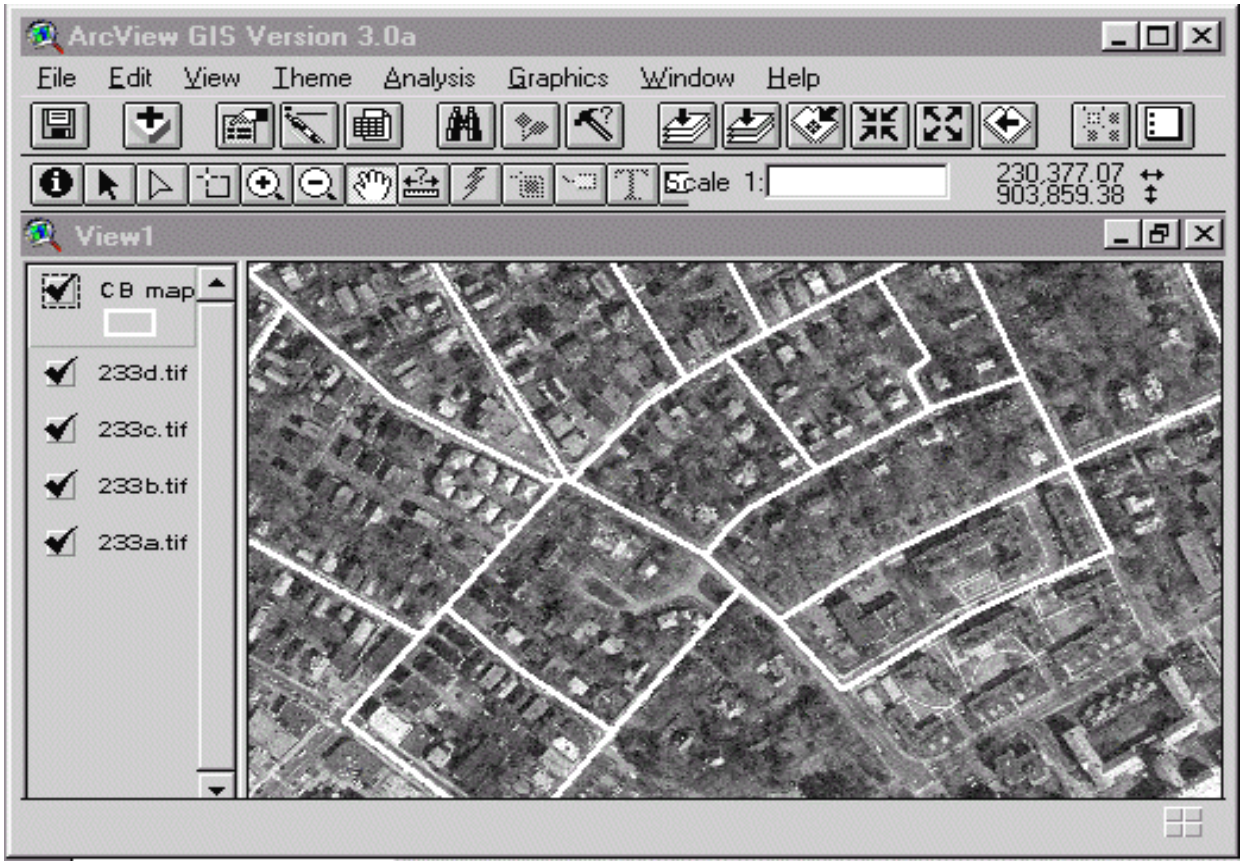
2.266. As described in the previous paragraphs, the development of orthophoto maps requires considerable technical expertise and specialized equipment. The use of orthophoto maps, in contrast, does not require significant additional training. A database for a city, for instance, may simply consist of a mosaic of several images on a CD-ROM 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 can 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.

iii. Application of air photos for census mapping

2.267. 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 of accuracy in the population estimates. Census staff can then delineate enumeration area boundaries that include a specified number of housing units. Since the orthophotos are correctly georeferenced, the resulting enumeration areas will also be registered in a proper map projection with known parameters. This means that possibly tedious georeferencing to make the digital boundaries compatible with other GIS data will be unnecessary.

2.268. Air photo interpretation is most often based on visual interpretation. Census cartographic staff therefore do not need to be trained in advanced image processing techniques. EA boundaries can be delineated on the air photo. Additional geographic features that provide the geographic reference for the enumerators can also be extracted from the photos. These features can be delineated interactively on the computer with the mouse or a similar pointing device (see Figure II.22). 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, sects. 4 (b) and 4 (c) on digitizing and scanning).

Figure II.22. Interactive delineation of census block boundaries on a digital orthophoto



(Source: MIT/MassGIS Digital Orthophoto Project http://ortho.mit.edu).

2.269. 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 addition to the EA maps, enumerators could be issued prints of digital orthophotos that show the EA boundaries to support orientation in their assigned area.

2.270. 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 to 30 cm). Resampled digital orthophoto images with pixel sizes between 0.5 and 2 metres are sufficient for delineating EAs in urban areas.

2.271. 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 already operational (Bossler and Schmidley, 1997). Digital frame cameras use arrays of a charge-coupled device (CCD) that can create images of 9,216 by 9,216 pixels, with a positional accuracy of 1 to 4 centimetres. 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 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 near future.

iv. Summary: advantages and disadvantages of air photos

2.272. The Advantages of air photos include the following:

- Air photos provide a large amount of detail and can be interpreted visually. Information about many types of features—roads, rivers, buildings—is shown concurrently;
- Data collection is faster and map data can therefore be produced much more quickly than by using cartographic ground surveys. Recent air photos are therefore a more reliable basis for census mapping compared to maps that are updated infrequently;
- Air photos can be used to produce maps for hard-to-reach areas or areas in which field work is difficult or dangerous;
- Topographic mapping using aerial photography can be cheaper than mapping using traditional surveying techniques. However, since the accuracy requirements for census maps are lower than for topographic mapping, the considerable costs are not necessarily justified if the products are used for census mapping only;
- Printed air 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.

2.273. The Disadvantages are as follows:

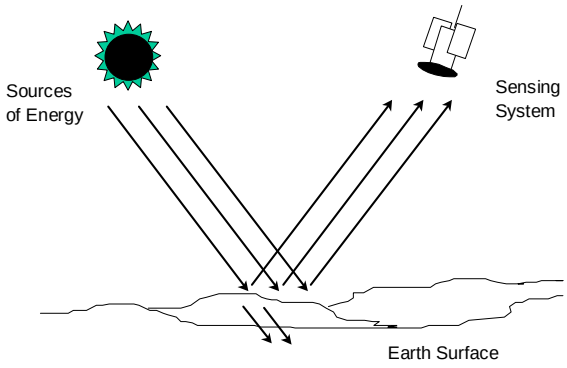
- Aerial photo processing requires expensive equipment and specialized expertise. Census offices therefore need to rely on outside support.
- Air photos still require information on the names of features that need to be extracted from possibly outdated maps. Aerial photography does not necessarily make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

(d) Satellite remote sensing

i. Principles

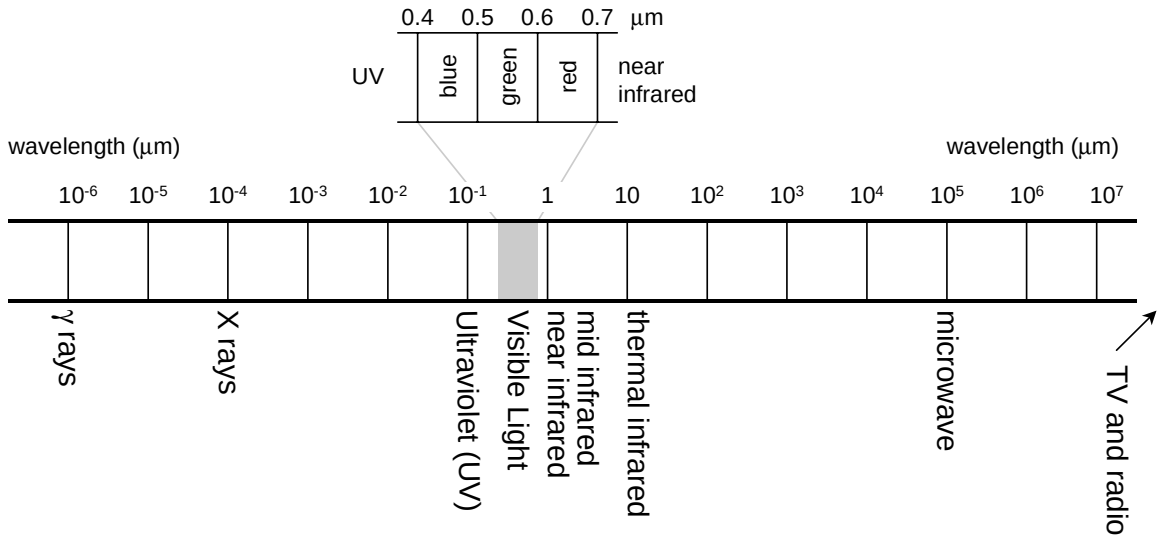
2.274. Some of the disadvantages of air photos—relatively small coverage on the ground and the need to conduct a special survey—do not apply to satellite remote sensing techniques (Lillesand and Kiefer, 1994: Jensen, 1996; and Gebizlioglu and others, 1996). Satellite images are collected from space based systems, most of which use so-called passive optical sensors to measure radiation reflected from objects on the earth's surface in the visible and invisible electromagnetic spectrum (Figure II.23 and Figure II.24). Satellite systems do not use photographic film to record the reflected energy. Instead, an electro-optical detector array—similar to a CCD camera—measures the intensity of electromagnetic radiation and records it digitally as a regular image of rows and columns.

Figure II.23. The remote sensing process



2.275. Satellite sensors operate in multi-spectral or panchromatic mode. *Multi-spectral* means that the satellite collects several images (or bands), each of which measures reflected energy in a different part of the electromagnetic spectrum, usually in the visible and near infra-red 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 mapping applications.

Figure II.24. The electromagnetic spectrum

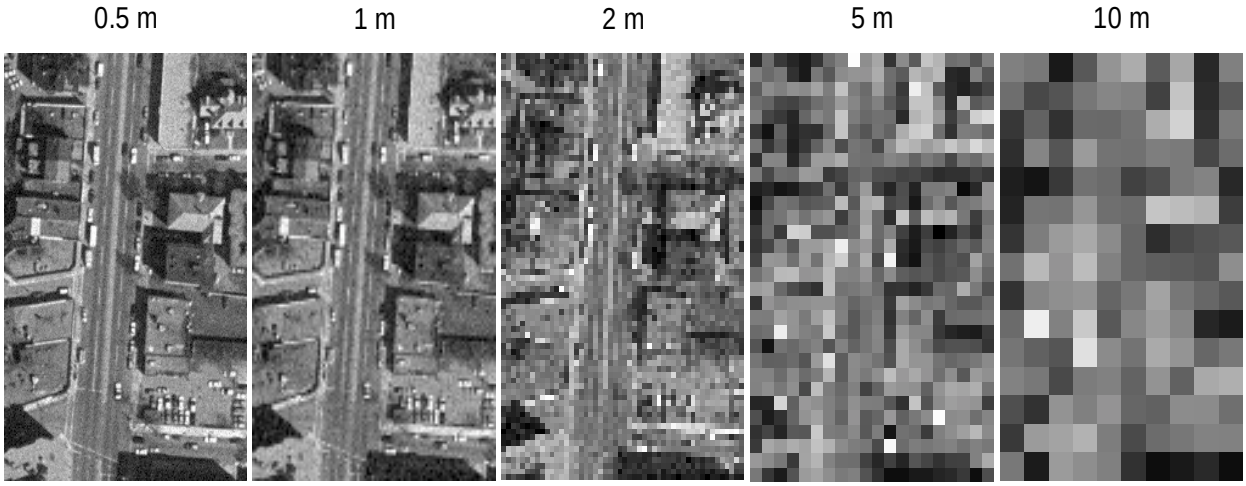


2.276. The digital data produced by the sensor systems consist of an array of numbers that indicate the level of energy reflected at the corresponding location on the earth's surface. The satellite sends these data to one of a system of earth receiving stations, where they are geometrically corrected and georeferenced. The resulting digital or printed images can be interpreted visually, similar to air photo interpretation discussed above. 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, however, 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.

ii. Resolution

2.277. Satellite image resolution is measured by the size of a pixel on the ground. Pixel size for commercial satellites varies from 10 to 80 metres for the most popular systems such as SPOT's panchromatic sensor and Landsat multi-spectral imagery. These resolutions allow mapping at cartographic scales of 1:25,000 to 1:50,000 or smaller. Figure 2.25 compares pixel sizes that were simulated from a 0.5 meter resolution digital air photo by aggregation. The image covers an area on the ground of 100 × 150 metres. Individual houses and even cars are distinguishable at a resolution of 2 m 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. Satellite data have also been used in rural regions of the developing world. One application has been the rapid development of topographic maps to support census mapping in the 1991 population census in Nigeria. Some 150 satellite image maps were produced from 90 SPOT images, covering an area of 110,000 sq km (Satellitbild, 1994).

Figure 2.25: Illustration of pixel size in aerial photographs and satellite images



2.278. Recently, higher-resolution satellite imagery has become commercially available. Russian and Indian satellites provide imagery at 2 metres and 5 metres resolution, respectively. Russian KVR 1000 images—a camera-based system—have been used for urban land use mapping and updating of city maps. Within the next few years, several private consortia will launch commercial satellites that promise to provide imagery with resolutions as high as 0.82 metres (Carlson and Patel, 1997). These companies expect that the availability of such high-resolution imagery will greatly expand the user base of satellite imagery. These satellite images may be cheaper and faster than aerial photography. However, since previous satellite systems were largely funded publicly, it is by no means certain that commercially operated systems can generate enough revenue to justify the large investments necessary to support their development, launch and maintenance.

2.279. Most of the commercial operators intend to provide several options for acquiring satellite images. 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 will be to obtain images on a less timely basis. Finally, over time, the satellite operators will build image archives, parts of which can be purchased at significantly lower cost. The price of imagery will also depend on the degree of processing of the raw data. This may include radiometric correction, geometric correction and geo-referencing, without or with ground control points. Raw image data will be considerably less expensive than a digital orthophoto map produced from satellite images.

iii. Applications

2.280. High-resolution satellite images show a level of geographic detail that is similar to digital orthophoto maps created from air photos. However, 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. Coarser-resolution satellite images may not show sufficient detail for dwelling counts.

2.281. Lo (1995) estimated population totals for urban wards in the dense urban area of Kowloon, Hong Kong based on the proportion of pixels classified as residential within each ward. While the overall errors were fairly low, since over- and underestimation cancel out to some extent, errors in each individual reporting units will often be unacceptable for census applications (see, also, Clayton and Estes, 1980, Lo, 1986; and Paulsen, 1992). Individual villages and major physiographic features outside dense cities can be extracted from coarser resolution satellite images, however. They can thus provide valuable information for EA map production and may show sufficient detail for delineation of EAs in rural areas. Pazner and others. (1994) provide an introduction to the extraction of information from remote sensing images.

2.282. As with air photos, acquisition of satellite images—though cheaper than air photo surveys—is quite expensive. 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.

iv. Advantages and disadvantages of remotely sensed data

2.283. The Advantages of remotely sensed data include the following:

- Up-to-date coverage of very large areas at relatively low cost with lower-resolution images;
- Large amount of information can be extracted from the images;
- Update of topographic maps in rural areas is possible; for example, identification of new settlements or villages that are missing on maps.

2.284. The disadvantages are as follows:

- Resolution of many systems is not sufficient for census applications;
- Cloud and vegetation cover restricts image interpretation;
- Low contrast between features—for example, dirt roads and traditional building materials in rural areas—makes their delineation difficult;
- Image processing requires a large amount of expertise.

4. Geographic data conversion

(a) Conversion of hard-copy maps to digital data

2.285. 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 is the step that requires, by far, the largest amount of time and resources.

2.286. The conversion of hard-copy maps or information from printed aerial photographs or remote sensing images to 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 usually a considerable amount of editing required to deal with any remaining errors or omissions. Following this step, the map coordinates, which are initially recorded in units used by the digitizer or scanner, need to be converted to the real- world coordinates corresponding to the source map's cartographic projection. Some systems allow the determination of the projection prior to digitizing. In this case, the coordinates are converted spontaneously during the digitizing process. The end result, of course, is the same.

2.287. 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, and so on) or a unique code that can be linked, for example, to a list of street names. In higher end-GIS software packages, this step is followed by the structuring of the database, also called building of topology. In this step, e 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 that are linked to the geographic database as needed.

2.288. The two main approaches for converting information on hard-copy maps to digital data are manual digitizing and scanning. The first involves the tracing of all required point and line features on a map, using a cursor or mouse. Digitizing techniques are also used to update existing digital maps on the basis of updated or marked-up map sheets. Scanning, in contrast, is the automatic process of converting a map into a digital raster image that can subsequently be converted into digital line work. The two approaches are discussed below in more detail.

(b) Digitizing

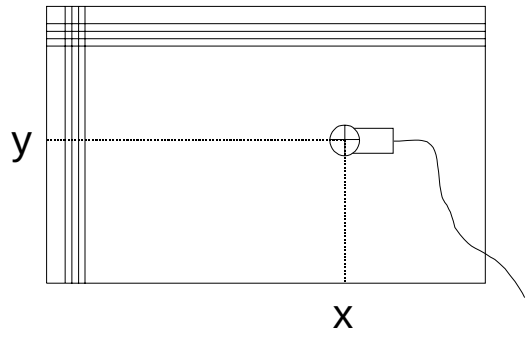
2.289. Manual digitizing has been the most common approach for spatial data automation. Manual digitizing requires a digitizing board that may range in size from small tablets of 30 x 30 cm to large digitizing tables of 120 x 180cm. Larger digitizing tables facilitate the digitization of larger map sheets. On a small tablet a large map will have to be digitized in several pieces that need to be combined later. In the process of digitizing, the map is fixed to the digitizing board using masking tape. Ideally, the map should be flat and not torn or folded. Paper often shrinks, especially in humid conditions, and this shrinkage introduces distortions that will be carried into the digital map database.

2.290. The first step is to determine a number of precisely defined control points on the map (usually at least four). These control points serve two purposes. Firstly, if a large map is digitized in several stages and the map has to be removed from the digitizing table occasionally, the control points allow the exact re-

registration of the map on the digitizing board. Secondly, control points are chosen for which the real-world coordinates in the base map's projection system are known. A good choice for control points are therefore the intersections of the graticule of latitude and longitude 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.

2.291. 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 II.26). 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 II.26. Digitizing table



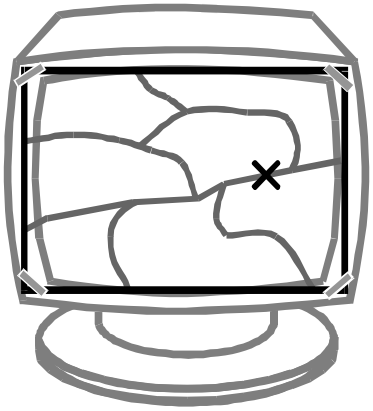
2.292. Coordinates are recorded in point, distance or stream mode. In point mode, the operator pushes a button on the cursor every time a line changes direction. For curved lines, the number of coordinates recorded will determine how smoothly the line will appear in the GIS database. In distance mode, 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.

2.293. Digitizing is 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 set-up. Consistent GIS software macro instruction that guide the operator, and quality control procedures, will minimize errors during digitizing and reduce the time required for later editing.

2.294. 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 so-called 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 attributes table which contains data about the polygon (see annex I).

2.295. A special type of data input without a digitizing tablet is sometimes called heads-up digitizing. The operator traces map features on a transparency and attaches this map to the computer screen (see Figure II.27). 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. This is a viable option in cases where a digitizing board or another standard coordinate input device is not available. However, the method is only appropriate if accuracy requirements are very low. In another type of heads-up digitizing the operator uses a scanned map, air photo or satellite image as a backdrop and traces features with a mouse. This method, which yields more accurate results is discussed in the following section.

Figure II.27. Heads-up digitizing



Advantages and disadvantages of digitizing

2.296. The advantages of digitizing include the following:

- 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; that is., there is usually no loss of accuracy compared to the source map.

2.297. The disadvantages are as follows:

- Digitizing is tedious possibly leading to operator fatigue and resulting quality problems that may require considerable post-processing;
- Manual digitizing is quite slow. Large-scale data conversion projects may thus require a large number of operators and digitizing tables;
- 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.

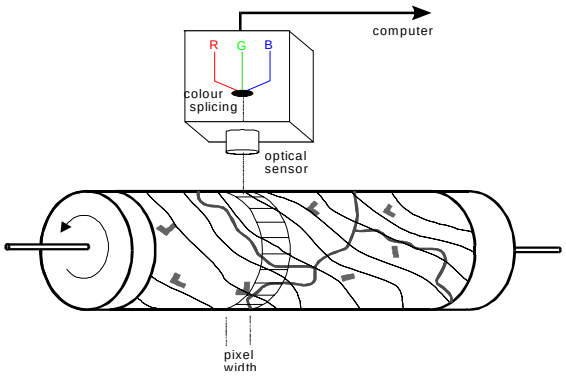
(c) Scanning

2.298. For many data input tasks, scanning has emerged as a viable alternative to digitizing. 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 grey-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, which can be stored in a standard image format such as a geographic interchange file (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.

2.299. There are three basic types of scanners in common use:

- Flat-bed or desktop scanners are currently found in many offices. They are of relatively small format so that larger maps must be scanned in several parts and joined in the computer. The document is placed upside down on a glass plate and the camera and light source move along the document beneath the glass. The strength of flat- bed scanners is their low cost and easy set-up and maintenance. They are useful for scanning text documents—for example data tables—which are later interpreted using optical character recognition software. They also provide a means to bring small graphics and maps into a computer. They are less suitable for large-scale map conversion tasks, where many large-format topographic and thematic maps need to be scanned. Scanning such maps in sections and joining the pieces later in the computer is time-consuming and may introduce a large number of errors.
- Drum scanners are more expensive and are used for professional applications that require very high 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 (see Figure II.28). While drum scanners provide very high precision, they are also very expensive and fairly slow. A single scan may take from 15 to 20 minutes.
- Feed scanners are currently 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.

Figure II.28. Principle of a drum scanner
(after Kraak and Ormeling, 1997)



2.300. 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 the anticipated further processing steps. The most important parameters are the following:

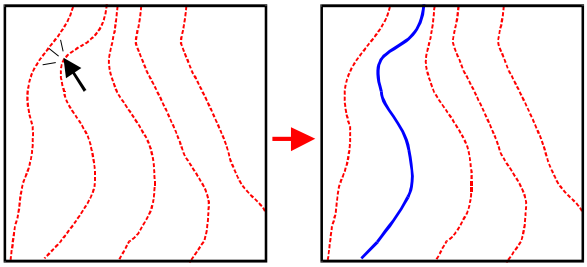
- Scanning mode. Binary or "line art" is appropriate for monochrome drawings or sketches, as well as for colour separations, where all features are basically of the same type. Grey-scale mode preserves variation on a map and subsequent image manipulation can be used to extract only features that have a certain reflectance value in a graphics or image processing system. This is even easier when the maps are scanned in 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 grey values or subtle colour tones are preserved. Higher contrast makes the image appear sharper, but may also lead to a loss of variation and detail. Threshold is a parameter that is used in binary mode to determine how grey 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 grey-scale range. This is often not the case. For example, the image might consist primarily of very bright and very dark areas. Gamma correction is a technique that considers the distribution of grey values in the image and adjusts automatically to brighten or darken areas, or to stretch cell values over a wider range of grey values. This technique can often help to preserve subtle variations in the image.

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

2.302. Raster-to-vector conversion can be performed in automatic, semi-automatic or manual mode. In automatic mode, the system converts all lines on the raster image into sequences of coordinates automatically. Since thick lines on the map result in lines on the raster image that are several pixels wide, the automated raster-to-vector process starts with a line-thinning algorithm. The next step is to determine the coordinates for each pixel that defines the line, followed, possibly, by the removal of coordinates that are redundant—that is., straight lines that can be represented by fewer coordinates. Conversion software also usually allows the user to specify tolerance levels. For example, features that consist of only one or a few pixels may actually represent dirt spots on the source maps and could be deleted automatically. Also, if the image has been scanned using a 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.

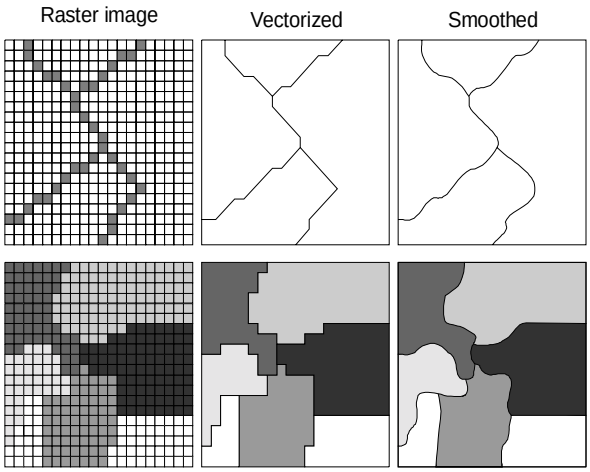
2.303. In semi-automatic mode, the operator clicks on each line that needs to be converted (Figure II.29). The system then traces that line to the nearest intersections and converts it into a vector representation. This has the advantage that the operator can select only a subset of features on the map, for example, all roads but not the rivers. Finally, in manual mode, the scanned raster image is simply used as a backdrop on the computer screen. Coordinates are created by tracing features on the scanned image using a mouse, similar to heads-up digitizing mentioned above.

Figure II.29. Semi-automatic vectorization



2.304. If linear or area features are converted automatically from relatively low-resolution raster images to 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 II.30 shows examples of a line and a polygon data set.

Figure II.30. Vectorization and smoothing of scanned image data



i. Some additional considerations

2.305. There are a number of considerations when planning a data conversion project based on map scanning. Extensive discussions of scanning techniques are provided in Pazner and others (1994), Hohl (1998) and United Nations (1997c). In the following paragraphs only a few major points are covered.

2.306. Proper preparation of the base map before scanning can significantly improve output quality. Maps should be flat and clean. Any tape residue that might be present on the map should be removed since it may leave traces on the scanner surface. Faint features on the map can be highlighted using a pen or marker. Similarly, the operator can retrace screened line symbols and fill cross-hatched polygons to produce solid lines and fills that 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 may make them invisible. For photographs, matte finishing will bring better results than glossy paper.

2.307. An additional step is often introduced for the conversion of relatively complex maps that show many different features (e.g., topographic maps) or of maps that are of bad quality. For such map data sources, accuracy can be improved and post-processing effort 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 subsequently scanned 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.

2.308. Despite these preliminary steps, the scanned images may 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.

2.309. GIS packages that support raster data provide raster-to-vector conversion routines. These are, however, largely 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 currently several commercial and non-commercial raster-to-vector packages available (Graham, 1997: United Nations, 1997c). 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 non-commercial product Map scan (United Nations, 1997c), for instance, offers most of the same functions that expensive commercial software provide. 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

2.310. The advantages of scanning include the following:

- Scanned maps can be used as image backdrops for vector information. For instance, scanned topographic maps can be used in combination with digitized EA boundaries for the production of enumerator maps;
- Clear base maps or original colour separations can be vectorized relatively easily using raster-to-vector conversion software;
- Small-format scanners are relatively inexpensive and provide quick data capture.

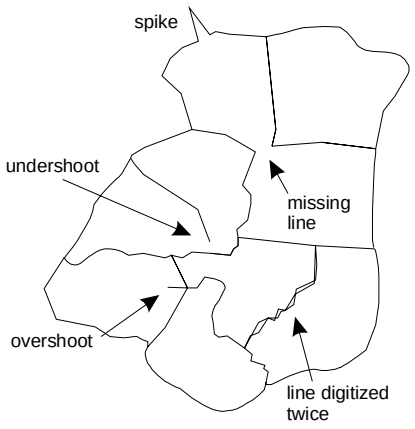
2.311. The disadvantages are as follows:

- Converting large maps with small-format scanners requires tedious reassembly of the individual parts;
- Large-format, high-throughput scanners are expensive;
- Despite recent advances in vectorization software, considerable manual editing and attribute labelling may still be required.

(d) Editing

2.312. The objective in converting geographic information from analog to 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 II.31. Similarly, after raster-to-vector conversion, disconnected line segments need to be manually joined. This happens, for instance, where small roads or rivers drawn with a thin line symbol cross major roads that are drawn as thick lines. If the minor roads or rivers are extracted into a separate map layer, there will be gaps in the road network at intersections with major roads.

Figure II.31. Some common digitizing errors



2.313. Some of the common digitizing errors shown in Figure II.31 can be avoided by using the digitizing software's so-called snap tolerances that are defined by the user. For example, the user may 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.

(e) Constructing topology

2.314. 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 I).

2.315. The user typically does not have to worry about how 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.

5. Digital map integration

(a)Introduction

2.316. A census mapping project should take advantage of all suitable cartographic data sources. These are likely to be stored in different formats, using varying map scales and cartographic projections. Integrating these heterogeneous data sources requires considerable knowledge of GIS data integration methods if the goal is to produce a complete and seamless digital census map database. The following sections discuss the most important methods that facilitate digital map data integration (for more details see Hohl 1998).

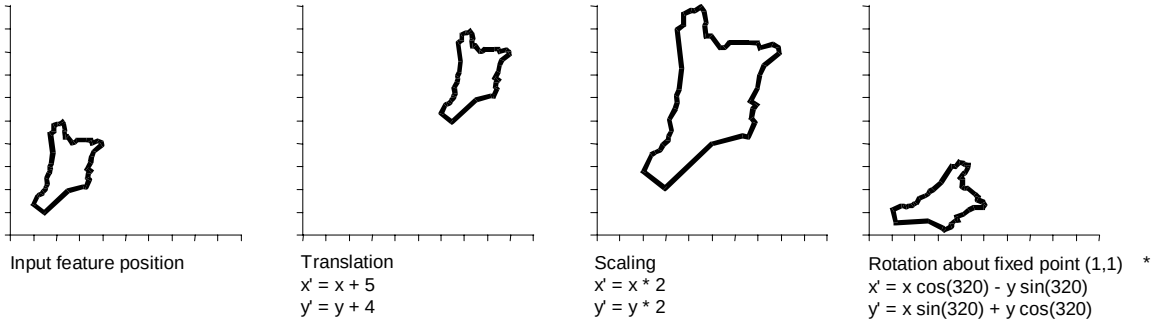
(b) Georeferencing

2.317. The coordinates captured with a digitizer or scanner are relative coordinates measured in the x and y direction usually in centimetres or inches from the data input device's origin—usually the lower left corner. If several adjacent map sheets are digitized, they will clearly not fit when their digitized map sheets are later pasted together in the database. In fact, they will be drawn on top of one on other since they are all referenced in the same segment of the digitizer's coordinate system. Similarly, existing georeferenced GIS layers for the same area or coordinates collected using a global positioning system will not be compatible with the digitized maps since they are referenced in a real-world coordinate system. For this reason, the digitized point and line coordinates need to be converted from digitizing units to real-world map coordinates are measured in metres or feet (see, also, annex II). As pointed out earlier, this step can be done in most systems either at the start of digitizing or after spatial data automation has been completed.

2.318. 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 II.32):

- 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 system's origin by a given angle. Rotation will make sure that the resulting digital map has the proper orientation even if the paper map has not been correctly aligned on the digitizing board.
-

Figure II.32: Translation, scaling, rotation

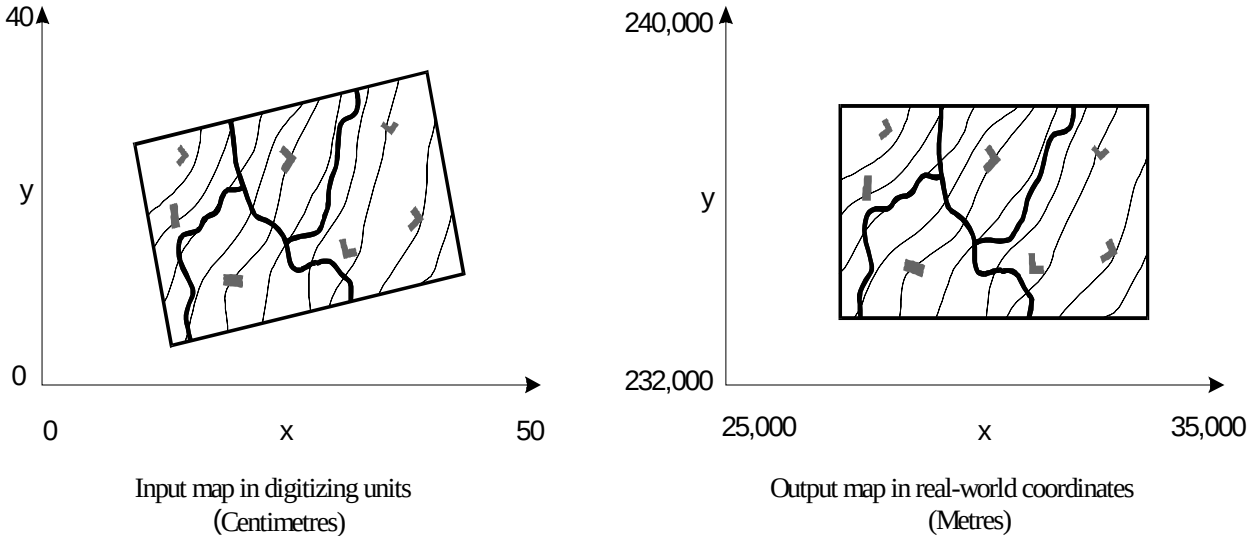


* Requires translation before and after rotation about the origin.
Rotation is positive anticlockwise.

2.319. 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 II.33). 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 in annex II.

Figure II.33. Map in digitizing units; map in real-world coordinates



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

2.321. 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. However, in some instances, where the input maps clearly do not conform to any well-defined projection, rubber sheeting is a viable option to make use of available geographic information. A good example in the context of census mapping is the georeferencing of hand-drawn sketch maps. Section F, Annex II, provides a practical example of georeferencing that illustrates the process of converting, for instance, a digitized map into a properly referenced digital database.

(c) Projection and datum change

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

2.323. Projection change is necessary when maps that were digitized from different map sheets need to be assembled into a seamless database. Often, maps issued at different map scales use different projections. In other instances, a mapping agency may have changed the standard projection used for mapping in the country, so that older map sheets may use a different projection

from those map sheets that were revised more recently. Similarly, the mapping agency may have modified the geographic datum, which establishes the reference framework for cartographic work in the country, so that older topographic maps, for example, use a slightly different coordinate system than do newer maps.

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

(d) Coding

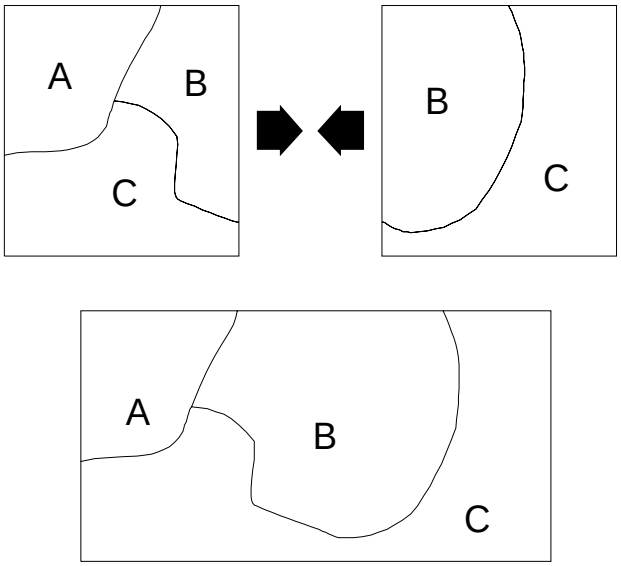
2.325. After the previous steps have been completed, the map database consists of a structured set of points, lines and polygons. Each geographic feature—that is, each point line, or area—has a unique identifier, which is used by the system internally. This internal identifier is not usually accessible by the user and should not be modified externally. What is needed is a more meaningful identifier that can be used to link the geographic features to the attributes recorded for them. For the enumeration areas and administrative units, this link is the unique EA or administrative identifier that is listed in the master file of all geographic areas relevant in the census.

2.326. 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.

(e) Integration of separate map segments

2.327. 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 GIS (see Figure II.34).

Figure II.34. Joining adjacent digital map sheets

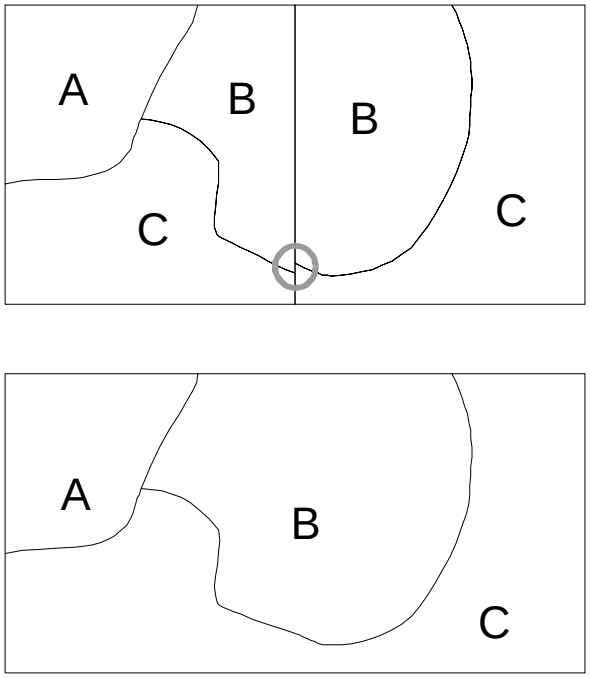


2.328. Usually this is straightforward. But the match between map sheets may not always be perfect. Features that span both sheets—for example,, roads or boundaries—might be displaced at the map boundaries (see Figure II.35). Errors could have been introduced during digitizing, or the errors may actually be present on the source map sheets. For instance, adjacent map sheets may have been produced at different times, so that newer features such as new roads do not continue across map sheet boundaries or they are represented by different symbols.

2.329. 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. Owing 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 judgement and experience.

2.330. 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.

Figure II.35. Edge-matching after joining adjacent map sheets



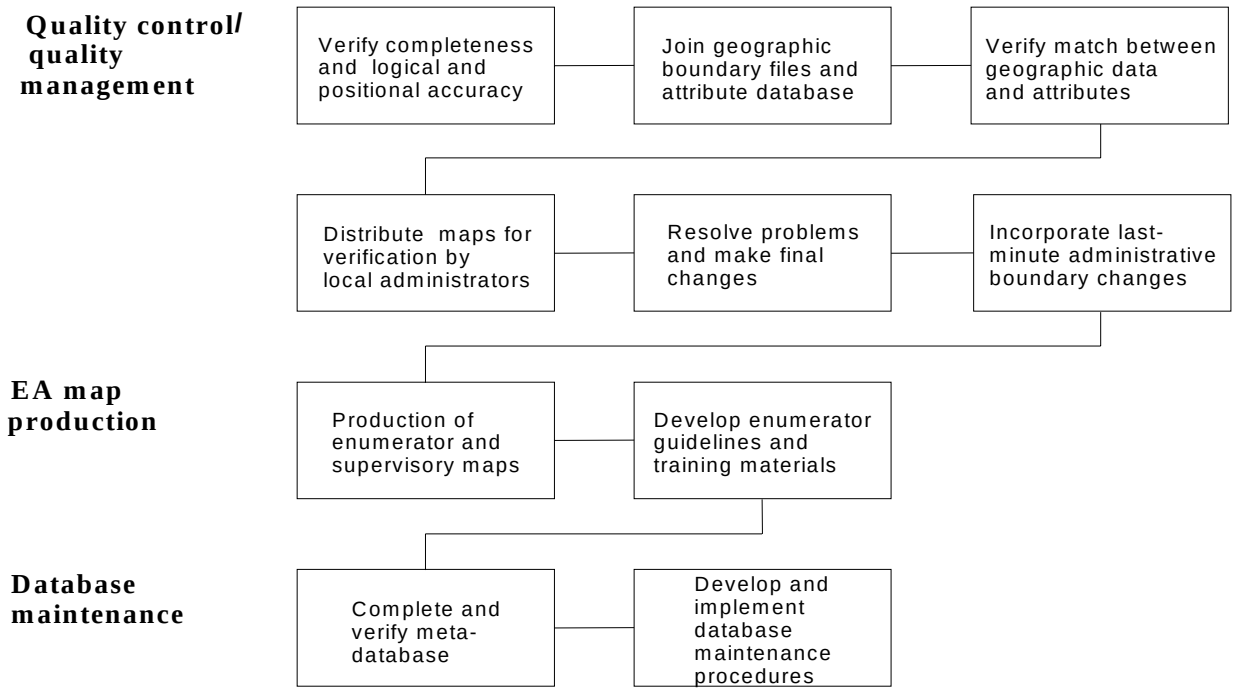
E. Quality assurance, enumeration area map production and database maintenance

1. Overview

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

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

Figure II.36: Stages in quality assurance, output production and database maintenance



2. Draft map production and quality assurance procedures

(a) Matching boundaries and attribute files and printing overview maps

2.333. In preparation of final map design and printing, the boundary data sets and geographic attributes file need to be matched if they are not already integrated in one consistent database. This step also involves checking the correctness of the match between boundary data and geographic attribute data. If both are correct, there should be at least one map feature (a point, line or polygon) for each record in the geographic attributes file. If this is not the case, there is either an error in the map database—that is, 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 (see sect.C.5 0 above).

2.334. Once the geographic data and the attribute information are correctly matched, labels need to be added to the map, and map symbols must be chosen that will identify features on the base maps (see, also, chap. III, sect. C.3, on thematic mapping). Labelling can be done interactively, semi-automatically or automatically, using a GIS package or a more specialized cartographic design software. In a very large census map production project, the labelling of features will be a time-consuming and tedious task. Especially when EA map design is quite complex—for example, many digital map layers are combined to produce each EA map—the resources required for proper label placement in terms of staff time and computer resources may be very large.

2.335. Most GIS and desktop mapping systems provide functions for automated label placement. The user simply specifies the attribute field in the GIS database's 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 one on other if features are too close together. However, in all but the simplest cases, some manual modification of the labels will still be required.

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

- 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.

2.337. 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.

(b) Quality assurance

2.338. 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.

2.339. Quality assurance refers to a final check of the digital map database before the products are released for the census operation. Quality assurance is similar to quality control, which was discussed in section C.5. 0 above. 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.

2.340. 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, described in the following paragraphs. Quality assurance is not a trivial task. It requires considerable time and resources and the census office needs to schedule and budget accordingly.

2.341. 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.
- 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.

(c) Verification by local authorities and final administrative unit check

2.342. 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—need to confirm that all settlements and parts of larger towns and cities are included in the geographic database. Involving local authorities in this process has the advantage that maps are reviewed by persons familiar with the local area. Naming and spelling conventions may vary in countries where several languages or dialects are in use. Approval of the maps by local officials will thus reduce the risk of errors of map interpretation by locally recruited enumerators.

2.343. A part of the verification process is also the confirmation of the administrative unit boundaries that are included on the EA maps. These boundaries change often. This poses problems for the census organization, which needs to produce summary statistics for these units. Several options are available to handle this problem:

- In the ideal case, administrative boundaries are frozen by government decree several months before the census. This provides stability of the reference framework for the duration of the census. The boundary structure that is current for this period is the one for which census tabulations will be produced.
- Continuous tracking of administrative boundary changes before the census is a second option. As changes occur, they are immediately committed to the digital map database. That way, the boundaries will be current at the time of enumeration. However, constant monitoring of changes and modification of boundary databases requires additional resources.
- In some countries, boundary changes are announced in advance. The census mapping agency can thus schedule work on those areas for a late stage in the census mapping process.
- The final option is for the census mapping agency to determine a freeze date and to revise all boundaries at a later stage, possibly after the census has been taken. If modified administrative unit boundaries cut through existing EAs, the household questionnaires for these units must be reassigned to the correct units. This introduces an additional step after enumeration and will therefore delay dissemination of census results.

1. Enumeration area map printing

2.344. After completion of verification and quality assurance procedures for all base maps and EA delineations, census cartography staff will print the final supervisory and EA maps. Supervisory maps will show several EAs and will be printed at a smaller cartographic scale. Defining the map layout for individual EAs is similar to the cut-out procedures in manual census mapping approaches (see BUCEN, 1978, p.149). EA maps should be simple, because they will be used by enumerators who have limited experience with maps. On the other hand, they must contain enough information to allow easy orientation. They should contain the following information:

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

2.345. Figure II.37 shows the components of a hypothetical urban EA map. All features are stored in

separate map layers in the same spatial reference system or as graphics templates. The main components are the street network, buildings and EA boundaries layer. In addition, annotation, symbols, labels and building numbers are stored in separate data layers, although these could also be added dynamically. The last component is a template consisting of neatlines and a legend that is consistently used for all EAs. Figure II.38 shows the complete EA map with all components overlaid on one map display. Depending on the scope of census mapping activities and the complexity of the enumerated area, EA maps may contain less or more information than this sample map.

Figure II.37. Sample components of a digital EA map

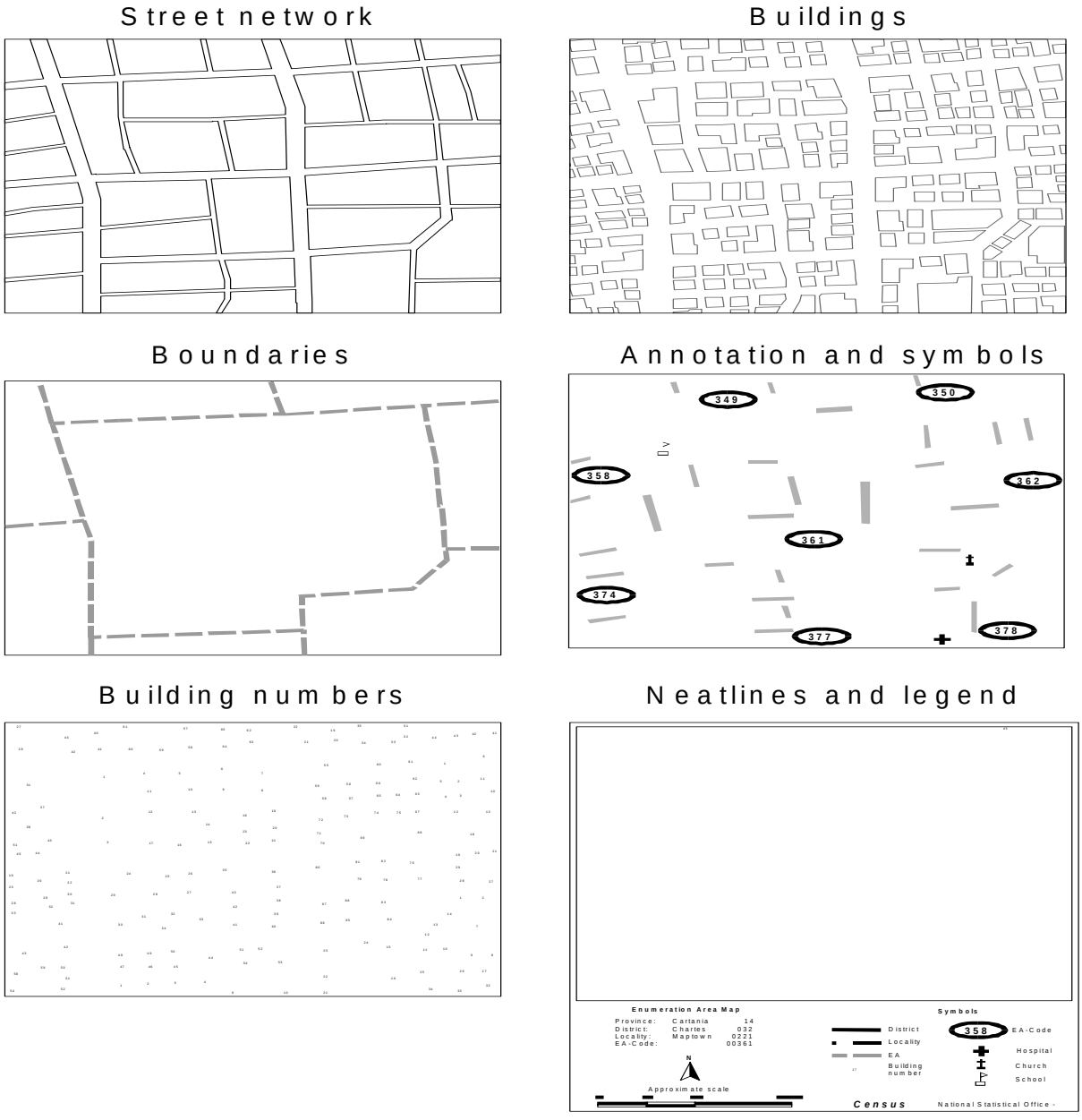
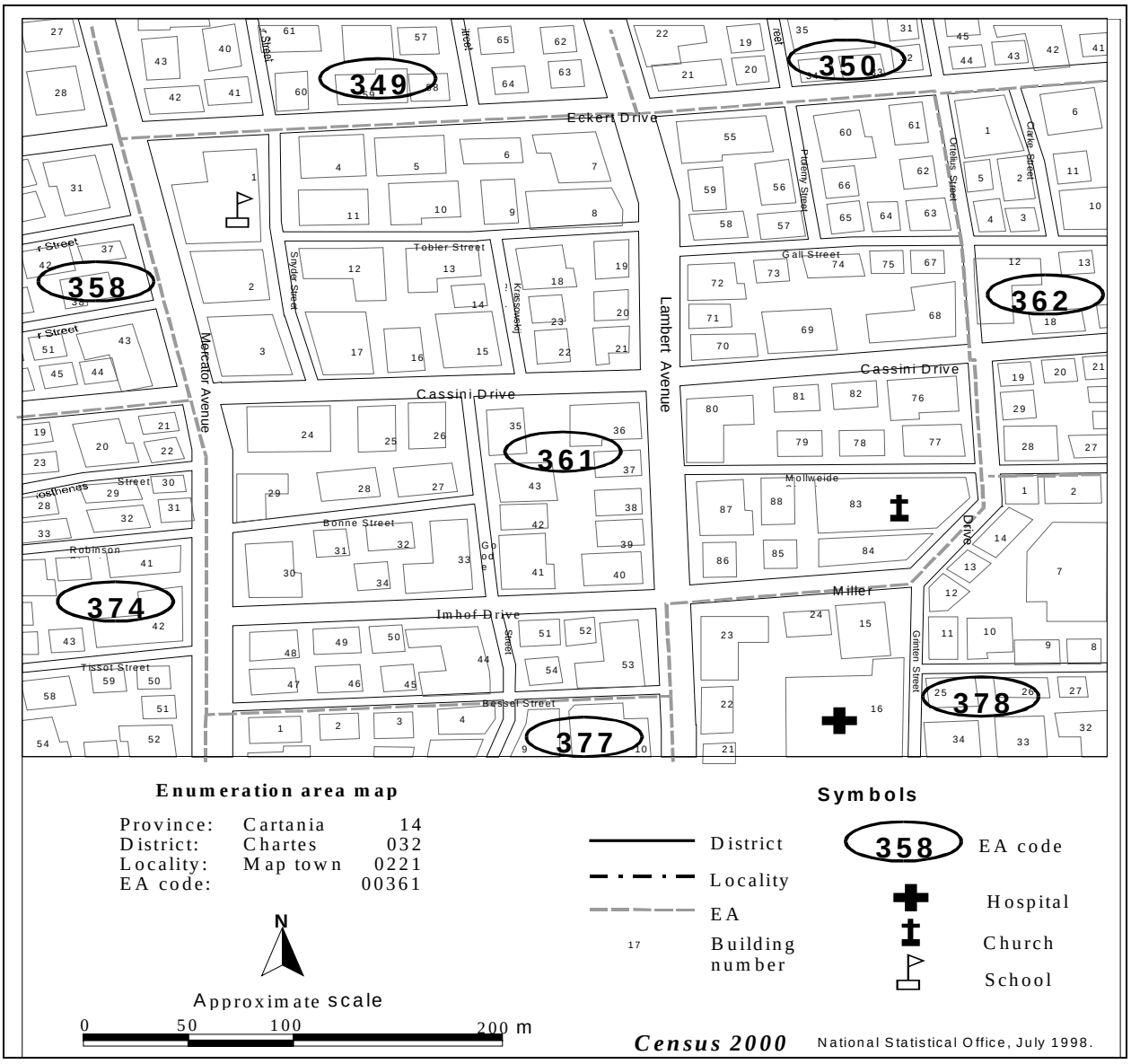


Figure II.38. Example of an urban enumeration area map



2.346. In many countries, EA map design may be simpler than in this example. For example, instead of a fully integrated digital base map in vector format, rasterized images of topographic maps may be used as a backdrop for EA boundaries. In some instances, map features may be more generalized, for instance by using only the centre lines for the streets, and polygons for entire city blocks rather than for individual houses.

2.347. Decisions must be made concerning format and colour. Given the high resolution available on laser printers, EA maps can usually be produced on A4 or letter-sized paper if the EAs are not too large and complex. Compared to larger-format printers or plotters, this has the advantage of lower cost and higher output speed. 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.

Box: II.5. Field map production for the 2001 census in the United Kingdom

2.348. In the United Kingdom, the census geography project is responsible for the provision of maps for data collection field staff to enable them to effectively manage the delivery and collection of census forms. Some 70,000 enumerators will be responsible for work within enumeration districts (ED). Census officers and assistant census officers (2,000 and 6,000, respectively) will manage groups of EDs, termed census districts (CD). Census area managers (approximately 120) will manage groups of CDs. Each level of staff will require maps detailing their areas of responsibility. For the 1999 census rehearsal and the 2001 census, the geography project will use a GIS- based system to plan and map enumeration and census districts.

2.349. *Map production requirements*. Each ED or CD map should display a 1:10,000 scale black and white raster map base to provide real-world context, simple coloured lines to define ED, CD or other statutory boundaries, ED and CD reference codes and, for CD maps, ward names. They will also display the office logo, orientation, map scale and copyright details. Each map should ideally fit on an A4 sheet for portability and the scale of the map background should be between 1:1,250 and 1:10,000. Where the map scale exceeds these limits, the map should be produced on an A3 sheet, then A2 and even A1 for certain rural areas. The 2001 census will require approximately 70,000 ED maps (90 per cent to fit A4 and 10 per cent at various sizes), 2,000 CD maps, plus other higher-level and ad hoc outputs.

2.350. *Map production approach*. The United Kingdom census geography office uses standard commercial GIS software, which provides adequate facilities for the creation and printing of map outputs. It incorporates an interactive procedure, whereby operators can collate a list of EDs they wish to plot and initiate a batch-plotting procedure. This procedure creates a plot for each ED, using templates of an appropriate size to maintain acceptable map scales. The plots are initially produced in postscript format (PDF), but are automatically converted into portable document format, which provides smaller file sizes and superior printing speed. The system then automatically sends the PDF file to a printer and places it into an archive directory. Refining the procedure from using postscript files only to PDF formats reduced overall time to output by a factor of about 10. Given the data volume, the time savings are very significant.

2.351. The benefits of the refined system include:

- Complete automation of batch map plotting from selection to output of print;
- End to end print times are considerably reduced;
- Plot file sizes are considerably reduced in comparison to postscript. Typically, reductions of 70 per cent can be achieved, although they can be as low as 20 per cent;
- The system can cater for varying plot sizes in the same batch run;
- Archived files can be easily retrieved, viewed in Acrobat and replotted.

Source: Office of National Statistics, United Kingdom.

2.352. A well-designed EA map will usually work well in black and white. Although colour printers are relatively cheap, their throughput is limited and supplies are often quite expensive. Good black and white maps can also be photocopied without loss of information, which allows the local staff to produce additional copies of EA maps as required. Where resources are available, however, colour can contribute to the clarity of map design. For instance, the EA boundary can be indicated by a brightly coloured line on the map.

2.353. Several copies of each EA map must be produced, in addition to back-up 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 four or five copies. If mapping activities are centralized in one or a few census offices, one approach is to distribute digital map files rather than hard-copy maps. These files can be transmitted to local census offices on diskette, CD-ROM or by way of the Internet. The local office will not need to have access to mapping software if the maps are exported to a generic file format such as the portable document format (PDF) or as a graphics file embedded in a generic word-processing format. Such

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

2.354. If the database is consistent and well organized, EA map production should be fast. Printing of EA maps will not require a high-end GIS package, but can be done using relatively inexpensive desktop mapping packages. Some of the process can be automated, using the built-in macro language of the software. For instance, a list of EAs can be accompanied by the bounding coordinates of the EA (the so-called map extent) in map units. The software can then be instructed to go through this list, include the content of the data layers into a pre-prepared template showing the legend and other marginal information, and print a specified number of copies.

F. Use of geographic information systems during census enumeration

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

1. Use of digital maps for census logistics

2.356. Maps are needed for many purposes in the census process. Among these, GIS can also play an active role in planning preliminary work and the logistics of the enumeration. Assignment of administrative units to operational areas, location of field offices and planning the travel of fieldworkers and enumerators are some of the tasks where GIS is potentially useful. If digital maps are to be used for these purposes, the census cartographic unit should develop a coarse-resolution GIS database very quickly. This system could consist of small-scale digital maps (at 1:500,000 or 1:1million) on settlements, roads, rivers and administrative divisions. These can be obtained from existing sources in most cases. Even the Digital Chart of the World, a consistent digital map at 1:1 million scale (Danko,1992: and Tveite and Langaas, 1995), can be useful for these purposes.

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

2.358. The use of GIS for logistical purposes is not quite as critical as the use of digital techniques to carry out the actual census cartography tasks. Many of the tasks can be done equally well by studying published maps. The advantage of using a GIS for these purposes is that distance and travel time estimates will be more accurate and that the census staff can quickly produce maps showing various aspects of the census planning process. Furthermore, the development of a small, coarse-resolution GIS database for the country is a good exercise for the much more challenging task of producing a detailed georeferenced census database.

2. Monitoring progress of census operations

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

2.360. Some countries implement a so-called 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.

2.361. In practice, any suitable summary statistic 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 have been completed. Census staff can then regularly link this information to the GIS database and prepare map output for evaluation by the overall census supervisors.

2.362. The key to this rapid quality control procedure is the fast flow of information from the

supervisors to regional offices and on to headquarters. The quickest way of exchanging this information is the Internet. If local and regional supervisors have access to the Internet, information can even be submitted through a password-protected database interface on the Web.

3. Updating and correction of enumeration area maps during enumeration

2.363. Even if a thorough quality control program has been carried out during the preparation of enumerator maps, it is likely that many of the maps are not perfect. For example, during initial fieldwork, buildings or streets may have been overlooked or registered incorrectly on the maps. Furthermore, since cartographic fieldwork needs to be conducted several months or even years ahead of census taking, new construction and infrastructure developments will not be considered in the enumerator maps.

2.364. In addition to training in data collection and basic map-reading skills, the census office should also instruct the enumerators to annotate the EA maps during enumeration to point out any errors or omissions. The census cartographic staff should collect the EA maps after the census and follow up on any suggested revisions. This may simply require making the corresponding corrections in the digital census database, or it may require some additional field checking. This process will ensure that the census office holds the most current information on the enumeration areas, which will reduce the workload for cartographic activities before future censuses or surveys.

III. Post-Enumeration

A. Introduction

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

3.2. If a complete digital census GIS database is available, statistical GIS databases for administrative or statistical units can be produced quickly by aggregation. Many countries, however, will not yet use digital techniques for the production of EA maps in the 2000 round. These countries may still choose to develop a digital georeferenced census database 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 aggregate statistical regions. At that level of aggregation, the resources required for producing a digital database are much lower than those necessary for a complete digital EA map database.

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

3.4. The three main sections in this chapter discuss management tasks relating to geographic databases after and between censuses and the development and dissemination of output products. The final section discusses some advanced topics such as urban and rural area delineation and techniques for dealing with incompatible geographic units.

B. Tasks after the census and during the intercensal period

1. Immediate tasks

(a) Incorporating updates and changes suggested by enumerators

3.5. The census mapping office should instruct enumerators to indicate any errors or inconsistencies in the delineation of EAs or base map features that are detected during enumeration on the EA maps. Local supervisors should then collect EA maps after enumeration and forward them to the census mapping office. The census geography unit can then correct the map database that was used for EA map production based on this information. This procedure will have two benefits.

3.6. Firstly, it ensures that tabulations and the development of digital and hard-copy map products are based on the EA delineation actually used during enumeration. Secondly, committing the modification of EA boundaries into the master digital map database will facilitate future census or other statistical data collection activities that are based on the same or similar geographic collection units.

(b) Reconciliation of collection units and tabulation or statistical units

3.7. The most important post-enumeration responsibility of the census mapping agency is to support the development of tabular statistical data produced from census returns. Census data are required for many different types of aggregated areas since census users from different sectors tend to use different geographic areas as the basis for planning and operations. EAs therefore need to be aggregated to these various reporting units as required for the development of a wide range of census output products.

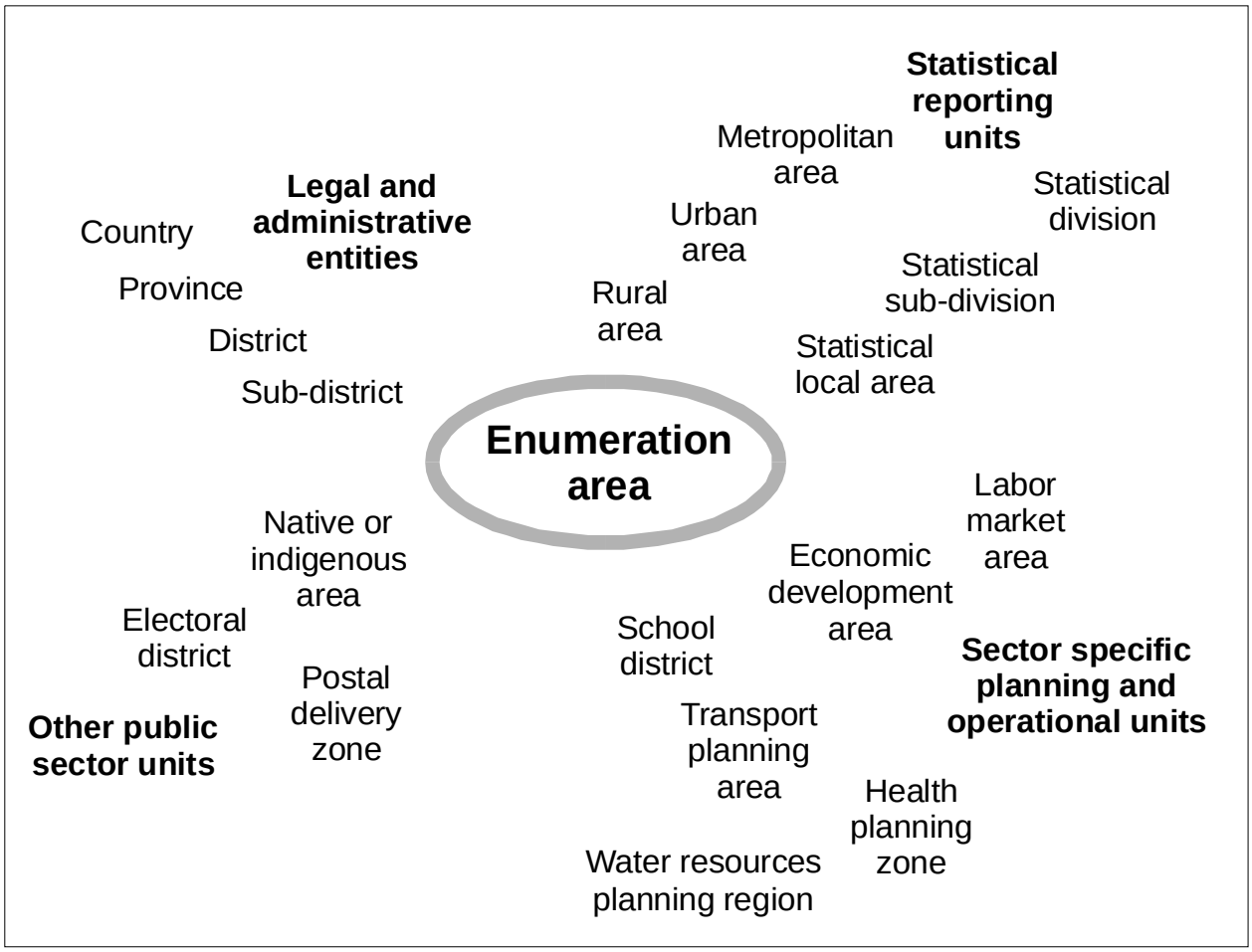
3.8. Matching of data collection and tabulation units requires the development of *equivalency* or *comparability files*. These files list for each tabulation unit the corresponding EAs that are part of that output unit. Once such lists have been defined, aggregation can be done using standard database operations.

3.9. The development of equivalency files is made easier if a consistent coding scheme has been implemented. This 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.

3.10. 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. Examples are health units, school districts, transportation planning regions, electoral districts, utility service zones, postal zones and environmental planning units (see Figure III.1). These may, in some instances, coincide with administrative areas, but often they will be incompatible with standard reporting units. In addition, special tabulation requests are likely to come from the private and academic sectors. Developing a consistent procedure for production and maintenance of equivalency files is thus an important task of the census mapping office.

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

Figure III.1. Examples of census tabulation and reporting units



2. Database maintenance

(a) Database archiving

3.12. After errors and inconsistencies have been addressed in the master digital census database, benchmark copies of all GIS data sets should be produced and archived. This database, for which the census geography has been frozen to reflect the situation at census time, will be the basis for all cartographic outputs, including reference maps, thematic maps of census results and digital extractions from this master database for distribution. All census results that are tabulated following enumeration will refer to the reference units in this database. This also implies that all documentation and 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 archived in a secure place immediately following completion of database work.

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

(b) Database maintenance: advantages of a continuous mapping program

3.14. As previously stated in this handbook, the benefits of a digital cartographic census program will outweigh the costs 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. Both 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 only possible if there is a high degree of continuity in the national census mapping program. Continuity of census geographic activities will therefore ensure that the investment in database development is preserved.

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

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

3.17. Digital cartographic data development requires special expertise in computer use, geographic concepts and specialized software packages. It is expensive to train personnel in all but the most basic GIS concepts and tasks. For a long-term census mapping program to be successful, staff continuity is therefore a critical factor. The census office needs to identify a core staff who will maintain the database in the intercensal period, provide GIS services for other statistical applications such as sample surveys, and serve as an institutional memory. This will facilitate a smooth operation of census GIS applications in the next enumeration. Core staff can, for instance, carry out the training of temporary staff recruited for digitizing or fieldwork. Retaining core staff will also reduce the start-up costs otherwise required for recruiting GIS experts who would then need some time to be fully integrated in the census cartographic process.

3.18. Again, the importance of a long-term view of census cartographic activities is emphasized. The benefits from pursuing a long-term strategy will be well worth the additional resources required to maintain a cartographic capability in the census office between censuses.

C. Dissemination of geographic census products

1. Planning data dissemination

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

3.20. The selection of suitable output products should be guided by a detailed assessment of customer requirements—that is, market research—that should be

carried out in the early stages of census planning. These plans for dissemination products should be made very early and published widely in order to get feedback from the user community.

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

3.22. Past experience of what has proved popular with census data users can only be a limited guide to the definition of output products. Demands change, partly in response to changing technical capabilities among data users. Digital map database products were rarely available after the last round of censuses, while they will be one of the most important outputs of the current census round. While demand for hard-copy maps may be larger in many countries than requests for digital information, this may well change in the near future. Thus, the census mapping agency needs to be flexible to respond to changing customer needs and special requests.

3.23. It is advisable to look ahead several years when planning the output strategy. For example, the Internet may not yet be a major data distribution channel in many countries. But this is likely to change in a few years as communications infrastructures are improving worldwide. Also, new user communities will emerge as new data products are created. 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.

3.24. The census office should also try to estimate the volume of possible demand for its products and services, which will allow some assessment of required capacity for servicing customer requests. Again, this is difficult since demand may increase as new products are introduced and as new users see census products and realize their potential for their own needs. Thus the census office needs to be prepared to serve a growing demand once products are made available. It is advisable to define clearly and early which census data users' needs *must* be served, whose *should* be served, and whose *will not* be served. A clear set of priorities will also facilitate the development of a timetable for census product distribution.

3.25. An open data dissemination policy—that is, 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 to serve.

3.26. Some census geographic data products will be required for internal and official use. This 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.

3.27. 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.

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

2. Required products

(a) Equivalency and comparability files

3.29. These 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.

3.30. 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

3.31. 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.

3.32. 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.

3.33. 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 Bureau of the Census, which is available on the Internet.

(c) Gazetteers and centroid files

3.34. Although it is usually the responsibility of the national mapping agency to produce gazetteers—a list of place names and their geographic location—a large-scale national mapping program implemented for census purposes may provide an improved or updated information base for a national gazetteer. 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.

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

3. Thematic maps for publication

(a) The power of maps

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

- Maps communicate a concept or an idea.
- Maps are often meant to support textual information. Some things are difficult to explain in words and a map display can help to explain complicated issues.
- Maps appeal to the viewer's curiosity. They provide eye-catching anchors on the pages of a report. These will get the reader's 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 for counties 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 hard 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 some commentary. 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 pretty. 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

3.37. GIS encourages a view of maps that is quite different from traditional cartography. In a computer, maps can be generated quickly on a computer screen. This supports a mode of work that is optimized for data validation, exploration of data patterns and data analysis. Maps created on a computer screen are sometimes called "virtual maps" to distinguish them from printed or drafted 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 II—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.

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

3.39. Table III. 1 shows a list of possible thematic maps that can be included in a census atlas or a census office's Internet Web site (see United Nations, 1998). Many other types of maps might be considered for publications on special topics or to highlight interesting aspects of census results in regions of the country. Just as tabulations of census data can be disaggregated by gender, age group or urban/rural areas, census maps can also be divided into population components. Maps that show comparisons over time, if comparable indicators are available from previous censuses, are also informative.

Table III.1. List of thematic maps for a census atlas

Population dynamics and distribution
- Percentage population change
- Average annual growth rate
- Population density (persons per square kilometer)
- 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
- Per cent of population age 0-14
- Per cent of population age 15-64
- Per cent of population age 65 and over
- Per cent female population in child-bearing age 15-49
- Total dependency ratio (per cent of population age 0-14 and 65 and over to population age 15-64)
- Marital status
- Birth rate
- Total fertility rate
- Death rate
- Infant mortality rate
- Life expectancy at birth

Socio-economic characteristics
- Educational level of population age 10 or over
- Literacy rates
- Illiterate population age 10 or over
- Unemployment rate
- Unemployed population (total number)
- Employment-population ratio
- Occupational structure by economic sector

Household and housing
- Average number of persons per household
- Average number of dwelling rooms per household
- Tenure status (owned, rented, etc.)
- Type of construction material
- Access to safe water
- Access to electricity
- Access to sanitation

(Source: *Principles and Recommendations for Population and Housing Censuses, Revision 1*. United Nations publication, sales No. E.98.XVII.8).

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

maps can be produced using census block or enumeration area-level data.

3.41. GIS and desktop mapping packages provide a wide range of cartographic functions and many commercial map-makers have switched to fully digital production techniques. Still, to achieve high-quality cartographic output requires considerable experience and know-how. Tools provided by computer-based mapping systems do not substitute for cartographic training. In fact, the availability of easy-to-use mapping packages has led to a proliferation of maps that violate many of the standard cartographic design principles. Initially, this was attributable to the lack of proper cartographic functions in early GIS packages. Today, it reflects the use of such packages by users who have had no training in cartographic techniques.

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

3.43. Owing to the widespread diffusion of GIS and desktop mapping software, thematic maps are increasingly produced by subject specialists with little or no training in cartographic design principles. Annex V therefore provides a summary of thematic mapping techniques. The information contained in annex V should be of interest to core cartographic staff as well as to people inside and outside the census agency who may produce maps from digital spatial databases only occasionally. Excellent additional references on cartography and thematic mapping are Robinson and others (1995), Kraak and Ormeling (1997) and Dent (1999). MacEachren (1994) produced a useful primer on thematic mapping, specifically targeted at GIS users with little formal training in cartography.

(c) Thematic map production and publication issues

i. Types of output

3.44. 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 (see sect. C.2 (b));
- 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 cheaper and 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 of maps;
- Digital census atlases that are a cost-effective alternative to printed versions in countries where computers are widely available. Census atlases can be based either on static pre-prepared maps or on a simple thematic mapping interface where the user can select the variables to map, the classification scheme, cartographic symbols and colours, and a basic layout;
- Maps can also be published on the Internet. The choice is again between static maps that are no different from other images or photos published on the Internet and dynamic mapping interfaces that give the user control over the thematic design process;
- Special purpose maps in various formats will be generated for in-house or outside census data users by special request. Such products will be printed in small numbers on in-house output devices such as laser or ink-jet printers;
- Presentation materials such as slide shows or large-format posters on census topics benefit from the inclusion of maps.

ii. Cartographic tools / software

3.45. 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 has 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 where the user can select styles from interactive menus, and map elements can be moved and resized using the computer's mouse. The on-screen map display shows quite realistically how the map will look on the printed page.

3.46. The cartographic design functions of modern desktop mapping and GIS packages will satisfy most user's requirements (e.g., Waldorf, 1995). For some applications, however, professional cartographers prefer to export the basic map from the GIS and import it to a graphic design or desktop publishing or graphics package. These packages provide sophisticated graphics functions such as 3-D effects, graduated fills or transparency, which give the cartographer greater flexibility in design. To copy from the GIS to the graphics package, two options are available. One is to use the standard cut-and-paste options in the Windows environment. The other is to go through an intermediate file in a standard format that can be imported by a graphics package (see section on output options below).

(d) Output options

i. Digital files

3.47. 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 Web sites as static map images and can also be exchanged as file attachments via electronic mail.

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

3.49. 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.

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

ii. Raster image formats

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

3.52. 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, and 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. For photos or photo-realistic graphical images, 16 or 24 bit image formats are more useful.

3.53. 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 that is contained in a small file header and contains the colour specification in a common colour model such as RGB.

3.54. Some standard raster file formats are:

- **BMP**. The Microsoft Windows device-independent bitmap (DIB) format. It allows Windows to display the bitmap image on virtually any type of display device. This is one of the most basic raster file

formats. Run-length encoding is supported, but file sizes are usually bigger than for other image formats.

- **TIFF**. Tagged 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, map extent and so on, while still complying to standard TIFF specifications. GeoTIFF is supported by most major GIS vendors, government agencies and academic institutions. The specifications are given in Ritter (1996)
- **GIF**. The graphics interchange file was designed for interchange of raster image graphics across hardware platforms. It supports a compression scheme that reduces file sizes significantly and is therefore optimal for exchange through computer networks. In fact, the format was developed by CompuServe for use in its 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 grey 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, that is not fully reversible. That means that a photograph that has been exported with a high degree of compression cannot be restored to show all details in the original photograph.

iii. Vector file formats

3.55. 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**. The Windows metafile is a graphics file format for use in the Windows environment. It is most often used for vector data, but can also store bitmap images. Enhanced WMF (EMF) files are a more comprehensive variation of WMF format developed for the 32-bit Windows environment (Windows 95 and NT). 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.
- **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 encoder 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 ink-jet and electrostatic printers, pen plotters were the most widely used output device for GIS projects that needed to print large maps.
- **DXF**. The drawing exchange format was developed by Autodesk, a software producer specializing in computer aided design (CAD) and GIS software. Initially designed to exchange Autodesk native files between platforms, DXF has become a standard exchange format that is supported by most GIS packages and many graphics software.
- **PS and EPS**. Postscript is essentially a programming language for describing vector data in a plain-text file. It is the most widely used page layout description. Postscript was developed by Adobe, a graphics software company. Optimized for scale-independent vector graphics, postscript files can also incorporate raster images. The main use of postscript is as an output format for sending documents and graphics to postscript printers. Postscript is thus fundamentally an output format. Many graphics packages support postscript import, but because the postscript codes are not completely standardized, it is often not possible to import postscript files for further editing if they were

created in a different computer program. This is especially true when the postscript file travels across hardware platforms. Sometimes, it is not even possible to import a postscript file created in the same software.

While it is often not possible to modify an imported postscript file, most software packages will be able to incorporate a postscript file in a document. Instead of the file content, only a 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.

- **PDF**. The portable document format was also developed by Adobe. Its initial use was for distribution of complex documents—containing text and graphics—on the Internet. PDF files can be created from any text-processing or graphics package using the Adobe Acrobat printer driver. The PDF reader can be downloaded free of charge from the Adobe Web site. 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.

iv. Personal printer

3.56. For small print runs or quality control plots, a census office should have one or a number of printers available. The following paragraphs briefly describe the most popular printer types (see, also, Cost, 1997):

- **Ink-jet** printers produce output by squirting electrically charged drops of colour through a nozzle onto the page. Liquid ink-jet printers use liquid ink that dries through evaporation. Ink is sent through the nozzle using hydraulic pressure in the so-called pulsed ink-jet technique. Thermal ink-jet, 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 ink-jet 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 ink-jet printers produce finer dots on the page compared to liquid ink-jet technology. Ink-jet printers work with plain paper, but to achieve highest possible output quality, specially coated paper designed for use with ink-jet printers is usually recommended. Owing to their reasonable cost and ease of operation ink-jet 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, which 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 (CMY) or four (CMYK) colours so 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.
- **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 only recently reached a price range to be considered for most graphics application environments. Their printing quality is not sufficiently high, however, to replace ink-jet printers as the most common colour printers for small and medium-sized GIS laboratories.
- **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.

3.57. 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 minute);
- Output resolution in terms of dots per inch (dpi) and number of colour or grey tones that can be produced;
- Media size;

- Supported media types (plain paper, specially coated paper, transparencies, and so on).

3.58. 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 dpi and more). They are ideal for printing reports and other documents that consist mostly of text, with some graphical illustrations and maps.

3.59. Colour printers are useful for printing complex maps for which monochrome shading and symbolization would be insufficient. Ink-jet printers are currently the most commonly used colour printers—from A4/letter-size desktop printers to large-format printers (e.g., 60 x 90 cm or 24 x 36 inches). They produce high-quality maps at 600 dpi. Printing speeds are still relatively slow for ink-jet printers. In the foreseeable future, colour laser printers, which at this point do not yet achieve the same printing quality, are likely to replace ink-jet printers as the most popular colour printing devices.

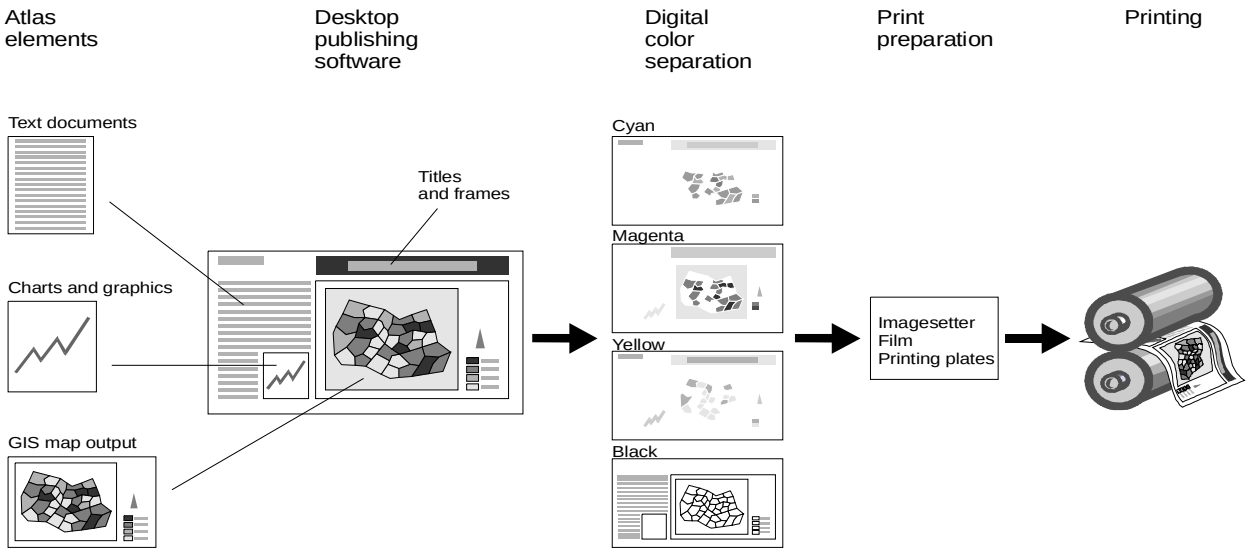
3.60. In deciding upon a suitable printer for a GIS project, cost is a major issue. One thing to keep in mind is that the purchase price of a printer is only one—often relatively minor—cost component. While printer prices have dropped considerably, the cost of ink cartridges and special paper have remained fairly high. In some cases, it appears that hardware prices are kept very low by printer manufacturers that hope to profit mainly from selling the hardware-specific supplies. In addition to the purchase price, the printing cost per standard page should also be compared (e.g., where 5 per cent of the page is covered by ink). Computer trade journals often publish comparisons.

v. Commercial printing

3.61. For larger print runs, personal printing devices are too slow and the costs per page are too high. Brochures, posters or census atlases will therefore be printed in an in-house or commercial print shop. If print volume is high, analog printing processes, where printing plates are produced and used in lithographic or similar printing machines, are currently still cheaper and faster than digital printing processes. This may change in the near future.

3.62. 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 the figure below (see Figure III.2). After an initial planning stage in which the text, graphics and map contents are specified, census cartographic staff produce all maps for inclusion in the atlas. These maps are stored in postscript format ready for printing. For complex map designs that incorporate graphics produced in external packages or photographs, the layouts may be produced in a high-end graphics package. Other census office staff members will write the text to accompany the maps, tables, references and other textual content in standard word-processing packages.

Figure III.2. The digital printing process



3.63. 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 may be done in-house or by an external service bureau.

3.64. Once the final atlas layout has been produced, it is saved in a digital output file. The most common file format is an encapsulated postscript file, but some software-specific file formats can also be used by commercial printers. Most high-end graphics and desktop publishing programs can also produce 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 so-called 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, which creates the film from which the printing plates are produced. Using digital files for producing the film will generally produce the best results. Camera-ready copies printed on a laser printer, which are reproduced by photographic techniques, may 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.

3.65. Useful references on digital pre-press—the preparation of material for printing—and digital printing are Romano (1996) and Cost (1997). Recent cartography books such as Kraak and Ormeling (1997) and Robinson and others (1995) also discuss digital pre-press and printing processes. Many vendors of printing hardware and software also provide extensive information and other resources on their Web sites.

4. Digital geographic databases for dissemination

3.66. For the time being, the publication of hard-copy census map products will remain one of the primary means of disseminating geographic census results. Access to computers varies by country, and even where computer use is widespread, many users will prefer a printed product. In parallel to the production of printed census maps, a census office should, however, also pursue a digital data dissemination strategy.

3.67. Demand for digital databases that consist of extractions of the census agency's digital geographic master database will continue to increase. Census data are an important input in policy planning and academic analysis in many fields. Health service provision, educational resource allocation, design of utilities and infrastructure, and electoral planning are some applications where government agencies require spatially referenced small area population statistics. Commercial users employ such data for marketing applications and location decisions.

3.68. The wide range of potential users of small area census data means that the census organization needs to pursue a multi-levelled digital data dissemination strategy. Broadly, we can distinguish between the following types of users:

- Advanced GIS users, who want to combine small area census data with their own GIS data on health facilities, school districts or sales regions, for example;
- Computer-literate users in the government, commercial or private sector. These users want to be able to browse the thematic information in a census database spatially. They will want to produce thematic maps and thus need to be able to perform simple manipulation of cartographic parameters. Simple analytical functions such as aggregation of census units to custom-designed regions should also be possible;
- Novice users, who mostly want to view pre-prepared maps on a computer and perhaps perform some basic queries.

3.69. The first group of users will want access to spatial and attribute information in a comprehensive digital GIS format. The census office needs to supply comprehensive documentation on the geographic parameters used for the GIS database, as well as on the individual census variables. The spatial information will be distributed in an open GIS format that can be easily converted into any number of commercial GIS formats.

3.70. The second group of users is best served with a comprehensive, pre-packaged application that is designed for a commercial or freely available desktop mapping package. Documentation requirements are somewhat smaller, since the users are unlikely to change the geographic parameters of the database or perform more advanced GIS operations.

3.71. For the third group of users, finally, the best data distribution strategy is to produce a self-contained digital census atlas. This atlas could consist of a series of static map images, for example, in the form of a slide show. Or it could be a 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

3.72. The first step in preparing the GIS databases for general release is to define the data content. The following questions need to be addressed.

i. Up to which level will data be released?

3.73. To maximize the overall benefits of census data collection, the objective of the census organization should be to release geographically referenced census data at the smallest level that does not compromise data privacy. Even at the enumeration area level, there may be special reporting zones that contain only a few households, and for which census data 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?

3.74. A high-resolution census GIS database will consist of thousands of reporting units. Such data volumes will be beyond the computing capacity of average data users. Instead of distributing one large database, the census organization should consider producing a family of census databases. At the medium resolution level—for instance districts—a national summary database can provide a sufficiently detailed overview of socio-economic conditions in the country. For each major civil division or even each district, separate databases that show indicators at the subdistrict and enumeration area level can be constructed. Individual databases can also be useful for major urban areas.

3.75. Finally, a point database of settlements in the country with associated census data will serve the needs of users who do not need the spatial resolution of a GIS database of reporting units. This database should at least contain all settlements classified as urban and the aggregate census indicators for each town or city. Ideally, a village-level database should also be constructed for the benefit of planners in the health, education or agricultural sectors. A village database can be based on a gazetteer of place names and locations if such information has been collected during census mapping.

3.76. Offering databases for subsections of the country will increase data use. Many users only need census information for a relatively small region. A subset of the national census database is easier to process by users with moderate GIS computing capacity. Also, in countries where the data access fees are larger than the cost of reproduction, smaller data sets are affordable to a larger number of non-commercial users.

3.77. If separate databases are distributed, care has to be taken that the individual parts or tiles are compatible. This means that the shared boundaries between database subsets need to match exactly. Separate pieces of the database should be in the same geographic reference system and have the same attribute database definitions. If the master database used by the census office is very detailed, it may be beneficial for some data users if a more generalized version of digital census maps is 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.

3.78. 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 be easily converted into other projection systems, if the user wants to use the census boundaries in combination with other data layers. Specific national projections and coordinate systems, in contrast, may not be supported by the GIS software. Users would then have difficulties in employing the census database for geographical analysis applications.

iii. How tightly should boundaries and the database be integrated?

3.79. Census GIS databases are characterized by their large number of attribute fields. Census questionnaires provide information that is stored in possibly hundreds of variable fields. Usually, it is impractical to store all of these in the same data table. A better approach is to select a small number of most important indicators in the geographic attributes, table and provide the remaining information in a series of separate tables. These external tables can be organized by topic—demography, household data and so on. The user can then link tables to the GIS by the common geographical identifier, as needed.

(b) Data formats
i. Coordinate data

3.80. GIS software packages differ widely in terms of the data formats that are supported. Each commercial package has its own native data format. In addition, import and export functions allow the user to convert data from a selected number of external data. In some instances, these conversion functions need to be purchased separately.

3.81. Despite some efforts by commercial and public GIS groups (see Open GIS Consortium, (1996), there is still no universally accepted and widely used generic data exchange format. The vector product format (VPF), which was initially developed for distribution of the

Digital Chart of the World, a 1:1 million-scale global base map, was intended as a general data exchange standard. However, it was never fully embraced by commercial GIS developers.

3.82. Instead, a number of exchange formats developed by leading GIS vendors have become de facto standards that are also supported by other software systems. The most important of these are described briefly as follows:

- AutoCAD DXF format (.dxf) originated in the CAD world. It is well suited to transfer the geographic coordinate data, but is not as good at converting attribute information.
- Arc/Info export format (.e00) was developed as a cross-platform exchange format for GIS databases produced by the Environmental Systems Research Institute (ESRI) Arc/Info GIS. Export files can be compressed to support smaller file sizes. However, to ensure maximum compatibility it is usually better to use the uncompressed export format. The resulting files can then be compressed using a standard compression and archiving program such as PKZIP. The .e00 format is not published, but many other GIS packages have developed import routines.
- ArcView shape files (.shp) are a simpler format used by ESRI's ArcView desktop mapping software. A shape file database consists of several files containing the coordinate data, a spatial index and attribute data, respectively. Their file formats are published and many other GIS systems are able to import shape files.
- MapInfo interchange format (.mif) is used for the exchange of files produced with MapInfo, a leading desktop mapping system. MIF files are in ASCII format and can be read by many programs.
- MicroStation design file format (.dgn) is used by Bentley's modular GIS environment (MGE) and geographics GIS packages. The format does not support attribute data directly but provides links to external database tables. A separate export format combines geographic and attribute files.

3.83. All of these formats support boundary and attribute information. Any commercial GIS will have an import function for at least one or two of these formats. Ideally, a census office should offer its public-release GIS databases in several formats to serve a wide range of users with varying GIS skills and different software platforms. The choice of distribution formats should be guided by information on which mapping systems are most widely used in the census user communities and by the flexibility and robustness of the data format.

3.84 Distribution of GIS data in their native, internal format—for example, a directory containing an Arc/Info coverage or a MapInfo workspace—is not usually a viable option. Data in native formats often cannot be transferred to another operating system, 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.

ii. Tabular data

3.85. Most GIS packages 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 format, which can be produced by most database management and spreadsheet packages, as well as by census tabulation packages, such as REDATAM and IMPS.

3.86. 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.

iii. Documentation

3.87. Consideration must also be given to the file formats for the data documentation. Simple ASCII text files can be read by any user. However, they do not support graphics, complex tables or formatting of text. The Adobe Acrobat system's PDF format is now becoming a standard for platform-independent distribution of formatted documents. Since the Adobe Acrobat reader is available free of charge, PDF documents are likely to be accessible by any user.

3.88. An alternative is to produce documentation in a format readable by Web browsers, which are also available free of charge from Microsoft and Netscape. HTML files allow a considerable degree of formatting and when located on a CD-ROM or hard disk can be accessed even when no Internet connection exists.

iv. File-naming conventions

3.89. Although the Windows 95, NT, Mac and UNIX operating systems all support long file names, it is good practice to use the DOS 8.3 file-naming conventions for all data and documentation files that are distributed. Some users may be working under DOS or Windows 3.1 or with older GIS software packages. Short file names can reduce incompatibilities, for example, with older network software, to a minimum. Consistent naming conventions that are explained in the documentation will make it easier for users to find the data they need quickly.

v. Compression

3.90. GIS files are often very large and, together with the tabular data, the set of distribution files may become quite voluminous. Especially for Internet data delivery or for distribution on data diskettes, file compression will greatly facilitate data distribution. The most widely used compression software in the Windows environment is the PKZIP utility. It is available on most computers, and utilities that can extract files from the compressed archives also exist for the UNIX operating system. Self-extracting files are more convenient for inexperienced users and do not require a decompression utility. However, they are operating system specific and should only be used if the target computer platform is known.

© *Documentation and data dictionaries*

3.91. 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 chap. II). 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 metadatabase, the user-targeted data documentation can be compiled very quickly. Data documentation may include the following information:

- Data set names and reference information, including all data sources;
- Narrative content of the data sets;
- Description of the hierarchy of administrative and reporting units and their relationship to other features (e.g., settlements). This should include a clear description of the statistical definition used for each type of reporting units. A complete list of all reporting units and their geographic codes is useful;
- Software and hardware requirements;
- General data format, decompression and installation guidelines;
- Geographic referencing information (all geographic data sets should be in the same reference system):
  - Cartographic projection with all required parameters such as standard parallel or meridian, false easting and northing, and so on;
  - Coordinate units (e.g., decimal degrees, metres, feet);
  - Source map scale; that is, the scale of the hard-copy maps from which boundaries were digitized;
  - Geographic accuracy information. For instance, any numeric accuracy information available for the source maps can be reported. If a quantitative assessment of data quality is not possible, accuracy can 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 (for instance, districts that consist of several islands; see chap. II);
- 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, and so on.

3.92. In addition, each GIS data set should be accompanied by a data dictionary that provides information for each individual GIS data layer or data table. This should list the following information:

- File names and file formats;
- Feature types (points, lines or polygons);
- Relationship between coordinate data files and associated external attribute data tables;
- For each field in the attribute table and in additional external tables:
  - Field name;
  - Description of field content (e.g., total population, 1995) 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 "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 to a given task.

3.93. The data documentation and data dictionaries can also be incorporated into a comprehensive users guide. A users guide might contain a more detailed explanation of database content, data lineage and quality. Step-by-step explanations of example applications or copies of census maps created with the database can also be included. A sample data dictionary is presented in annex IV.

(d) Preparation of deliverables

3.94. Quality control is an important step before release of the final product for reproduction. After producing the final version of all databases in the form in which it will be distributed (e.g., compressed), the database should be tested on all target platforms (e.g., Windows environment, UNIX and Macintosh).

3.95. At the time of writing, CD-ROM is the most appropriate distribution medium for large data sets. A CD-ROM can hold up to 630 MB, and most computers are equipped with CD-ROM readers. CD writers are also quite inexpensive, so that digital masters can be produced in-house. This also allows distribution of customized data sets for which only a few copies are required. For wider distribution of large data sets, CD-ROM offers the advantage of low per-unit cost of production, durability and readability on multiple hardware platforms.

3.96. In the future two technologies may supersede CD-ROM technology. One is digital video/versatile disk (DVD) technology. A DVD can hold two gigabytes or more of data. DVD writing technology is also quickly progressing, although there is still some uncertainty about standards. These are very likely to be resolved over the next few years.

3.97. In the longer term, most data distribution will be done via the Internet. Currently, limited bandwidth—the amount of data that can be transferred in a given time period—is still hindering distribution of very large files. Download times are often unacceptable owing to shortcomings in the Internet infrastructure in many countries. The main bottleneck, however, is modem connections from homes or offices to the main Internet cables. Large files can be transferred to academic, government or commercial users that have dedicated high-speed Internet access.

3.98. Internet data distribution eliminates much of the cost of reproduction for the census organization. The remaining costs are for development of the software interface, maintenance of the Web site and incremental use of Web server resources. Census GIS databases can thus be provided to the user at very low cost or free of charge. Some organizations may, however, decide to charge for on-line data. One reason may be to cross-subsidize a publication program for users without access to the Internet. Another is where the organization intends to recover parts of the cost of data collection and compilation of the census data.

(e) Legal and commercialization issues

i. Data copyright

3.99. A copyright is the exclusive and legally guaranteed right to publish, reproduce or sell a piece of work—in this context, a digital geographic database. Because digital data are easy to reproduce, copyright issues concerning GIS databases are a more pressing issue than they have been for paper maps (see Antenucci and others, 1991). The census office thus needs to develop a data access policy for tabular as well as cartographic census information.

3.100. 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 licence agreement.

3.101. 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.

ii. Trade-offs in the commercialization of geographic data

3.102. Copyright laws differ from country to country. At one extreme, some Ggovernments 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, GIS data produced by public organizations are 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.

3.103. 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—ironically—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.

3.104. The benefit of this development has been a very wide use of census data for geographic applications. The increased number of users has in turn encouraged the commercial development of easy-to-use desktop mapping packages and the provision of value-added services. The overall economic benefits of this development are high as tax revenues increased and improved access to information led to productivity gains and better decision-making in the public and private sectors. These benefits have justified the royalty-free release of data, which was essentially a public subsidy for private companies.

3.105. 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. These prices may reflect the commercial value of such data to, for example, financial institutions and businesses. Yet, they may price small companies or non-commercial users out of the market for census information and may limit the overall use and therefore benefits of census GIS data. As Prevost and Gilruth (1997) point out, cost-recovery efforts that put census GIS products out of reach of non-commercial users often lead to illegal copying of data sets, time-consuming duplication of data development from original source materials, or use of alternative, cheaper and lower-quality data.

3.106. 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 owing 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 a realization that the benefits of charging higher prices do not justify the cost of enforcement of copyrights and of lost societal benefits because of the reduced use of vital information.

3.107. Access to the data and secondary uses are also often restricted where the census office collaborates with a private data producer or where data from public or private data producers are used to produce census maps. For instance, the census agency may enter into an agreement with a private mapping firm that absorbs part of the cost of digital map production for the census. The firm will only be able to recover its investment if it is awarded an exclusive right to market the geographic data (this will, of course, not be an issue where the agency simply purchases the services of the company, and all outputs remain the property of the census organization).

3.108. If data from other agencies—for instance the national mapping agency or local authorities—are used for producing census maps, pricing, copyright issues and the definition of source and credit information shown on the census maps need to be clarified in detail. Conflicts over copyright issues should be avoided especially, because the census agency will likely require the collaboration of those agencies for future census mapping activities.

3.109. In most countries, the trade-off between widest possible access to census data and the pressures to recover some of the cost of data collection will lead to a compromise between the two extreme positions just described. For instance, special arrangements can be made between government agencies that want 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, and so on, 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.

3.110. Several chapters in Rhind (1997) relate the experience of national mapping agencies in developing copyright and data distribution strategies for geographic information in a digital world. Onsrud (1992a and 1992b), Rhind (1992) and Onsrud and Lopez (1997) discuss the pros and cons of cost recovery and open access policies in the context of spatial databases.

iii. Liability issues

3.111. 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 resulting from missing or erroneous information on topographic maps. For instance, Lynch and Foote (1997) provide examples, where plane crashes and accidents at sea were caused by erroneous information on navigational maps. Map design and information content is guided by their intended use, but maps are sometimes used for purposes not anticipated by the data producer. For instance, to adapt an example used by Lynch and Foote (1997), a census organization might publish data for reporting units, together with a street network database. Because the road information is not critical to census data use, quality control of this information may have been much less rigorous than if the road information had been compiled for an emergency services routing system. If the imperfect data are used for such unintended purposes, damages may well occur.

3.112. Another example related to liability issues that is very relevant to census data dissemination is the violation of privacy of information. Usually, a census organization publishes only aggregate data at a level that does not reveal the information for an individual, a household or a very small group of persons. If the census organization reaggregates the microdata for several small area geographies—for example, enumeration areas, postcode sectors, health or education districts—there is a possibility that clever GIS operations can isolate information for groups of persons smaller than the lowest disclosure level (see sect. iv). In some countries, this may be grounds for legal action by the concerned individuals.

3.113. Interestingly, Johnson and Onsrud (1995) argue that selling GIS data and restricting secondary uses of data may increase liability by a data provider. The fee would imply a guarantee by the data provider that the material is error-free and fit for the purposes intended. Placing data into the public domain, in contrast, may shield the agency from such claims.

3.114. 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 (see, also, ESRI, 1995):

- A statement that the information was believed to be accurate at the time of collection and to have been obtained from reliable sources, but no warranty can be given as to the accuracy;
- Warnings that information is subject to change and notification of actual changes;
- If any parts of the geographic database were created by an external agency, this should be stated clearly;
- Mention that use of data implies acceptance of disclaimers and agreements.

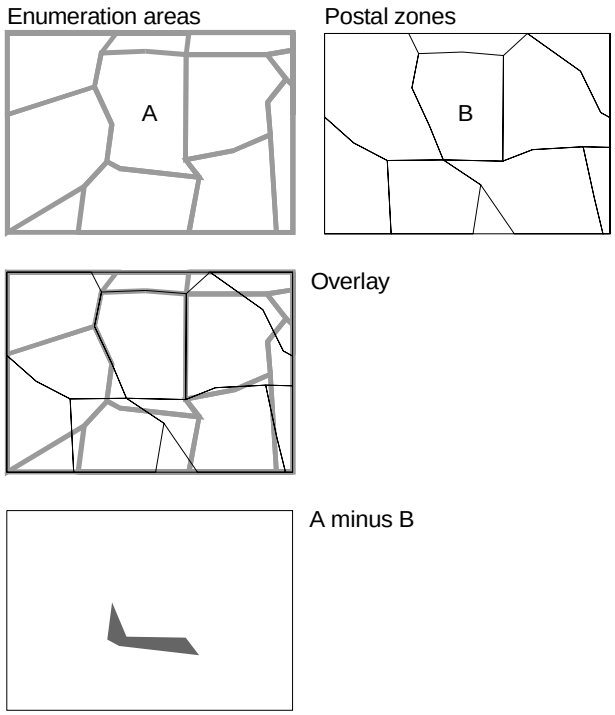
iv. Data privacy considerations: the differencing problem in statistical disclosure

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

3.116. 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.

3.117. 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 III.3 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 III.3: The differencing problem in statistical disclosure



3.118. To avoid data disclosure problems, the census organization should carefully review the boundaries of alternative census geographies. In instances where differencing appears possible, additional data protection must be introduced. Duke-Williams and Rees (1998) 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 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.

(f) Marketing of digital map products

3.119. Countries that aim at recovering some of the costs of developing census GIS databases and in which there is a strong commercial demand for small area statistical data may want to explore the possibility of entering into a marketing agreement with a private data vendor. Potential collaborators include the local distributors of the major GIS software producers. Most of the leading GIS vendors produce and sell GIS data sets on many topics. This is partly an additional revenue source and partly a way to facilitate the use of their software products by providing data sets in the software's data format. These private vendors sometimes collaborate with national mapping and statistical institutes to produce professionally designed GIS databases.

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

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

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

beneficial agreement will be the basis of a public-private or public-public partnership.

(g) Outreach

3.123. To ensure broad awareness of data availability and the widest possible distribution of georeferenced census data, the national statistical office may want to develop an outreach plan. Part of that plan could be printed brochures and posters featuring census maps. These can be widely distributed to schools, universities, commercial enterprises and national and local government offices.

3.124. 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.

5. Digital census atlases

3.125. While a more generic GIS database is targeted at users who have considerable experience in GIS, a digital census atlas is aimed at the general public, schools and other non-expert users. Two approaches for producing a digital census atlas are considered in the following paragraphs. A *static* census atlas consists of a collection of maps and other materials that have been prepared by the census office. It is essentially a presentation in which the user can change the sequence of viewing the content, but cannot change the content itself. A *dynamic* census atlas, in contrast, combines a digital GIS database and census data in a simple mapping package. The user can use the data to produce custom maps, which can be printed or copied into other applications packages.

(a) Static census atlases

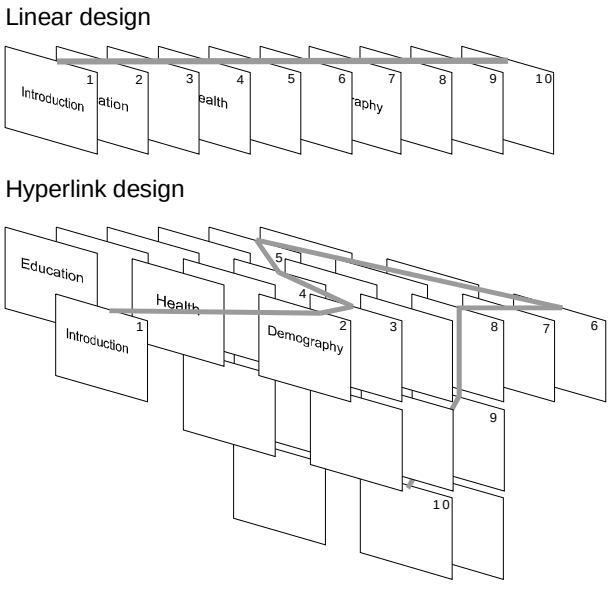
3.126. A static digital census atlas can bring together maps, tables, graphs and possibly multimedia products such as photographs or movie clips in a visually appealing, user-friendly environment. The presentation can be put together in a standard presentation software. Some presentation graphics packages 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, which can be distributed on computer-readable media or via the Internet. Maps can be produced in a desktop mapping package and incorporated into the presentation software using a graphical interchange format or simply the cut-and-paste commands in the Windows environment.

3.127. 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.

3.128. The presentation design may 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 may be taxed when they have to go through many slides with material that they are possibly not interested in.

3.129. Most presentation packages provide a better design option which is based on *hyperlinks*. These links allow the user to jump between different sections of the presentation. They also allow the user to integrate additional sources and information that may 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 III.4. Presentation design options for a static digital census atlas



3.130. The hyperlink concept is illustrated in Figure III.4, where it is contrasted with the linear design approach. In the hyperlink design, several parallel topics are presented, which are interconnected by links as appropriate. For instance, the three parallel story lines or chapters that follow the introduction page (1) could be

on education, health and demographic indicators. The user might follow a path—indicated by the grey 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 15. From here, links might be provided to maps showing child health indicators (4), educational facilities (5) and so on.

3.130. 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. An interesting overview of information design that uses these concepts is given by Wurman (1997).

3.131. Hyperlinks are, of course, familiar to anyone who has used the World Wide Web. In fact, rather than using a presentation software package, a static census atlas can also be implemented in the standard Internet browser language HTML. Web page design tools give the developer considerable flexibility in the design of the census database. One tool that can make the presentation more interesting, for example, is a clickable map. For instance, the entry screen might show an overview map of the country, with instructions to click on the province of interest for more detailed maps at the subnational level. Web technology also allows the inclusion of multimedia content and links to information outside 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.

3.133. One advantage of using Web design tools is that the same static census atlas can be distributed on CD-ROM or diskette for stand-alone use, and it can be posted on the census office's Web site for viewers anywhere in the world. More advanced Internet mapping applications are described in section 6.

(b) Dynamic census atlases

3.134. An alternative to a static census atlas is to publish a digital map and database, together with mapping software that allow the user to produce custom maps of census indicators. This, of course, requires some knowledge of cartography on the user side. A dynamic census atlas will include digital boundary files at a lower resolution than the full census database to allow fast drawing and low disk usage. The closely integrated attribute table should contain only a selected number of census indicators. Densities and ratios that are appropriate for mapping should already be calculated.

3.135. This approach will serve the needs of users who do not have the GIS expertise and skills required to make use of the complete digital census GIS database, but who want more flexibility in exploring and utilizing geographical census information than is possible with a pre-packaged static census atlas.

3.136. The problem, of course, is that such users may not have a desktop GIS package available that can be used to create maps. The data provider should therefore provide an easy-to-use package, together with the boundaries and data. The use of that package should require minimal training and experience. Essentially, the application should be "plug-and-play"—after installation, the user should immediately be able to produce maps.

3.137. Some census offices have developed map viewing software in-house and distribute these with their census data products. The maintenance of such programs is expensive, however, and binds resources that could otherwise be spent on data development or dissemination. Some GIS vendors are now selling GIS software tool kits that can be put together to produce custom applications or to integrate GIS functions in other software products (e.g., spreadsheets or database applications).

3.138. As an alternative, there are now several mapping packages available that are free of charge and can be distributed with a database. One of these is the PopMap software developed by the Software Development Project of the United Nations Statistics Division of the United Nations Secretariat, with funding from the United Nations Population Fund. PopMap is a desktop mapping package geared towards population application, even though any other information can, of course, be integrated as well. The system has geographic data input options (digitizing and drawing), a spreadsheet-like interface to manipulate attribute data and extensive cartographic mapping functions. The program is targeted at people who are not GIS experts and is thus easy to learn.

3.139. PopMap encourages the development of digital geographic census databases for distribution through its stand-alone mapping module. A census organization can produce a digital census atlas, package it together with the mapping software and distribute the resulting product to any interested user royalty free. An example application is the digital census atlas produced by the National Statistical Office of Uganda (see Box III.1).

Box III.1 The Uganda census atlas

3.140. Uganda carried out a population and housing census in 1991. After completion of census data processing, the census office decided to produce a digital census atlas with initial support and training from the United Nations Statistics Division's Software Development Project[a/]. Using a standard personal computer, a 12 by 18 inch digitizing board, a colour desktop printer and the PopMap desktop mapping software, two staff members, with the assistance of a technical adviser, produced digital maps for 38 districts, 163 counties and 809 subcounties. For each reporting zone, a selection of 36 census variables was compiled and integrated with the maps. For some indicators, data were also available for 1969 and 1980, allowing an analysis of change over time.

3.141. The census atlas was completed in less than 12 months. The census office produced a comprehensive users guide and has distributed the atlas to national and local government authorities and to private users. With relatively minor resources, the census office was able to provide a useful outlet for the census data, in addition to the printed census volumes.

[a]/See Vu and others (1994).

3.142. 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 ArcExplorer package produced by ESRI Inc. of (Redlands, California). ArcExplorer is a mapping interface for data created in the Arc/Info and ArcView GIS packages. In contrast to PopMap, the system does not have data input options and is thus strictly a map viewer.

3.143. The ArcExplorer interface is easy to use and the system provides basic mapping functions for producing thematic maps that can be exported as bitmaps or Windows metafiles. ArcExplorer 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 reside on remote Web sites. Analytical functions are limited, but the system does support different types of data query—interactive or using SQL-like commands—and address matching.

3.144. 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, emphasis in the documentation should be placed more on the attribute information and less on the technical geographic details.

6. Internet mapping

3.145. Many national statistical organizations have embraced the World Wide Web as a means to disseminate information and data. Web pages range

from simple lists and tables of census results to sophisticated query interfaces, in which the user can request special cross-tabulations.

3.146. The Internet is also suitable for presenting and distributing geographic information. The simplest option is to present static map images that were produced by the statistical office. For instance, a series of maps showing census variables can be produced using a desktop mapping package. Most packages allow the user to save maps in a standard image format such as GIF or JPEG. These images can then be integrated into Web pages just like any other graphic or photo. Such Web sites can give data users access to useful information. However, they do not allow the user to manipulate the data and to produce custom maps for specific geographic areas. The following sections concentrate on approaches that allow a significant degree of user interaction with the census geographic database.

3.147. A comprehensive discussion of Internet mapping options is provided by Plewe (1997) (see, also, ESRI, 1997). Foote and Kirvan (1997) present a more concise overview. Internet technology is changing fast, however, so the most up-to-date information is likely to be found on the Web sites of the major GIS software vendors, whose products are reviewed regularly in GIS trade journals.

3.148. 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.

3.149. 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 licences of commercial GIS packages that are run from a central server, staff members can access geographic information through their browser software.

3.150. There are three main options for implementing Internet mapping:

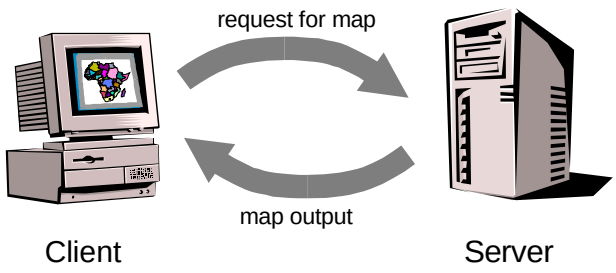
- In server-side strategies, the user sends a request for a map to the server holding the database. Mapping software on the server processes the request, produces a map—for example, in GIF format—and sends it back to the user.
- In client-side strategies, in contrast, most of the processing tasks are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

(a) Server-side approaches

3.151. Sometimes called "thin client/fat server" architecture, these strategies put most of the data-processing load onto the server that is located at the data distributing organization. This is similar to traditional mainframe architecture, where a powerful central computer handles data management, storage and processing for a number of users that are connected by dumb terminals.

3.152. The principle of a server-side strategy is summarized in Figure III.5. The user connects to a Web site and enters a request for a map. User-defined specifications for the output map include the geographic region of interest—which is specified either through the name of the region such as the district's name or through coordinates that form a bounding rectangle—, the variable to be mapped, the classification and colour scheme and additional data layers that provide context such as roads, rivers or administrative boundaries.

Figure III.5. Internet mapping – the server-side approach



3.153. The user's request is sent through the Internet to the server and routed to a GIS package. The GIS software can either be located on the Web server or it can reside 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.

3.154. The server-side approach has a number of advantages:

- The user does not need a powerful computer to access possibly very large spatial databases. Even fairly complex GIS procedures such as address matching or network routing can be carried out quickly if a powerful server is available. All that is required by the user is a basic Internet browser and an Internet connection.
- File sizes of the output maps in compressed image formats are much smaller than the database that would need to be transferred in client-side applications.
- Data integrity is maintained since the user cannot manipulate the database itself. The user is also always assured access to the most recent information.
- The data provider has more control over what users can see and how they can see it. Cartographic design choices can be pre-set to ensure that even non-expert users will obtain acceptable map output.

3.155 The Disadvantages are:

- Every change in map specification results in a new request. Even small changes in the geographic

region displayed (e.g., panning or zooming) will need to be requested specifically.

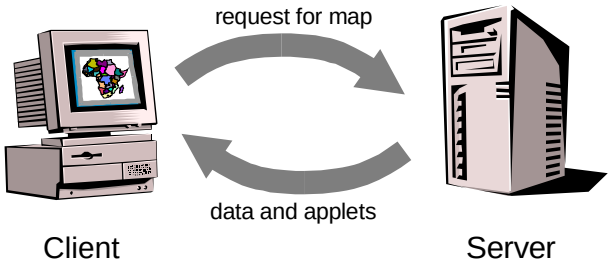
- On busy servers, repeated requests may be slow to execute when network traffic is high.
- Processing resources available on the user's computer are not utilized.

(b) Client-side approaches

3.156. Client-side approaches—thick client architectures—transfer much of the required processing to the user's computer. The server is mainly used to hold the database and send required pieces of the database, possibly together with mapping modules, to the user. Two variations of the client-side approach are available.

3.157. In the first, no mapping capability resides on the user's computer. After the user's request has been submitted, the server sends the geographic data, as well as a small program or applet that enables mapping or geographic analysis (see Figure III.6). 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 III.6. Internet mapping – the client-side approach



3.158. In an alternative client-side approach, a mapping package, applet or browser plug-in resides permanently on the user's computer. A plug-in is a program that extends the Internet 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.

3.159. The advantages of client-side approaches are:

- 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.

3.160. The Disadvantages are:

- Data and program files may be very large, requiring a fast Internet connection.
- Users with less powerful computers may not be able to execute more complex mapping and analysis tasks.
- Users with limited GIS or geography training may be unable to make use of the flexibility that client-side approaches can provide.
- Client-side approaches may allow users to save the raw geographic data that is requested from the server on their computer. This is a problem if some or all of the geographic data on the census bureau's server are copyrighted.

(c) Hybrid approaches

3.161. Server-side approaches are good at providing access to relatively simple maps to a large, non-expert audience. They would thus be most suitable for a census office's presentation of census maps to the general public. Client-side strategies, on the other hand, are preferable for Intranets, where a smaller number of users, with relatively comprehensive knowledge of GIS and mapping, access complex databases. They would thus be suitable for in-house GIS data access for census office staff.

3.162. 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

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

obtained that allows the data provider to custom-design the map server interface.

3.164. Capabilities of Internet mapping packages are likely to increase dramatically in the coming years. With increased network capacities, larger data sets and program modules can be transferred to users, and more users can be served simultaneously. The problems inherent in both client-and server-side solutions should be overcome with faster Internet connections. Client computers can have frequent communication with servers without delays leading to near instantaneous execution of user requests. In addition, limitations to the size of data sets that can be distributed should be reduced.

3.165. While current Internet mapping packages usually create GIF images that can be saved by the user, future packages are likely to support cashing or downloading of vector information to the user's computer. It will depend on the system whether the user will have access to this vector data or not. Clearly, this has ramifications for data copyright. If the census bureau charges for digital geographic data, access to an Internet map server that delivers vector data may be implemented on a fee basis.

3.166. For census data, the best Internet data access and distribution strategy will depend on the capabilities and expertise of the user. A flexible system will provide services for any level of user:

- "Power users" who want to obtain the entire database for use on their own computer using commercial GIS software. These users are served by conventional data distribution methods such as purchase of CD-ROMs or Internet download options of "raw" census GIS data sets;
- Active users with some expertise in GIS but who do not have local GIS capabilities. These users want to download parts of the database, together with GIS program modules (applets) that can perform the required tasks;
- Passive users who simply want to obtain a pre-designed map. The user request is executed by the server and the resulting information is sent to the user through the Internet in a suitable format—for example, raster image or postscript files for maps and spreadsheet or database files for the data.

3.167. A flexible census data distribution system on the Internet could look like the following:

- 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. The level of detail shown is determined by the map scale that corresponds to the current map extent on the user's screen;
  - A region that is defined by a geographic query. For instance, a commercial user who wants information about demographic characteristics of potential customers could request demographic information for a circular area of a 5 km radius surrounding a shopping centre location. A government planning agency may request data on the population living within 5 km of a proposed highway corridor;
- The user specifies the variables of interest and the type of output desired. Options may include maps for which the user can specify basic cartographic designs such as 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;
- The database server interprets the user's request and creates the appropriate subset of the database. For regions specified using geographic names, this will simply involve a logical selection of, for example, all census enumeration areas within a given district. For requested areas that do not match the standard census geographic hierarchy, some further processing is required. In some countries, dwelling unit GIS databases are now available or under construction, where every residence is associated with a geographic coordinate. A GIS on the server can then compile a custom tabulation by selecting all households that fall into the user-defined geographic area. In cases where this is not possible, the server-based GIS needs to perform an

areal interpolation, using techniques such as those described in section D below;

- The result of the query is returned to the user either as base data that can be manipulated further by the user using GIS applets, or as a map or database report that can be used directly by the user. Of course, in addition to the database or maps, data documentation and other relevant information must be available as well.

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

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

3.170. 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. An example is the Data Access and Dissemination System (DADS), referred as the "American fact finder" for the 2000 census of the United States. According to the United States Bureau of the Census design plans for DADS, geography is the integrating principle for the data, using both standard geographic areas and non-standard geography based on centroids or coordinates, as appropriate.

D. Advanced topics: geographic analysis of census data

1. Urban area definition/delineation

3.171. Definitions of urban versus rural areas vary widely from country to country (see United Nations, 1993). The most common approach is to use a population threshold to classify towns and villages into urban or rural settlements. The threshold population may be as low as 300 or as high as 5,000 people. A second approach that is frequently employed is to use a functional definition of an urban settlement. A town is classified as urban if it provides certain administrative, educational and commercial functions for a surrounding hinterland.

3.172. In some instances, it will also be useful to obtain a more general delineation of urbanized areas. Using administrative units as the basis for the urban/rural classification may mean that relatively large districts are entirely classified as urban even if they contain substantive agricultural or forest areas in addition to a major town or city. Some countries therefore also produce a finer delineation that groups only those areas that are densely settled and in which the majority of the population is not primarily engaged in agricultural activities. The resulting regions are variously called urbanized areas or densely inhabited districts.

3.173. If a comprehensive digital enumeration area database is available, GIS functions can aid the design of such areas. Ooishi and others (1998), for instance, describe an automated system used in Japan that groups basic unit blocks—the smallest area for which data are compiled—into larger regions. The system, which is implemented in a standard GIS package, uses a number of defined criteria based on a population density threshold and a contiguity constraint. An area threshold is used to determine whether an area that is surrounded by densely inhabited blocks should be included or left out. After aggregation of census blocks to densely inhabited districts, census staff can produce any number of summary census statistics for these areas that can be published in tabular or map form. Maps of these regions for two censuses can show increases or decreases in urbanized areas.

3.174. A more morphological approach to the delineation of urbanized areas is possible where recent air photos or satellite images are available for the most densely populated areas of the country. If these images are in digital form, urban areas can be delimited by tracing the boundary between built-up and agricultural, savanna or forest areas. Census staff can then derive statistics for these areas by overlaying enumeration areas and aggregating data either for all EAs that fall into the urbanized area or by using some form of areal interpolation, as described in the previous section.

2. Reconciling small area statistics with similar information from previous censuses

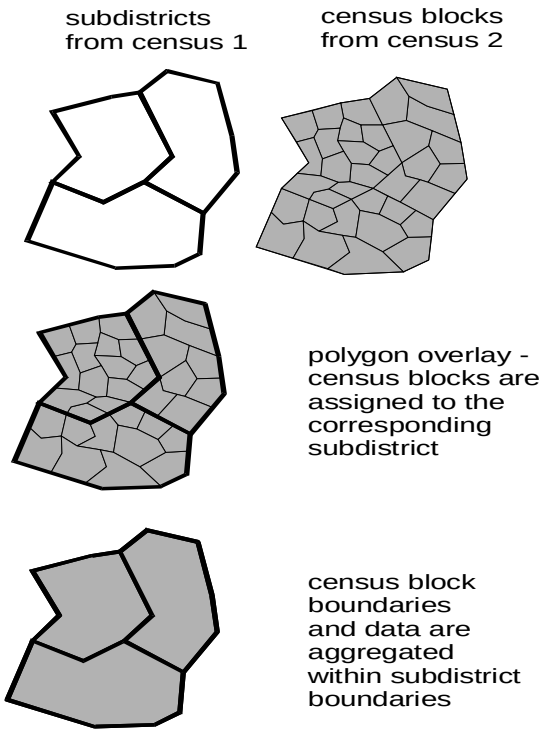
3.175. Censuses yield information about the demographic and social conditions in a country at one point in time. To obtain an indication of changes in the country, data from the current census need to be related to information from previous censuses. This is usually done for nationally aggregated indicators and perhaps for some large and relatively stable geographical subdivision such as states or provinces. However, local government agencies and private data users can also benefit from change information at the local level.

3.176. Unfortunately, administrative and census boundaries change over time. The lower the level of aggregation—that is, from province to district to ward to enumeration area—the larger the number of changes between censuses. To create time series of small area data, change maps at the local level, or summary statistics, the boundaries and data from two or more censuses need to be reconciled. The following paragraphs describe two options for doing this.

(a) Aggregation of old enumeration areas to new district boundaries

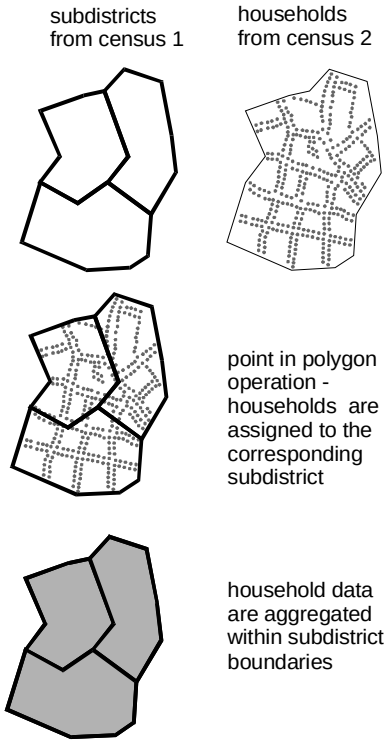
3.177. The task is relatively simple if there is a set of boundaries that are identical in both censuses—that is, a "lowest common denominator". For instance, in the most recent census, only the census block boundaries may have changed, but not the ward or subdistrict boundaries. Population data can then be compared for subdistricts simply by aggregating the census block data. However, census blocks may have been reassigned to different subdistricts between the censuses without actually changing the boundaries. In this case we have to determine into which subdistrict in census 1 each census 2 census block falls. GIS polygon overlay operations can help us in this task (see Figure III.7).

Figure III.7. Data aggregation when higher-level boundaries match



3.178. The task is even easier if the census office has produced a national household or dwelling unit database in GIS format. For each address, a point in a GIS database is available that is referenced to the census data in the microdatabase. The task of reconciling data for small area units is then simply a point in polygon operation in GIS, followed by an aggregation of the data for households that fall into the same reporting unit (see Figure III.8). Instead of point locations representing households, points could also represent centroids of small enumeration areas. Although the exact boundaries of these areas may not be available, aggregation to relatively large subdistricts or similar areas may yield reasonable estimates.

Figure III.8. Data aggregation with point data

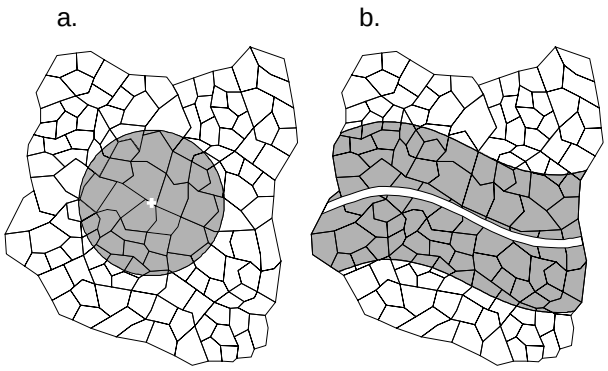


(b) Areal interpolation where boundaries are incompatible

3.179. If the boundaries of reporting units for the two censuses are not nested at some geographic level of aggregation, some form of areal interpolation is required to obtain compatible census data. Areal interpolation is the process of transferring data—for example, population totals—from one set of areal units to another, incompatible set of units.

3.180. The two sets of areas could be of the same type—for example, census units that have been considerably revised between two different censuses, or they could be very different—for example, where demographic census data need to be estimated for land cover zones or watershed areas. Areal interpolation is also required when a database query is defined as a spatial proximity operation. For instance to obtain demographic characteristics of the population residing within a circular distance around a point location such as a hospital (Figure III.9a.) or within a certain distance of a river that frequently floods (see Figure III9b), a buffer region is first determined. Demographic data available for small census units are then interpolated to derive data for the buffer region.

Figure III.9. Deriving data for areas that do not match reporting unit boundaries



3.181 Descriptions of areal interpolation techniques are given by Flowerdew and others (1991), Goodchild and others (1993) and Fisher and Langford (1995). In the following paragraphs, the set of zones for which data are available is termed *source zones*, while the second set of zones for which estimates need to be derived is termed *target zones*. Which areal interpolation method is most appropriate depends on whether we can assume that the variable is evenly distributed in the source zones, the target zones or a third set of zones—named *control zones*. The next paragraphs discuss these three cases. The example discussed refers to the interpolation of population values, but other variables can also be interpolated.

3.182. It is important to note, however, that *no interpolation method can provide error-free estimates* of target zone socio-economic indicators. In fact, the errors may often be unacceptably large for applications requiring high accuracy. Areal interpolation should thus be seen as a method of last resort, where more accurate options—such as reaggregation of small data collection units—are unavailable.

i. Source zones homogeneous

3.183. In the simplest case, we can reasonably assume that the source zones have a relatively constant population distribution. Census offices often design reporting units to be internally homogeneous. So this assumption is often quite reasonable unless extreme terrain conditions—steep mountains, swamplands or deserts-cause a very uneven distribution of the population. If the assumption of equal population

densities in each source zone is credible, we can assume that if, for example, 65 per cent of source zone A overlaps with target zone I, then 65 per cent of source zone A's population will reside in target zone I (see Figure III.10). In other words, population is distributed in proportion to the areas of overlap between source and target zones. The method is therefore also called areal weighting.

3.184. To illustrate the approach, consider Figure III.10 and Table III.2. The information we require—in addition to the population totals for each source zone—is the areas of overlap between source and target zones. A standard GIS polygon overlay operation produces these figures quickly. The GIS combines the polygons representing source and target zones and calculates the area of each new polygon. The first columns of Table III.2 show the resulting information. The area figures in square kilometres now need to be converted to proportions of overlap. For instance, 65 per cent of source zone A falls into target zone I and 35 per cent into target zone II. All that remains is to multiply these proportions of overlap by the source zone population figures to yield target zone estimates. For instance, target zone I's population estimate is 0.65 × source zone A's population plus 0.75 × source zone B's population (no part of source zone C's area overlaps with target zone I). The result is 28,500.

Figure III.10. Areal interpolation – homogeneous source zones

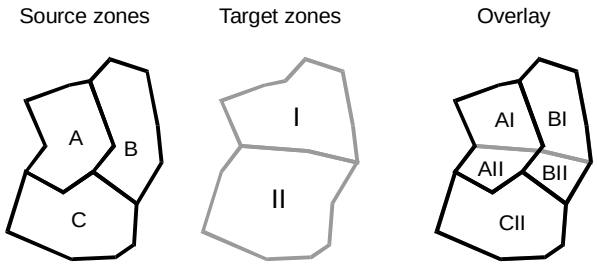


Table III.2. Areal interpolation – illustration of computations for homogeneous source zones

	Area of overlap (sq km)			Overlap proportions		Population in source zones	Population in target zones	
	I	II	Total	I	II		I	II
A	117	63	180	0.65	0.35	15 000	9750	5,250
B	150	50	200	0.75	0.25	25 000	18750	6,250
C	0	210	210	0.00	1.00	12 000	0	12,000
Total	267	323	590			52 000	**28 500**	**23 500**

3.185. This description illustrates the principle of areal weighting. In practice, there is an easier way of implementing this approach in a GIS;

- First, we compute source zone population densities by dividing the total population by the zone's area, which can be calculated by the GIS. The result is stored in a data field in the GIS database's attribute table.
- We then perform the polygon, overlay operation and let the GIS calculate correct surface area for each new polygon of intersection.
- For each new polygon we now derive a population estimate by multiplying its population density computed in the first step by the new area figure. All that remains is to sum the population figures for all areas of overlap that belong to the same target zone.

3.186. Densities are often not constant within areal units so there will always be an error in the target zone population estimates. This error can be quite significant, therefore, it depends on the specific conditions in the study area whether areal interpolation is appropriate or not.

3.187. If information about uninhabited areas is available, this can be incorporated into the areal weighting procedure by subtracting the empty areas first. For example, boundaries of lakes, agricultural lands, dense forests or other uninhabited areas might be

available from additional GIS layers that can be used to improve the target zone population estimates significantly. In cartography, this technique is called *dasymetric* mapping. To obtain more realistic population density figures, the cartographer masks unpopulated areas before producing a choropleth map (see for example, Plane and Rogerson, 1994).

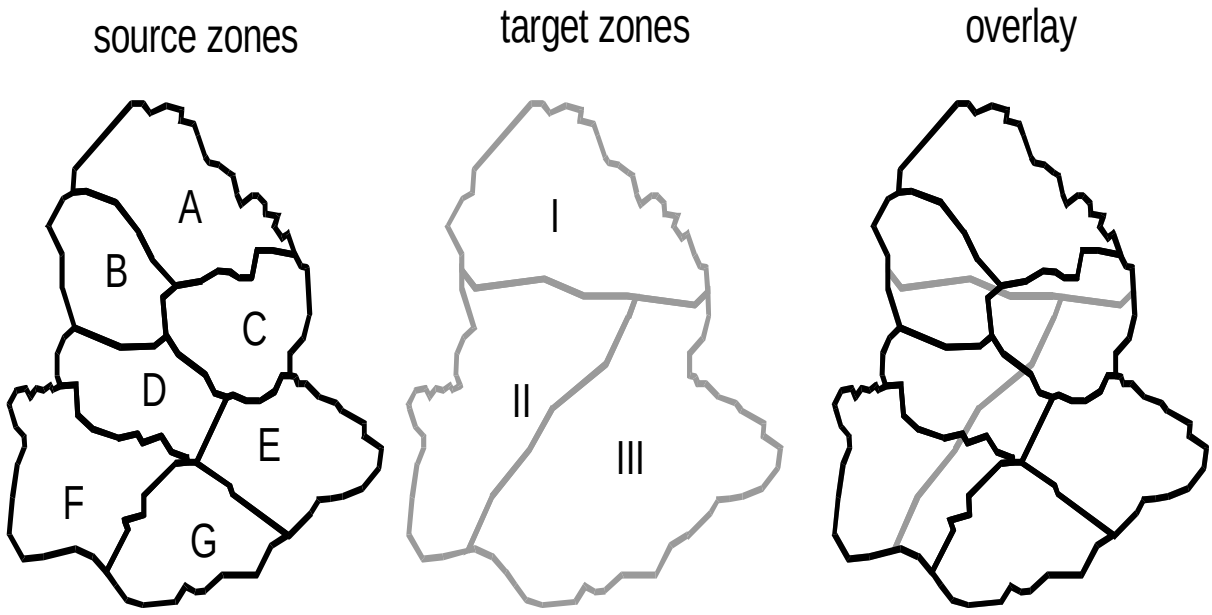
ii. Target zones homogeneous

3.188. Areal weighting will not produce very good results if the target zones rather than the source zones have constant densities. For instance, source zones may be fairly large and heterogeneous districts, while target zones represent land use or land cover classes. It is likely that land cover classes such as urban, agricultural and forest have a fairly uniform population distribution. Provided that the number of target zones is smaller than the number of source zones, we can produce target zone population estimates using statistical regression techniques.

3.189. More specifically, target zone population densities can be estimated as the coefficients of a linear regression through the origin—that is, without a constant term. The dependent variable is the source zone population, while the predictor variables are the areas of overlap between source and target zones. We can then derive target zone populations by multiplying the estimated densities by the corresponding areas of overlap and subsequent summation over all areas of overlap with the target zone. Regressions can be carried out in any spreadsheet or statistics package.

3.190. Figure III.11 And Table III.3 present an example. We need to estimate population totals for three target zones that are assumed to have a homogeneous population distribution. Total population is available for each source zone. The GIS overlay operation yields the areas of overlap between each source and target zone. The linear regression yields coefficient (population density) estimates of 16.0, 15.5 and 21.5. Multiplying these by the respective total target zone areas gives us population estimates of 18,179, 17,422 and 38,194. These values do not add up correctly to the known total population in the overall area. This is attributable to the error in the regression, which can be significant. A simple uniform adjustment can correct this problem.

Figure III.11. Areal interpolation – homogeneous target zones



TableIII.3. Areal interpolation – illustration of computations for homogeneous target zones

		Area of overlap (sq km)			
		Target zones			Total
Source zone	Population	I	II	III	area (sq km)
A	9 692	735	0	0	735
B	14 614	258	198	0	456
C	7 422	140	131	268	539
D	5 092	0	330	151	481
E	11 686	0	466	212	678
F	6 503	0	0	539	539
G	19 561	0	0	707	607
Total	74 570	1 133	1 125	1 777	4 035
Estimated densities		16.0	15.5	21.5	
Estimated population		**18 179**	**17 442**	**38 194**	

3.191. If the number of target zones is greater than the number of source zones, further assumptions need to be made or target zones need to be aggregated. In most cases, where densities do not vary extremely across administrative units, the regression can simply be performed in a spreadsheet program after importing the necessary data from the GIS. If extremely low densities are present in the study area, however, it may happen that negative regression coefficients result, implying negative population densities. In this case, specialized regression techniques need to be employed or we can set the densities of the affected target zones to zero or to some other externally estimated value (i.e., a constrained estimation).

iii. Control zones homogenous

3.192. Finally, in cases where neither source nor target zone densities can be assumed homogenous, we can incorporate auxiliary information such as a GIS database of units that are assumed to have constant densities. Examples are digitized land use maps (Moxey and Allanson 1994) or a classified remote sensing image (Langford and others, 1991). These control zones do not have to match either source or target zones. The control zone approach combines the regression estimation for homogeneous target zones with the areal weighting that is appropriate for homogeneous source zones.

3.193. Provided that the number of control zones is smaller than the number of source zones, the control zone densities can be estimated using a linear regression through the origin, as described in the previous section. With these estimated densities, we can then estimate target zone populations by using the areas of overlap between control zones and target zones. This is identical to areal weighting, as described earlier. In this second step, the number of target zones is not restricted by the number of source zones.

iv. Summary

3.194. The three variants of the areal interpolation technique outlined above, together with the extensions suggested in the literature cited, provide a comprehensive set of tools for transferring data from one set of areal units to another, incompatible set. It should be reiterated, however, that depending on the variability of densities within the study area and the quality of auxiliary information, the *errors inherent in the estimation can be quite considerable*. Collecting additional information at higher-resolution levels (if available) is thus always the preferable approach when high data accuracy is required.

3.195. A number of other methods of areal interpolation have been suggested. Some of these do not estimate the population of target zones directly. Instead, the population in each source zone is first distributed over a fine mesh of grid cells—that is, a raster GIS data layer—that is draped over the study area. Several rules for this process are possible.

3.196. Population could be assumed to be distributed very smoothly. In any given district, we would then expect that more people live in areas bordering other districts with higher population densities than close to those with lower densities. This is the principle of Tobler's smooth pycnophylactic interpolation (Tobler, 1979). "Pycnophylactic" means mass-preserving and it

means that the total population of all raster grid cells within a district will sum up to the known district total after the iterative interpolation has distributed population so that the surface generated is maximally smooth. Alternatively, the conversion of polygon to raster data can be guided by additional information such as land use, road infrastructure, settlement patterns and other indicators of population density. Bracken and Martin (1989), Martin (1991), Langford and Unwin (1994) and Deichmann (1996) review these and several other approaches in more detail.

(c) Temporal geographic information system databases

3.197. Over the long run, a census organization should attempt to minimize incompatibilities of statistical reporting units used for different censuses. This involves a targeted strategy to collect and manage data over time. The ideal outcome is a temporal GIS database that links boundaries and data collected at different points in time (see the comprehensive review by Langran 1992). There are three basic strategies to deal with changes in administrative or reporting area boundaries in spatially referenced databases:

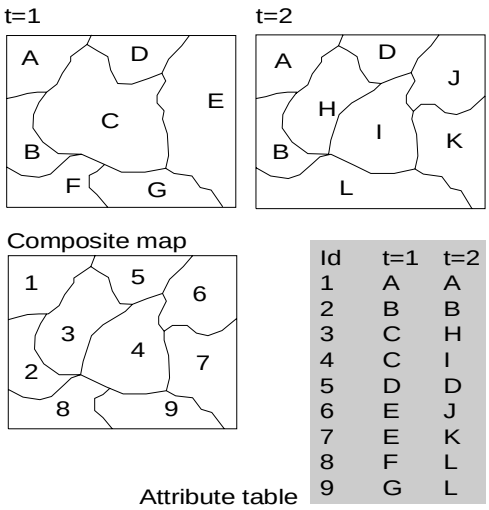
- Storing boundary data sets for each time period separately;
- Enforcing consistency of historical data with the latest available set of boundaries;
- Integrating information about the complete time series in the database.

3.198. In the first of these, the administrative boundaries for each census are stored in a separate GIS data layer. The disadvantage of this scheme is that there is no direct link between the data for different time periods. In order to calculate intercensal growth rates for units whose boundaries have changed over time, for example, significant additional manipulation is necessary.

3.199. The second approach involves the reconciliation of the data at different points in time to match the administrative unit boundaries for the latest available census. In cases where smaller units for a previous census are merged to form a new, larger reporting zone, the data can simply be aggregated as well. In most cases, however, it is more likely that districts were split between censuses or that completely new boundaries were introduced. In both cases, the construction of a consistent time series of data requires a data homogenization scheme. Either the individual districts are aggregated to the lowest level at which the two sets of boundaries match ("the lowest common denominator"), or some form of areal interpolation as described in the previous section.

3.200. A third option, a fully integrated spatio-temporal database, relies on storing the complete information about boundary changes over time within the database (see Figure III.12). In such a system, the spatial data set consists of a set of elementary polygons, each of which only belongs to one administrative unit at any given time. The elementary polygons form what is termed a *space-time composite* - that is, an overlay of all boundary data sets considered. Each polygon has a unique identifier and one or more entries in a transition table that record the time period in which the areal unit belonged to a specific administrative unit. For any given query, the system selects the appropriate records in the transition table and aggregates elementary polygons that belonged to the same county at the particular time. The resulting data set can then be linked to a specific data table for the corresponding census for mapping or further query.

Figure III.12. Simple spatio-temporal database



3.201. This data model maintains a log of the boundary changes over time. However, it does not solve the problem of creating consistent time series for the census indicators. Since the data tables for each census are in separate, often incompatible data tables, some form of areal interpolation is still required to compare census data over time.

3. Population data by grid cells

3.202. EAs or administrative units are represented in a GIS as irregularly shaped polygons. For some applications, the varying shape and size of these reporting units has some disadvantages. National statistical organizations in several countries have therefore developed census databases for regular grid cells (see Figure III.13). The grid cell sizes range from 100 metres used in the United Kingdom, to 1 kilometre

grids for Japan and the Republic of Korea, to 5 kilometres used for some international databases.

Figure III.13. Population density on a regular raster grid

3.203. One motivation for doing so in the past has been that raster data can be more easily stored and manipulated in a computer. Instead of storing boundary coordinates, raster GIS data consists basically of a long list of data values. A small header tells the computer how many data values (columns) are stored in each row, as well as the bounding coordinates and grid cell size in real-world units. Maps of raster GIS data could easily be produced on line printers simply by printing out the data values or a text symbol in a regular array of rows and columns.

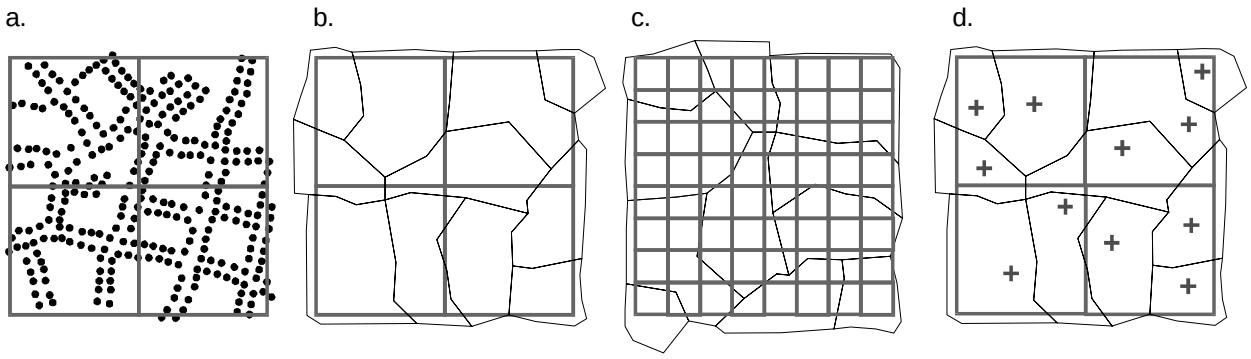
3.204. But gridded population data has some other advantages as well. Many environmental data sets are stored as raster data, including elevation and climate indicators. Analysis of population and environmental indicators is therefore greatly facilitated by storing both types of data as grids. Equivalent areas of all reporting units also provides, a more uniform appearance for thematic mapping. A good example is the gridded population density maps in the Population Atlas for China (Population Census Office, 1987) and the National Atlas of Sweden (Statistics Sweden, 1993).

3.205. There are several approaches for creating grid cells from census reporting unit data (e.g., Ohtomo, 1991). The most precise results will be achieved if individual households or housing units are allocated to grid cells (Figure III. 14a). In some instances this may be straightforward, specifically, if the census organization maintains a georeferenced address register that is linked to the census microdata set. With falling prices of GPS units one can imagine that more countries will produce such data, for example, by equipping each enumerator with a GPS during the census so that the exact coordinate of the household is captured, together with the household characteristics. If manual techniques are used, large-scale maps for each settlement are required, in which individual households can be identified. Clearly, this is an extremely labour-intensive task and few census offices will have the resources to create grid databases in this way.

3.206. A second option is to simply allocate census reporting units to a grid square, if more than half of its area falls into that cell (see Figure III.14b). On the other hand, a large enumeration area may coincide with several much smaller grid cells (see Figure III.14c. In this case, the EA data could be assigned in total to the grid cell that contains the population centroid of the EA. The population centroid must be assigned interactively. It defines a representative point in the EA that should coincide with the largest population concentration in the area. Alternatively, the data can be distributed evenly across all grid cells that fall into the enumeration area.

3.207. The centroids or representative points can also be used directly to assign EA data to grid cells. The analyst can choose to allocate the data of an EA to that grid cell into which its representative point falls. Using population-weighted representative points will generate better results than using GIS-calculated geometric centroids (see Figure III.14d).

Figure III.14. Alternative methods for producing population data for grid squares



3.208 Finally, we could also use the areal interpolation techniques described earlier to estimate census data for grid cells. Here, the source zones are the census enumeration areas and the target zone consists of a polygon data layer representing a regular grid. Provided that the enumeration areas are small, simple areal weighting should generate satisfactory results. As for all the other methods, this technique can be implemented using standard GIS overlay functions.

3.209 In terms of data storage, there are two options. One is to store the grid cells in vector format where each grid cell is essentially a square polygon. This allows easy storage of census indicators in the polygon attribute table of the GIS database. However, storage in this way means that the advantages of raster GIS—faster and easier processing, compatibility with environmental data sets—cannot be exploited. An option that combines the advantages of both, the relational database capabilities of vector format and the versatility of raster grids, is to use a GIS package that has the ability to manage attribute tables for raster grid cells. Only one grid is required, in which each cell is assigned a unique identifier. The identifier points to an attribute table that contains all census indicators. The GIS can then dynamically access any of these indicators for mapping or analysis.

Bibliography and references

Ahmed, M.M. (1996). Geographical information system (GIS) and its statistical applications in Egypt. *Journal of Economic Cooperation among Islamic Countries*, vol. 17, no. 1-2, pp. 25-39.

Antenucci, J.C., and others (1991). *Geographic Information Systems: A Guide to the Technology.* New York: Van Nostrand Reinhold.

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.

Batini, C., S. Ceri and S.B. Navathe (1992). *Conceptual Database Design. An Entity-Relationship Approach*. Redwood City, California: Benjamin/Cummings.

Batty, M. (1992). *Sharing Information in Third World Planning Agencies* NCGIA Technical Report 92-8. Buffalo, New York: National Center for Geographic Information and Analysis. (ftp://ftp.ncgia.ucsb.edu/pub/Publications/ Tech_Reports/92/92-8.PDF)

_____, D.F. Marble and A. Gar-On Yeh (1995). *Training Manual on Geographic Information Systems in local/regional planning.* Nagoya, Japan: United Nations Centre for Regional Development.

Becker, P., and others (1996). *GIS Development Guide*. Local Government GIS Demonstration Grant, Erie County Water Authority. Buffalo, New York: National Center for Geographic Information and Analysis, GIS Resources Group Inc. (www.geog.buffalo.edu/ncgia/sara/)

Ben-Moshe, E. (1997). Integration of a national GIS project within the planning and implementation of a population census. Euro-Mediterranean workshop on new technologies for the 2000 census round. Ma'ale Hachamisha, Israel, 16-20 March. (www.cbs.gov.il/mifkad/euromedit.htm)

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

Boehme, R. (1991). *Inventory of World Topographic Mapping.* Essex, United Kingdom: Elsevier Science Publishers.

Bond, D., and L. Worrall (1996). Geographical information systems, spatial analysis and public policy – the British experience. *Proceedings of the Fifth Independent Conference of the International Association for Official Statistics.* Reykjavík, 1-5 July.

_____, and others, eds. (1994). *GIS, Spatial Analysis and Public Policy.* Conference proceedings. Coleraine, United Kingdom: University of Ulster.

Bossler, J.D., and R.W. Schmidley (1997). Airborne system promises large-scale mapping advancements, *GIS World*, vol. 10, no. 6, pp. 46-48.

Bracken, I., and D. Martin (1989). The generation of spatial population distributions from census centroid data, *Environment and Planning A*, 21, pp. 537-543.

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

Broome, F.R., and others (1995). Automated mapping at the United States Census Bureau: 1980-1994 (parts I and II), *Cartography and Geographic Information Systems*, vol. 22, no. 2.

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

_____ (1997). Information technology operational plan for the decennial census 1998-2002., Washington, D.C.: United States Department of Commerce, Bureau of the Census. 7 November.

Bugayevskiy, L.M., and J.P. Snyder (1992). *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.

Carlson, G.R., and B. Patel (1997). New era dawns for geospatial imagery. *GIS World*, vol. 10, no. 3, pp. 12-15. (www.geoplace.com/print/gw/1997/ 0397feat.html)

Clarke, D. (1997). Mapping for the reconstruction of South Africa. In *Framework for the World*, D. Rhind, ed. Cambridge, United Kingdom: GeoInformation International.

Clayton, C., and J. Estes (1980). Image analysis as a check on census enumeration accuracy. *Photogrammetric Engineering and Remote Sensing*, no. 46, pp. 757-764.

Coiner, J.C. (1997). Transferability of the Qatar enterprise GIS model: experience in Viet Nam and Jamaica. *Proceedings GIS/GPS Conference 97*. Doha, 2-4 March. (www.gisqatar.org.qa/conf97/links/f3.html)

Cost, F. (1997). *Pocket Guide to Digital Printing*, Albany, New York: Delmar Publishers.

Dana, P.H. (1997). Global positioning system overview. Austin, Texas: The Geographers Craft Project (on-line). (www.utexas.edu/depts/grg/gcraft/notes/gps/gps_f .html)

Danko, D.M. (1992). The Digital Chart of the World Project, *Photogrammetric Engineering and Remote Sensing*, vol. 58, no. 8, pp. 1125-1128.

Deichmann, U. (1996). A review of spatial population database design and modelling, Technical Report TR 96-3. Santa Barbara, California: National Center for Geographic Information and Analysis. (ftp://ncgia.ucsb.edu/pub/Publications/ tech_reports/96/96-3/)

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

Duke-Williams, O., and P. Rees (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, pp. 579-605.

Eritrea National Statistical Office (1996). Preliminary design of census geographic processes. Asmara: National Statistical Office.

Espejo, A.B. (1996). The use of geographic information systems in Mexican censuses. *Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-scale Demographic Surveys*, 22-26 April. The Hague: Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund.

ESRI (1995). Data publishing guidelines for ESRI software, White paper. Redlands, California: Environmental Systems Research Institute. (available at www.esri.com)

_____ (1997). The future of GIS on the Internet, White paper. Redlands, California: Environmental Systems Research Institute. (available at www.esri.com)

EUROSTAT (1996). Statistics and Geography. *Sigma – The Bulletin of European Statistics* (summer).

Falkner, E. (1994). *Aerial Mapping: Methods and Applications*. Boca Raton, Florida: CRC Press.

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

_____ (1997b). The subcommittee on cultural and demographic data. Washington, D.C.: Federal Geographic Data Committee. (www.census.gov/geo/www/standards/scdd/ index.html)

Fisher, P.F., and M. Langford (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation, *Environment and Planning A*, 27, pp. 211-224.

Flowerdew, R., M. Green and E. Kehris (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science*, no. 70, pp. 303-315.

Foote, K.E. and A.P. Kirvan (1997). WebGIS. NCGIA Core Curriculum in GIScience. (www.ncgia.ucsb.edu/giscc/units/u133/u133.html, posted 13 July, 1998)

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

French, G.T. (1996). *Understanding the GPS.* Bethesda, Maryland: GeoResearch.

Gebizlioglu, L.Ö., H.M. Aral and N. Teksoy (1996). Impact of remote sensing on official statistics, *Journal of Economic Cooperation among Islamic Countries*, vol. 17, no. 1-2, pp. 1-23.

Geomatics Canada (1994). *National Topographic Database. Standards and Specifications.* Québec: The National Surveys, Mapping and Remote Sensing Organization, Natural Resources Canada.

GIS World (1998). *GIS Source Book*. Fort Collins, Colorado: GIS World. (updated annually since 1989), (www.geoplace.com)

Goodchild, M.F., L. Anselin and U. Deichmann (1993). A framework for the areal interpolation of socio-economic data. *Environment and Planning A*, 25, pp. 383-397.

Graham, L.A. (1997). Modern-day magic: options abound for raster-to-vector conversion, *GIS World*, vol. 10, no. 7, pp. 32-38 (www.geoplace.com/gw)

Hall, T., and others (1997). Comparison of GPS and GPS+GLONASS Positioning Performance. Proceedings ION GPS-97. Kansas City, Missouri, 16-19 September. (satnav.atc.ll.mit.edu/papers/timsept97/sep97tim.html)

Heine, G. (1997). Geographical information standards. Luxembourg: European Commission, Directorate XIII/E (www2.echo.lu/oii/en/gis.html)

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

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

Johnson, J., and H.J. Onsrud (1995). Is cost recovery worthwhile? *Proceedings of the Annual Conference of the URISA*, San Antonio, Texas, July. (www.spatial.maine.edu/onsrud.html)

Johnson, L.E. (1997). Factors for a successful implementation. *GIS World*, vol. 10, no. 2, p. 57.

Jones, C. (1997). *Geographical Information Systems and Computer Cartography.* Harlow, Essex, United Kingdom: Longman.

Kennedy, M. (1996). *The Global Positioning System and GIS: An Introduction*. Ann Arbor, Michigan: Ann Arbor Press.

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

Lang, A. (1997). Accuracy specifications affect application success. *GIS World*, vol. 10, no. 8, p. 58.

Lange, A. (1997). Put low-cost GPS receivers to the test, *GIS World*, vol. 10, no. 6, p. 36.

Langford, M., D.J. Maguire and D.J. Unwin (1991). The areal interpolation problem: estimating population using remote sensing in a GIS framework. In *Handling Geographical Information: Methodology and Potential Applications*, E. Masser and M. Blakemore, eds. Harlow, Essex, United Kingdom: Longman.

_____, and D.J. Unwin (1994). Generating and mapping population density surfaces within a geographical information system. *The Cartographic Journal*, no. 31, (21-25 June).

Langran, G. (1992). *Time in Geographic Information Systems*, London: Taylor and Francis.

Larsgaard, M.L. (1993). *Topographic Mapping of Africa, Antarctica and Eurasia.* Provo, Utah: Western Association of Map Libraries.

Leick, A. (1995). *GPS Satellite Surveying*, 2nd edition. New York: John Wiley and Sons.

Li, L. (1997). Dwelling frame feasibility study. *GIS/LIS Proceedings.* Denver, Colorado, 19-21 November.

Lillesand, T.M. and R.W. Kiefer (1994). *Remote Sensing and Image Interpretation*, 3rd edition. New York: John Wiley and Sons.

Lo, C.P. (1986). *Applied Remote Sensing*, London: Longman.

_____ (1995). Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. *International Journal of Remote Sensing*, vol. 16, no. 1, pp. 17-34.

Lynch, M., and K.E. Foote (1997). Legal Issues Relating to GIS: The Geographer's Craft Project. Austin: University of Texas. (wwwhost.cc.utexas.edu/ftp/pub/grg/gcraft/contents.html)

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.

Martin (1991). *Geographic Information Systems and their Socio-economic Applications.* London: Routledge.

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

Michael, J. (1997). Digital orthophotography – Principles, project design, issues, utility, accuracy, economics, *Proceedings GIS/GPS Conference 97.* Doha, 2-4 March. (www.gisqatar.org.qa/conf97/links/h1.html)

Misra, P. (1993). Integrated use of GPS and GLONASS in civil aviation. The Lincoln Laboratory Journal, vol. 6, no. 2, pp. 231-248. (satnav.atc.ll.mit.edu/papers/LLjournal/Misra.html)

Moellering, H., and R. Hogan, eds. (1997). *Spatial Database Transfer Standards 2: Characteristics for Assessing Standards and Full Descriptions of the National and International Standards in the*

World. Amsterdam: International Cartographic Association, Pergamon, Elsevier Science.

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

_____ (1996). *How to Lie with Maps*, 2nd edition. Chicago: University of Chicago Press.

Montgomery, G.E., and H.C. Schuch (1994). *GIS Data Conversion Handbook.* Fort Collins, Colorado: GIS World.

Moxey, A., and P. Allanson (1994). Areal interpolation of spatially extensive variables - A comparison of alternative techniques. *International Journal of Geographic Information Systems*, vol. 8, no. 5, pp. 479-487.

Murray, J.D., and W. van Ryper (1994). *Encyclopedia of Graphics File Formats.* Sebastopol, California: O'Reilly & Associates, Inc.

National Research Council (1997). *The Future of Spatial Data and Society: Summary of a Workshop*. Washington, D.C.: National Academy Press. (www.nap.edu/readingroom/books/spa)

NCGIA (1998). *GIS Core Curriculum.* Santa Barbara, California: National Center for Geographic Information and Analysis. (www.ncgia.ucsb.edu/giscc)

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

NIDI (1996). *Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-scale Demographic Surveys*, 22-26 April. The Hague: Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund. (www.nidi.nl/innotec/index.html)

Nordisk Kvantif (1987). *Digital Map Data Bases. Economics and User Experiences in North America*. Arendal, Norway: Joint Nordic Project – Community Benefit of Digital Spatial Information, VIAK A/S.

_____ (1990). *Economics of Geographic Information.* Helsinki: National Board of Survey.

Ohtomo, A. (1991). Small area statistical databases. Second Interregional Workshop on Population Databases and Related Topics. Jakarta, 14-19 January. New York: United Nations Department of Technical Cooperation for Development and the Statistical Office.

Onsrud, H.J. (1992a). In support of open access for publicly held geographic information. *GIS Law*, 1992, vol. 1, no. 1, pp. 3-6. (www.spatial.maine.edu/onsrud.html)

_____ (1992b). In support of cost recovery for publicly held geographic information. *GIS Law*, 1992, vol. 1, no. 2, pp. 1-7. (www.spatial.maine.edu/onsrud.html)

_____ and X. Lopez (1997). Intellectual property rights in disseminating digital geographic data, products, and services: conflicts and commonalties among European Union and United States approaches. In *Geographic Information: The European Dimension* I. Masser, and F. Salge, eds., London: Taylor and Francis.

Ooishi, T., and others (1998). Automated census system for densely inhabited districts. Proceedings of the Eighteenth Annual ESRI Users Conference. Redlands, California: Environmental Systems Research Institute.

Open GIS Consortium (1996). *The OpenGIS Guide: Introduction to Interoperable Geoprocessing*. Wayland, Massachusetts: Open GIS Consortium, Inc. (www.opengis.org)

Openshaw, S., ed. (1995), *Census Users Handbook*. Cambridge, United Kingdom: Geoinformation International.

Ordnance Survey (1993). *Address-Point User Guide.* Southampton, United Kingdom: Ordnance Survey.

Padmanabhan, G., J. Yoon and M. Leipnik (1992). *A Glossary of GIS Terminology*. Technical Report 92-13. Santa Barbara, California: National Center for Geographic Information and Analysis.

Paulsen, B. (1992). *Urban Applications of Satellite Remote Sensing and GIS Analysis*, Urban Management Programme Discussion Paper No. 9. Washington, D.C.: World Bank.

Pazner, M., N. Thies and R. Chávez (1994). *Simple Computer Imaging and Mapping.* London, Ontario: Think Space, Inc.

Plane, D.A., and P.A. Rogerson (1994). *The geographical analysis of population.* New York: John Wiley and Sons.

Plewe, B. (1997). *GIS Online: Information Retrieval, Mapping and the Internet.* Santa Fe, New Mexico: OnWord Press.

Population Census Office (1987). *The Population Atlas of China*. Hong Kong: Population Census Office and Institute of Geography, Chinese Academy of Sciences, Oxford University Press.

Prévost, Y., and P. Gilruth (1997), *Environmental Information Systems in Sub-Saharan Africa*. Building Blocks for Africa 2025, Paper No. 12. Washington, D.C.: World Bank and New York: United Nations Development Programmes and United Nations Statistical Office.

Rajani, P. (1996). Simple models reflect GIS market segmentation, *GIS World*, vol. 9, no. 12, p. 130.

Rhind, D. (1991). Counting the people: the role of GIS. In *Geographical information systems - principles and applications*, D.J. Maguire, M.F. Goodchild and D.W. Rhind, eds., vol. 1, pp. 127-137. London: Longman.

_____ (1992). Data access, charging and copyright and their implications for geographical information systems. *International Journal of Geographical Information Systems*, vol. 6, no. 1, pp. 13- 30.

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

Ritter, N. (1996). The GeoTIFF Web Page. (http://home.earthlink.net/~ritter/geotiff/geotiff.html)

Robinson, A.H., and others (1995). *Elements of Cartography*, 6th edition. New York: John Wiley and Sons.

Romano, F.J. (1996). *Pocket Guide to Digital Prepress*. Albany, New York: Delmar Publishers.

Satellitbild (1994). Reference project: national population census in Nigeria. Kiruna, Sweden: Swedish Space Corporation. (www.ssc.se/sb/ssc_sb.html)

Schmidt, J.J. (1996). Evaluation of hand-held GPS for registration of cadastral maps. *GIS/LIS Proceedings.* Denver, Colorado, 19-21 November.

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

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

Statistics Sweden (1993). *National Atlas of Sweden.*, Stockholm: SNA Publishing.

Steffey, D.L., and N.M. Bradburn, eds. (1994). *Counting People in the Information Age.* Washington, D.C.: National Academy Press.

Suharto, S., and D.M. Vu (1996). Computerized cartographic work for censuses and surveys. *Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-scale Demographic Surveys.*, 22-26 April. The Hague: Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund. (www.un.org/Depts/unsd/softproj/papers/sv01.htm)

Thygesen, L. (1996). GIS and official statistics – Synergy or clash? *Proceedings of the* Fifth *Independent Conference of the International Association for Official Statistics.* Reykjavík, 1-5 July.

Tobler, W.R. (1979). Smooth pycnophylactic interpolation of geographical regions. *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 519-530.

Tripathi, R.R. (1995). Updating and improvement of census base maps using global positioning systems. Presented at the TSS/CST Workshop on Data Collection, Processing, Dissemination and Utilization, New York, 15-19 May.

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

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

Tveite, H., and S. Langaas (1995). Accuracy assessments of geographical line data sets: The case of the Digital Chart of the World, *Proceedings from the Fifth Scandinavian Research Conference on Geographical Information Systems.* Trondheim, Norway, 12-14 June. (see, also, http://ilm425.nlh.no/gis/dcw/dcw.html)

United Nations (1988). *The Geography of Fertility in the ESCAP Region.* Asian Population Studies Series, No. 62-K. Bangkok: Economic and Social Commission for Asia and the Pacific.

_____ (1993). *World Urbanization Prospects.* Sales No. 93.XIII.II.

_____ (1997a). *Geographical Information Systems for Population Statistics.* Studies in Methods, No. 68. Sales No. E.97.XVII.30. (www.un.org/Depts/unsd/demotss/intro2.htm)

_____ (1997b). *PopMap – Users' Guide and Reference Manual.* (www.un.org/Depts/unsd/softproj/index.htm)

_____ (1997c). MapScan Manual. (www.un.org/Depts/unsd/softproj/index.htm)

_____ (1998). *Principles and Recommendations for Population and Housing Censuses, Revision 1*. Statistical Papers, No. 67/Rev. 1. Sales No. E.98.XVII.8.

United Nations Environment Programme (1997). *A Survey of Geographic Information Systems and Image Processing Software.* Sioux Falls, South Dakota: Environmental Assessment Program. (http://grid2.cr.usgs.gov/survey)

Vu, D.M. (1996). PopMap: geographical census software for developing countries, *Proceedings of the Expert Group Meeting on Innovative Techniques for Population Censuses and Large-scale Demographic Surveys.* 22 - 26 April. The Hague: Netherlands Interdisciplinary Demographic Institute and United Nations Population Fund. (www.un.org/Depts/unsd/softproj/papers/vdm961.htm)

_____, P. Gerland and D. Castillo (1994). PopMap – a step toward better utilization and dissemination of population data – case study of a national census atlas. Working Paper No.37. Work Session on Geographical Information Systems, 27-30 September. Voorburg, Netherlands: Statistical Commission of the Economic Commission for Europe, Conference of European Statisticians.

Waldorf, S.P. (1995). Commercial cartography: custom design and production, *Cartography and Geographic Information Systems*, vol. 22, no. 2.

Waters, H. (1995). Feasibility studies for mapping projects in developing countries. *The Cartographic Journal*, no. 32, (December), pp. 143-147.

Wood, C.H., and C.P. Keller, eds. (1996). *Cartographic Design: Theoretical and Practical Perspectives*. New York: John Wiley and Sons.

Worrall, L. (1994). Justifying investment in GIS: a local government perspective, *International Journal of Geographical Information Systems*, vol. 8, no. 6, pp. 545-565.

Wurman, R.S. (1997). *Information Architects*. New York: Graphics Inc.

Annex I. Geographic information systems

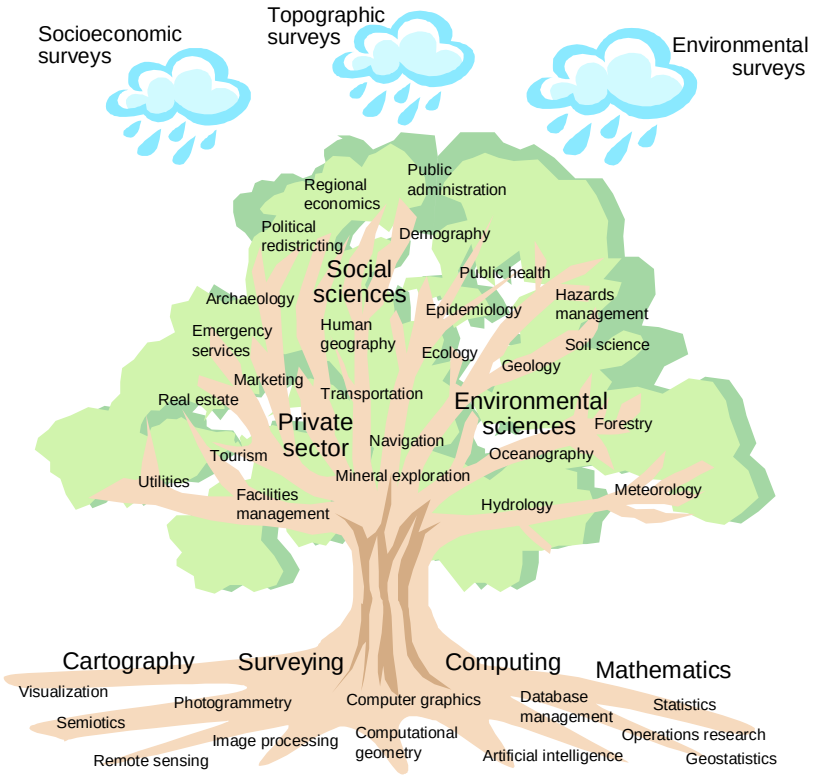
Geographic information systems overview

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.

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 geographic data. For example, there are now GIS degree courses at several universities; the term "geographic information science" is increasingly used to refer to academic research on computer-based geographic software and procedures.

Several fields have contributed to the foundations of GIS, as illustrated in figure A.I.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 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.

Figure A.I.1. Foundations of GIS
(after Jones, 1997)



Inventory-type applications are found in the utilities sector, where, for example, a phone company manages and maintains its physical infrastructure using a GIS database. Land titling systems operated by local and regional government agencies are another example. In some fields, GIS is used to support data collection. The use of digital mapping 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, as well as 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.

1. Hardware, software and data

Hardware and software issues are discussed in chapter II in the context of census mapping. In general, 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.

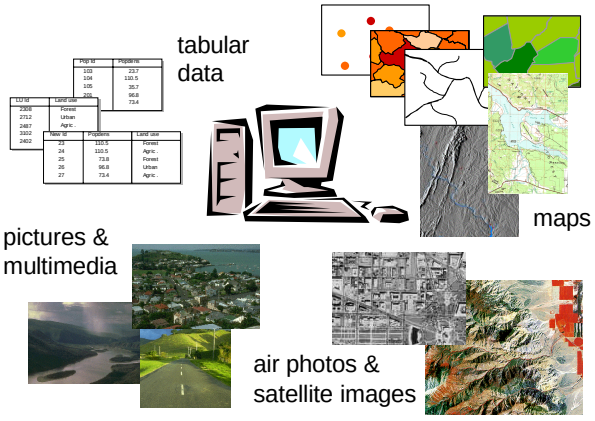
GIS software has developed rapidly in recent years from command-line oriented systems that were very hard to learn, to menu-driven, easy-to-use packages that can be employed by anyone, with minimum training. High-end packages are used by GIS analysts who create new databases and carry out advanced spatial analysis. At the medium level, there are now a number of desktop mapping packages that combine a standard Windows 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 packages 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.

One recent development is toolboxes of GIS-related software routines or “objects” sold by several GIS vendors, which 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.

Current GIS software trends appear to be focused on two aspects: Internet mapping and modular design that allows integration of GIS functions in any application. Users may soon be able to carry out GIS data query and analysis on remote geospatial databases, using their Web browser and software that is downloaded on demand. For high-end applications, there may be a further 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 may thus gradually disappear.

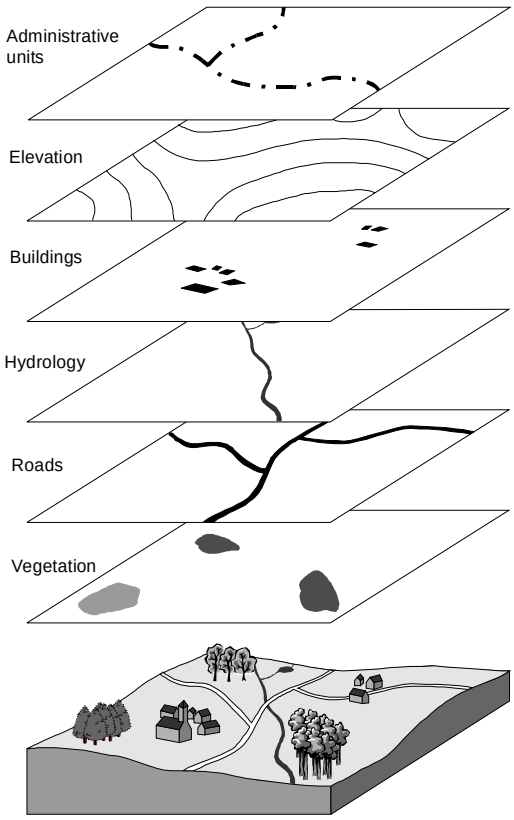
Data are what fuels GIS applications (see figure A.I.2). Many of the most common GIS data sets are digital equivalents of paper maps such as topographic maps showing roads, rivers, elevation contours and settlements. Thematic information includes socio-economic attributes referenced by administrative 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 great detail in a data table that is 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 A.I.2. Types of information stored in a GIS



Another source of geographic information is remote sensing. Pictures or images taken from low-flying aircraft or from satellites can be 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 can also be integrated in GIS. 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.

Figure A.I.3. Data layers—space as an indexing system



2. Geographic data layers

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. 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 (see figure A.I.3).

A comprehensive GIS database will 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 that 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.

In the process of creating a multi-layer GIS database, features may 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 that 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 (see section C). The ability of integrating 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 geographic information system.

GIS data models

Despite the heterogeneity of the information that can be stored in a GIS, there are only a few common methods of representing spatial information in a GIS database. 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: the *vector* data model, which is used to symbolize discrete features such as houses, roads or districts, and the *raster* data model, which 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 plane-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

Vector GIS systems represent real-world features using a set of geometric primitives: points, lines

and polygons (see figure A.I.4). A point is represented in a computer database by an x,y coordinate. A line is a sequence of x,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 may 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.

The simplest vector data models store the data without establishing relationships among the geographic features (see figure A.I.5). This is sometimes called the "*spaghetti model*" (e.g., 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.

Figure A.I.4. Points, lines and polygons

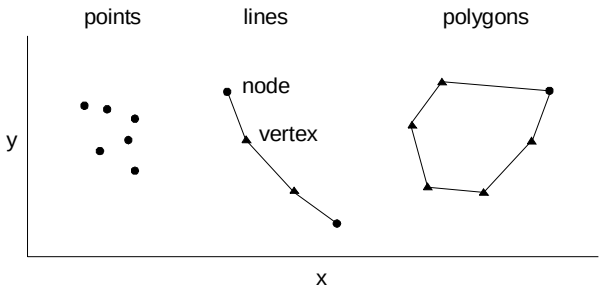
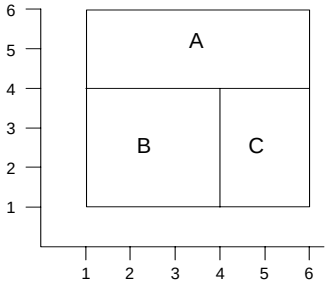


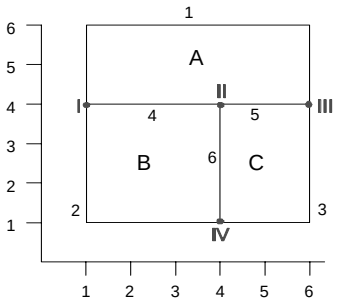
Figure A.I.5: Vector data models: spaghetti versus topological

"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

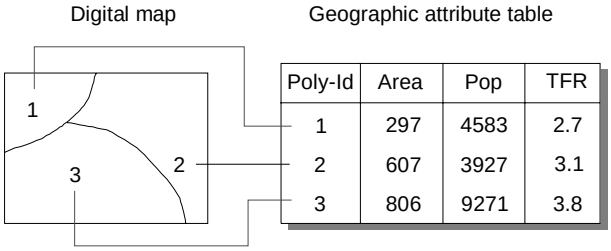
Poly	Lines
A	1,4,5
B	2,4,6
C	3,5,6

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

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 fast queries on individual data objects and their relation to other data objects. For example, to quickly identify all neighbours of a particular enumeration area, the system would simply go through the list of lines that define this EA and find all of the remaining EAs that are also bounded by these lines.

High-end GIS packages 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.

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



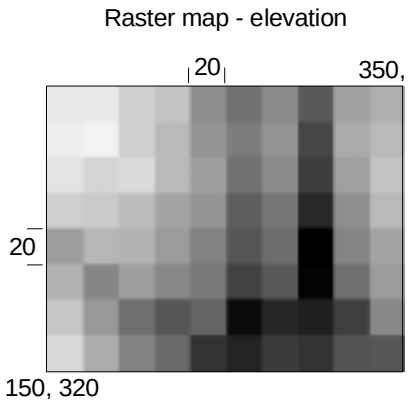
Every feature in the database is labelled internally with a unique identifier that links the geometric feature with a corresponding entry in a data or attribute table (see figure A.I.6). The user can add information about each feature to the corresponding database record. For points representing houses, 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 (chapter II discusses these issues in more detail). 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.

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

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 and 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 A.I.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 A.I.7. Example of a raster data file



ASCII raster file

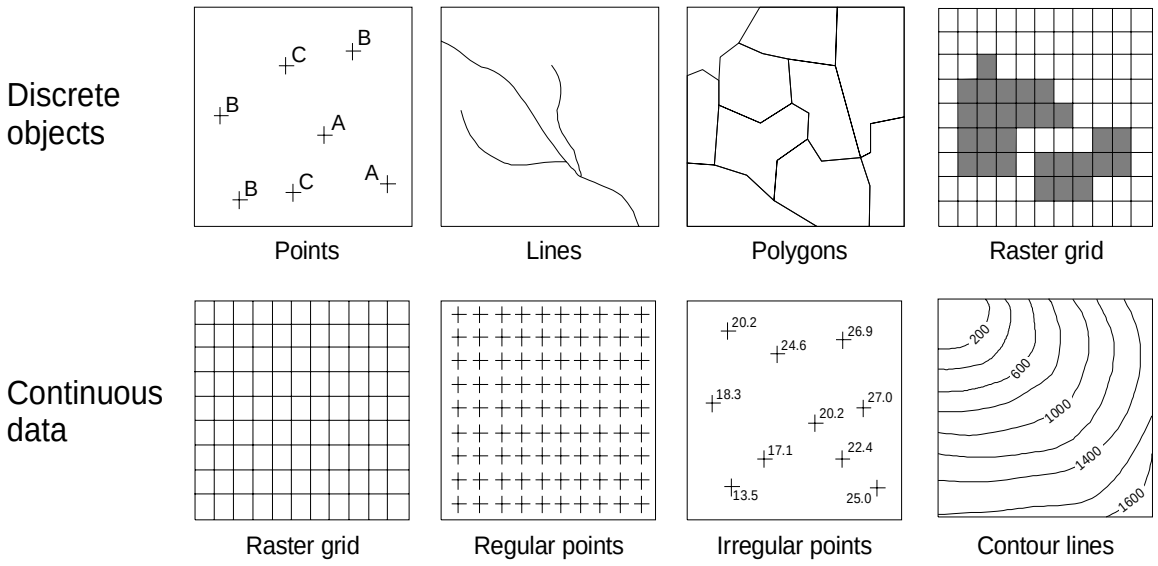
Number of columns	10
Number of rows	8
Lower left corner x	150
Lower left corner y	320
Cell size (square cells)	20

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

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 systems 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.

Raster data are most often used to store continuously varying data or images that show many continuous grey 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 is contour lines, which show elevation on topographic maps. Other examples are shown in figure A.I.8.

Figure A.I.8. Vector and raster can both be used to display discrete and continuous data



3. Advantages and disadvantages of vector and raster data models

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.

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 that 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.

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 area boundaries come in raster form. Chapter II, 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

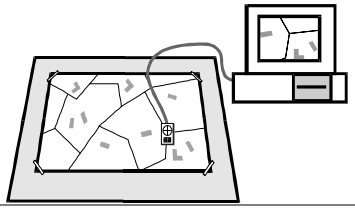
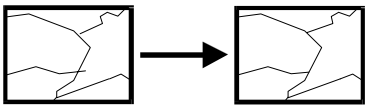
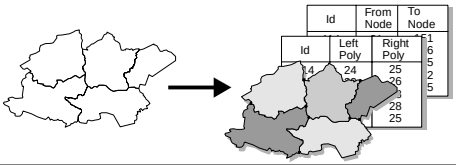
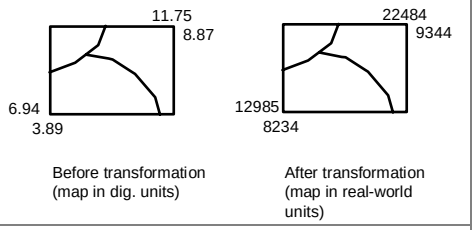
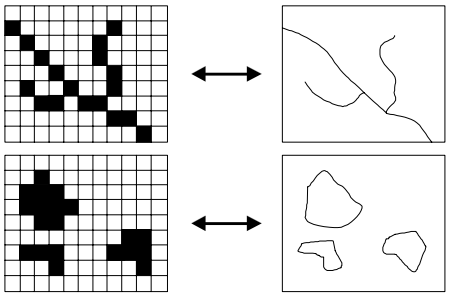
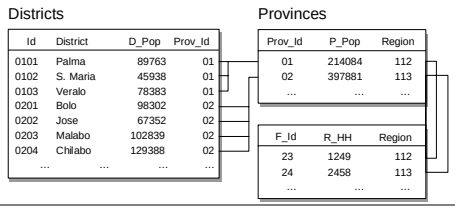
The terms precision and accuracy are often used interchangeably, even though they have different meanings. Accuracy means freedom from error. In a spatial context, for example, an accurate point coordinate in a GIS database is registered at the correct place with respect to the point's true location 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 accurate to the nearest metre. If we have a more precise measuring tool, we can obtain point coordinates that are accurate to the nearest centimetre or millimetre.

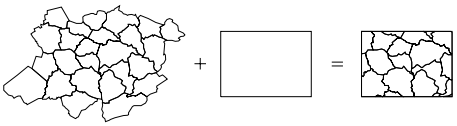
In practice, the precision by which coordinates can be stored in a vector GIS is virtually infinite, because they use double precision data types (8 bytes for each floating point number) for storing the geographic coordinates. The accuracy of spatial coordinates, however, depends largely on the data 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.

GIS capabilities

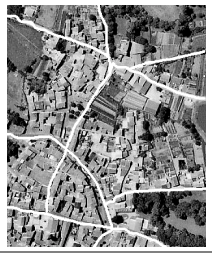
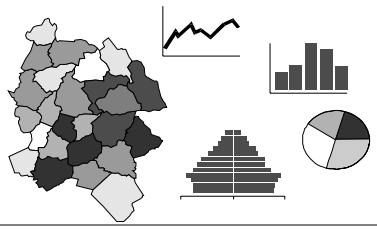
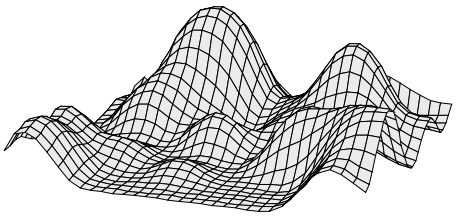
The following table provides an overview of GIS capabilities. The list is by no means complete, since high-end GIS packages and even desktop mapping packages offer numerous specialized functions for data entry, manipulation, analysis and display.

Data input and management

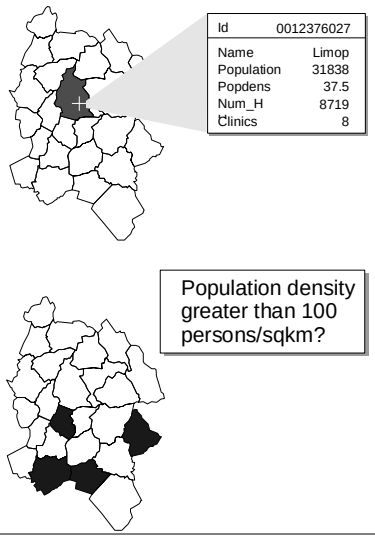
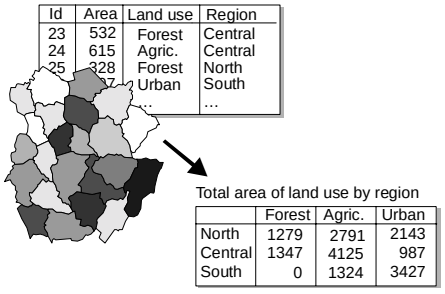
Line tracing, coordinate data input	Still the most common form of coordinate data entry is by means of a digitizing table. Lines are traced on the paper map with a cursor and captured in the GIS or digitizing software. Alternatively, maps can be scanned to create raster bitmaps that are then converted 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 relationship 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 conversion	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 conversion is also used for the automatic conversion of scanned maps. The opposite operation—vector to raster—is required for analysis and modelling in a raster GIS.	
Attribute data management	Each feature in the database is labelled 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.	

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, cookie-cutting	Apart from subset selection based on queries, GIS can also create custom subsets using so-called cookie-cutting operations.	

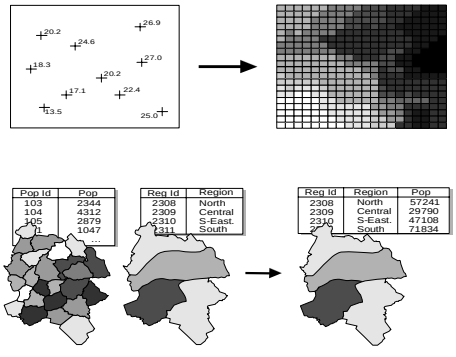
Display

	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.	Cartographic functions
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 valuable, especially 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 3-d visualizations, using wire frames onto which other features can be draped.	

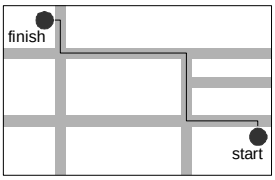
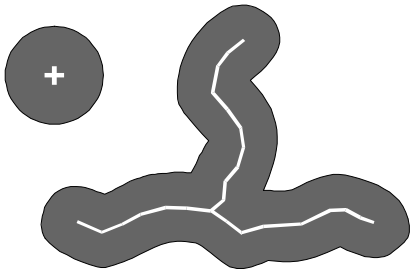
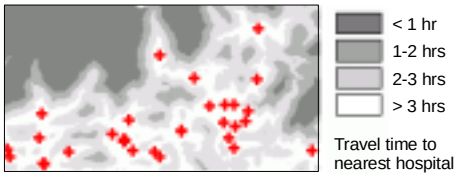
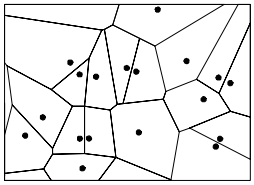
Query

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. GISs 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

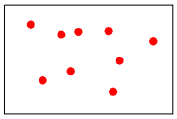
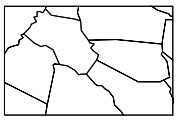
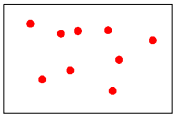
Interpolation	Sometimes called *basis change*, interpolation allows us to create a complete coverage from sample data. For example, based on a set of station precipitation surfaces, we can create a raster surface that shows rainfall in the entire region. More important for socio-economic applications is so-called areal interpolation. For example, using population by district, we would like to estimate population for environmental monitoring regions whose boundaries do not match the districts.	

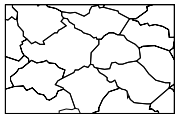
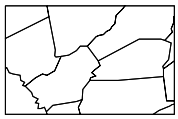
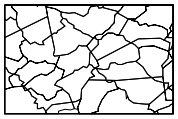
Distance operations

Simple distance computations	Distance computation is one of the fundamental GIS operations. Distances can be computed as straight lines 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 used where we need to identify the closest of a number of features of a given category. For example, we would like to compute for all locations in a district the distance to the nearest hospital. 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

<table>
<tr>
<td>Point or line in polygon operation</td>
<td>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 would like to combine the survey information with data from the census that is available by enumeration area. 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).</td>
<td> 
<table>
<tr><th>Cluster Id</th></tr>
<tr><td>12</td></tr>
<tr><td>13</td></tr>
<tr><td>14</td></tr>
<tr><td>15</td></tr>
<tr><td>16</td></tr>
</table>
+

<table>
<tr><th>EA Id</th><th>Avg household size</th></tr>
<tr><td>507</td><td>4.3</td></tr>
<tr><td>508</td><td>3.8</td></tr>
<tr><td>601</td><td>2.9.</td></tr>
<tr><td>602</td><td>5.2</td></tr>
<tr><td>603</td><td>4.6</td></tr>
</table>
=

<table>
<tr><th>Cluster Id</th><th>Avg household size</th></tr>
<tr><td>12</td><td>4.3</td></tr>
<tr><td>13</td><td>3.8</td></tr>
<tr><td>14</td><td>2.9.</td></tr>
<tr><td>15</td><td>5.2</td></tr>
<tr><td>16</td><td>4.6</td></tr>
</table></td>
</tr>
<tr>
<td>Polygon overlay</td>
<td>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.</td>
<td> 
<table>
<tr><th>Pop Id</th><th>Popdens</th></tr>
<tr><td>103</td><td>23.7</td></tr>
<tr><td>104</td><td>110.5</td></tr>
<tr><td>105</td><td>35.7</td></tr>
<tr><td>201</td><td>96.8</td></tr>
<tr><td>202</td><td>73.4</td></tr>
</table>
+

<table>
<tr><th>LU Id</th><th>Land use</th></tr>
<tr><td>2308</td><td>Forest</td></tr>
<tr><td>2712</td><td>Urban</td></tr>
<tr><td>2487</td><td>Agric.</td></tr>
<tr><td>3102</td><td>Agric</td></tr>
<tr><td>2402</td><td>Urban</td></tr>
</table>
=

<table>
<tr><th>New Id</th><th>Popdens</th><th>Land use</th></tr>
<tr><td>23</td><td>110.5</td><td>Forest</td></tr>
<tr><td>24</td><td>110.5</td><td>Agric.</td></tr>
<tr><td>25</td><td>73.8</td><td>Forest</td></tr>
<tr><td>26</td><td>96.8</td><td>Urban</td></tr>
<tr><td>27</td><td>73.4</td><td>Agric.</td></tr>
</table></td>
</tr>
</table>

Annex II. Coordinate systems and map projections

A. Introduction

The previous review of GIS concepts 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 socio-economic 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. This 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.

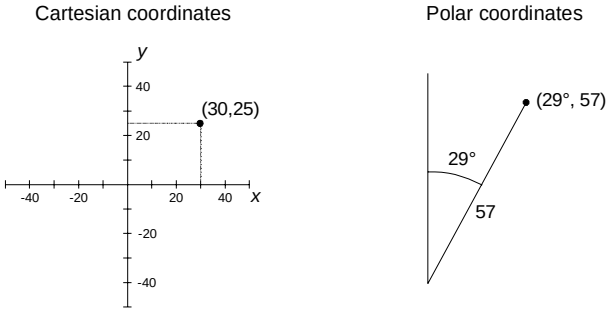
In building a GIS database—for instance, a census GIS—the data developer must ensure that spatial coordinates and boundaries captured from hard-copy data sources, digital gazetteers or during fieldwork are registered in a proper coordinate system in a process referred to as *georeferencing*. This will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

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. The present annex provides a brief review of important cartographic concepts. Cartography textbooks such as Robinson and others (1995), Kraak and Ormeling (1997) and Dent (1999) provide much additional information. More specialized treatments on the topic can be found in Canters and Decleir (1989), Snyder (1993) and Bugayevskiy and Snyder (1992).

B. Coordinates

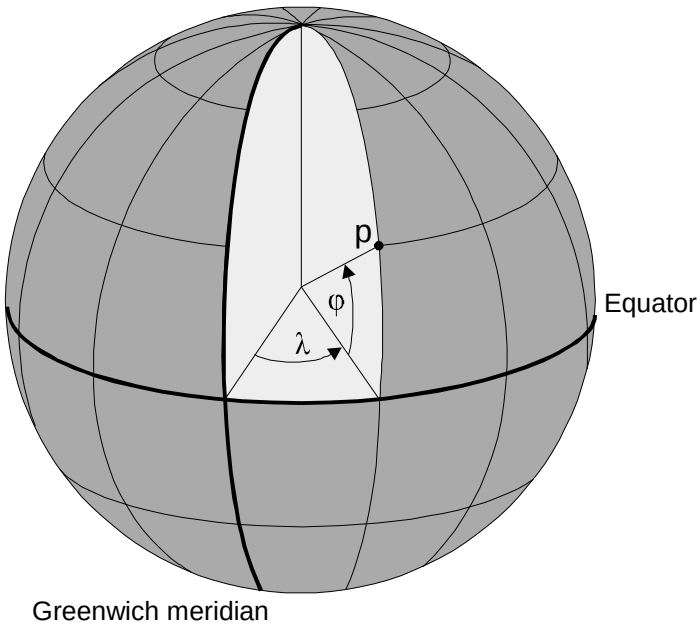
In cartography, the method by which positions of objects on the earth's surface are measured is called the geographic coordinate system. In two-dimensional geometry, the most common coordinate system is the so-called *Cartesian coordinate system*, named after the French scientist René Decartes (1596-1650). 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 A.II.1).

Figure A.II.1. Planar and polar coordinate systems



A flat map, whether on paper or on 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 A.II.2).

Figure A.II.2. Coordinates on the sphere: the latitude/longitude reference system

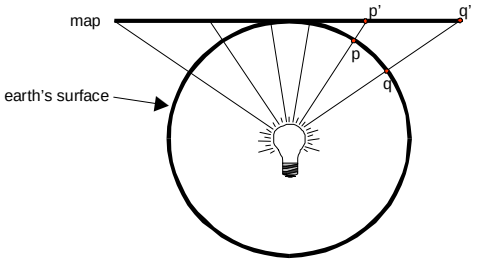


To produce paper maps of the world or some part thereof, these spherical latitude and longitude coordinates need to be translated in some way into a planar coordinate system. A recent book on map projections calls this process of producing a two-dimensional representation of a part of the three-dimensional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

Map projections

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 A.II.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 can not 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.

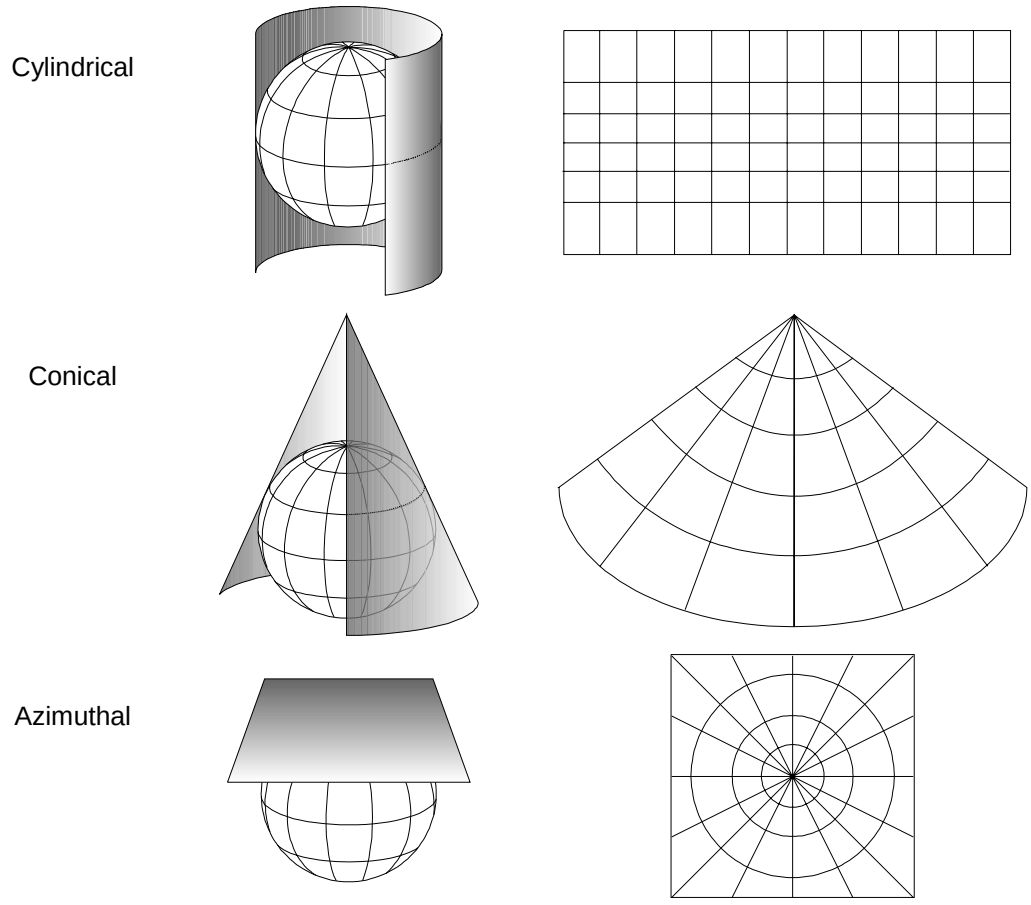
Figure A.II.3. Illustration of the map projection process (azimuthal projection)



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 A.II.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.

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.

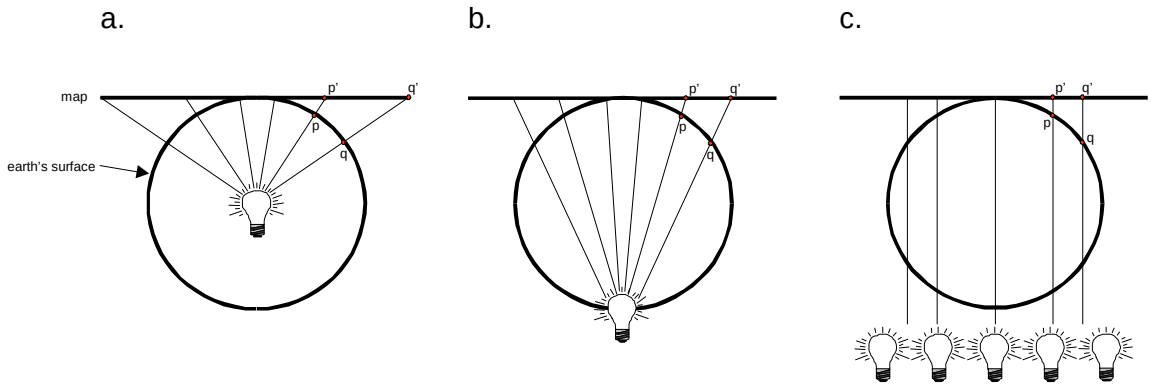
Figure A.II.4. Map projection families



The hypothetical light source is not always located at the centre of the globe (see figure A.II.5a), but could be located at the far pole (see figure A.II.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 A.II.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 A.II.5. Different ways of constructing the projection



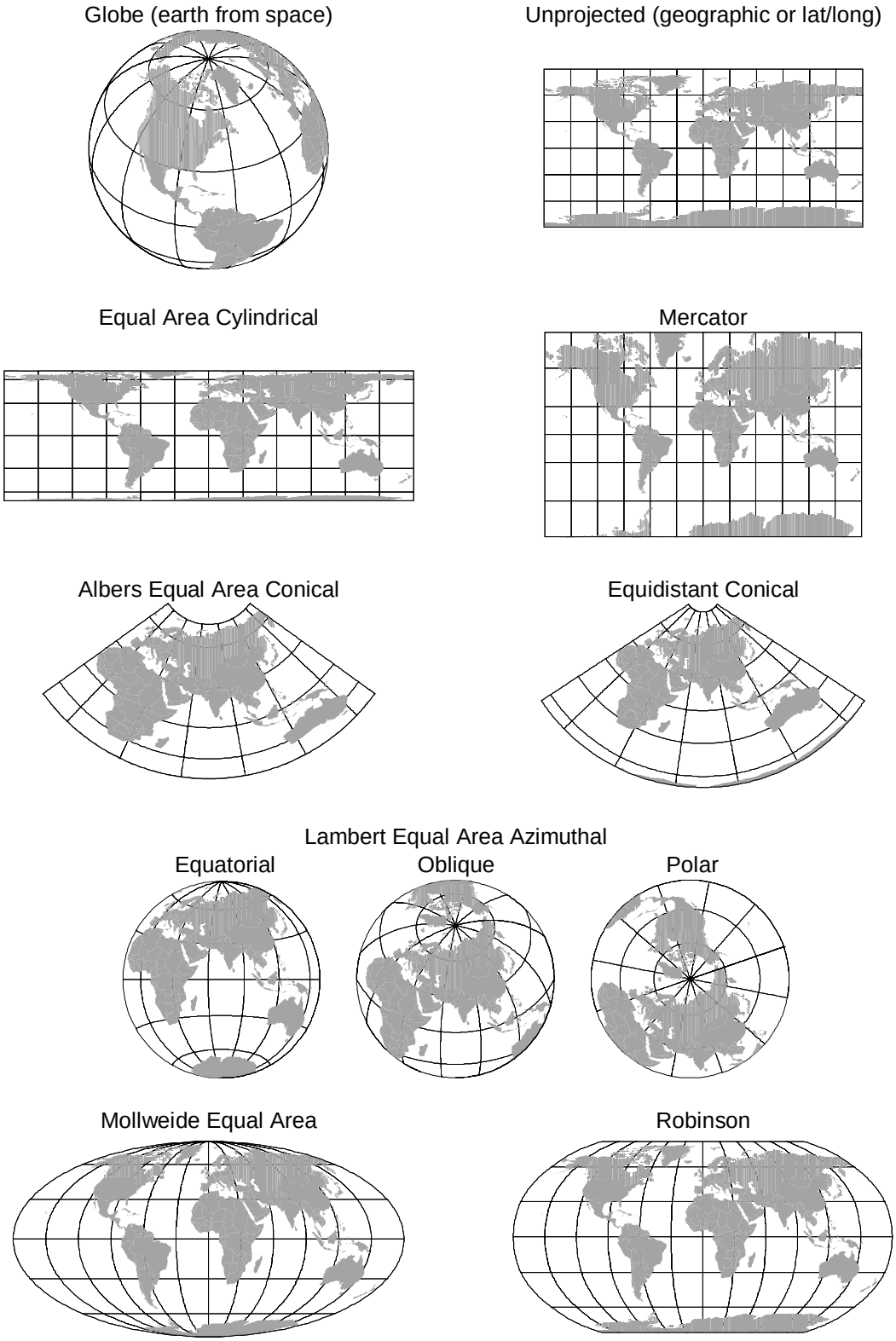
Map projection properties

Although, the imaginary light source is a good way of showing the principle of map projections, these are, in practice, 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.

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 Arabic 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 Arabic peninsula. Maps that show the relative area of all features correctly are called equal area projections. An example is the 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. An example is the 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. An example is the Mercator projection.

Figure A.II.6. Common map projections



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).

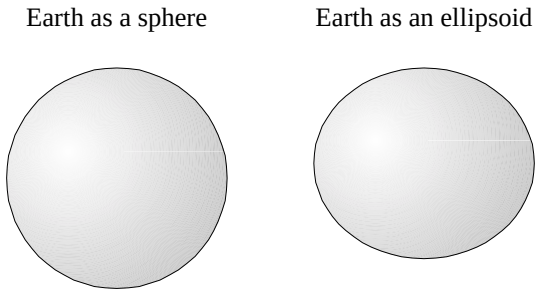
Cartography textbooks and many GIS manuals have comprehensive lists that show which applications are best served by which map projection. In some instances, the best choice may be a projection that does not preserve any property perfectly. The Robinson projection that is popular for global maps, for 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.

Figure A.II.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 that 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.

More precise mapping: geographic datums

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 A.II.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).

Figure A.II.7. Sphere versus ellipsoid



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 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 past 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.

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 A.II.1 shows the coordinates of the United Nations Headquarters Building in the Universal Transverse Mercator (UTM) coordinate system, which is discussed in more detail below. The latitude and longitude coordinates of the United Nations building 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 A.II.1. Projected coordinates of the United Nations Secretariat Building in New York using different reference ellipsoids

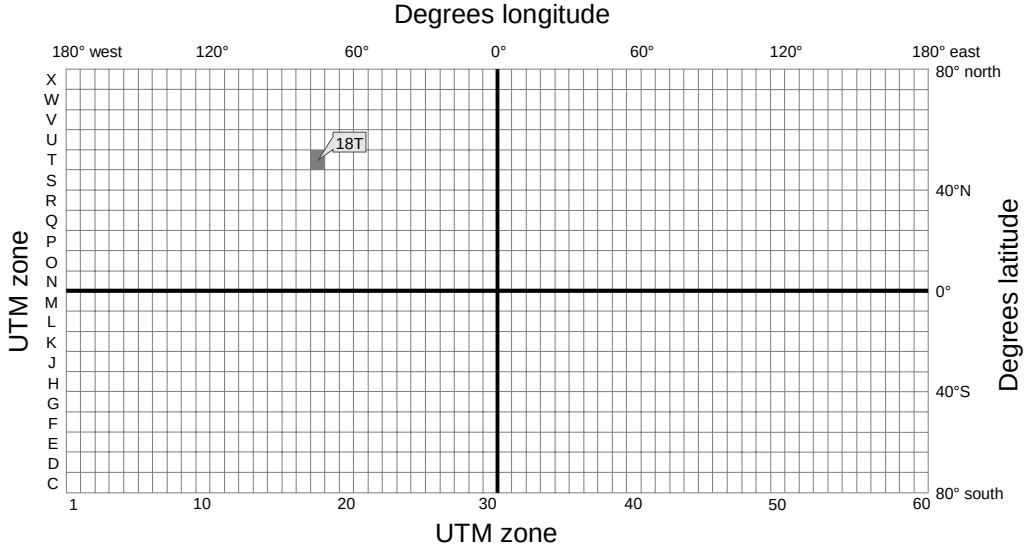
Reference ellipsoid	UTM coordinates (metres) easting (x)	northing (y)
Clarke, 1866	587141.3	4511337.1
Clarke, 1880	587142.6	4511245.1
WGS84	587139.0	4511549.7
Bessel	587128.5	4511095.4
Sphere	586917.2	4529920.6

Universal Transverse Mercator reference system

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 60 zones of longitude (see figure A.II.8).

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°. 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.

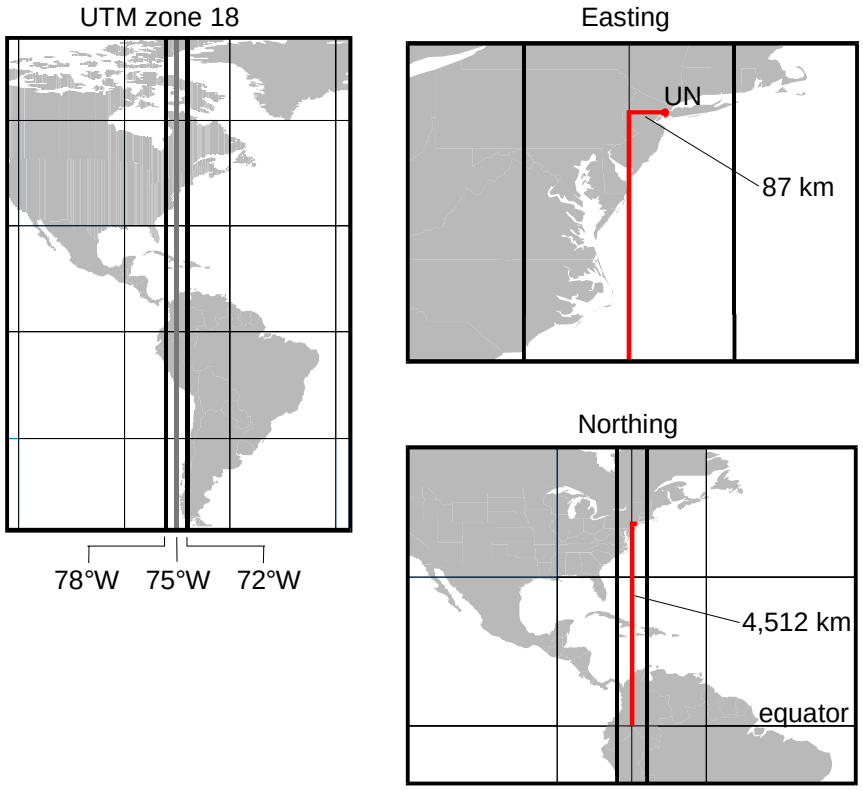
Figure A.II.8. The UTM system



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. To avoid negative numbers, 500,000 is added to the easting. For the same reason, 10 million is added to the northing, but only for coordinates in the southern hemisphere. Such offsets are called "false easting" and "false northing".

To illustrate the use of the UTM system, an example is shown in figure A.II.9. The United Nations Headquarters Building 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 587,139.0 and 4,511,549.7. This means that the United Nations Building 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.

Figure A.II.9. The location of the United Nations Headquarters Building in the UTM reference system.



E. Cartographic scale

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.

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, small-scale and large-scale are often confused, because in colloquial use, "large" and "small" refer 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".

Some common map scales:

Map scale	1 cm on the map represents	
1:5,000	*50 metres*	*larger scale*
1:25,000	*250 metres*	
1:50,000	*500 metres*	
1:100,000	*1 km*	
1:500,000	*5 km*	
1:1 million	*10 km*	*smaller scale*

Digital geographic data are essentially scaleless. 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.

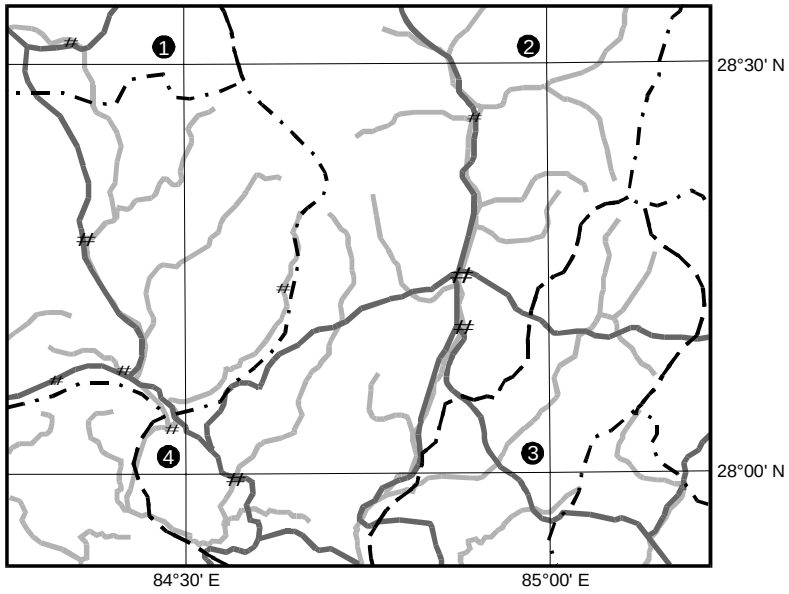
Nevertheless, it is important to keep in mind that the data were likely derived from source material (maps, images, etc.) at a given source scale. Printed maps at different scales, for instance, will show varying degrees of detail. Individual buildings that make up a village will be shown on a 1:25,000 scale map. On a 1:500,000 scale map, the entire village will be displayed as a point, if it is shown at all.

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

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 in the section on digital map integration. To illustrate the process of georeferencing, the following paragraphs will describe a realistic example. Figure A.II.10 below 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 in the previous sections of the present annex.

Figure A.II.10. Control points on a map sheet



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 this so-called 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, 27.0. Note that because GIS programs use planar coordinates, we have to specify longitude/latitude (i.e., x/y) pairs rather than latitude/longitude. 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.

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° .

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.

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.

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 A.II.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).

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 and x and y are the digitizer coordinates of the control points; 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.

Table A.II.2 shows for each control point the coordinate pair in the input coordinate system (digitizer units) and in 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). We see that the transformation error is 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 owing 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 of control point identifiers. Overall, the process should be done with much care, since it will greatly impact the accuracy and thus the usefulness of the resulting GIS database.

The table also provides an indication of the overall error in the transformation. This is the root mean squared (RMS) 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 on 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 A.II.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*	*48 936.2*	*55 529.6*	*-14.59*	*7.80*
2	*26.670*	*20.661*	*97 871.5*	*55 835.2*	*14.60*	*-7.81*
3	*27.696*	*3.824*	*98 333.0*	*409.3*	*-14.55*	*7.78*
4	*12.751*	*2.810*	*49 166.9*	*102.3*	*14.54*	*-7.77*
RMS error (input, output)			*0.005034, 16.524*			

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.

Practical considerations

Any large digital mapping project (e.g., census mapping) requires that map information from many different sources needs to be integrated. For that reason, a standard projection and coordinate system needs to be chosen. Ideally, the reference system that is chosen should be the same as that 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.

Almost all GIS packages provide functions for transforming coordinates from one reference system to another (e.g., from metres to feet or from digitizing units to map units), for converting digital maps from latitude/longitude to 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, which are discussed in detail in chapter II, 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.

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 GIS 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 are properly documented and that the metadata—information about the data—are kept with the digital map data set.

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. The location of the United Nations Headquarters Building 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 the United Nations Headquarters as:

$$40 + \frac{\left(45 + \frac{1}{60}\right)}{60} = 40.7502778$$

$$73 + \frac{\left(58 + \frac{4}{60}\right)}{60} = 73.9677778 \quad .$$

Since the longitude of the 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.

To convert, for example, the latitude back to degrees, minutes, seconds:

Degrees: 40
*Minutes: 0.7502778 * 60 = 45.016668 = 45*
*Seconds: 0.016668 * 60 = 1* .

Annex III. Data modelling

The present annex contains a discussion of 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 presented in annex IV.

A. Definition of key terms

A *spatial data model* is the description of the 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 of any specific software package. The user can therefore implement the data model in any comprehensive GIS package.

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—that is, 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.

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), for example, was initially developed for the AutoCad software package. It is now supported by virtually all commercial GIS software packages.

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.

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.

B. Example template

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 A.III1. 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 an administrative unit (district, province, etc.), or the surface type of a road. These pre determined 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 might 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 that this 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 areas (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 is represented 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 modelled, a graphic illustrates the relations of that entity with various other entities.

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 GIS

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.

To clarify the contents of a data dictionary, table A.III.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.

Table A.III.1. 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 attributes table through the administrative unit identifier.

Authorized combination of attribute values:

Province, Not applicable
District, Not applicable
Locality, Urban
Locality, Rural

Note: Only these combinations are possible. For example, there are no urban provinces or rural districts.

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 that the maximum is indeterminable (indicated by N, meaning any number).

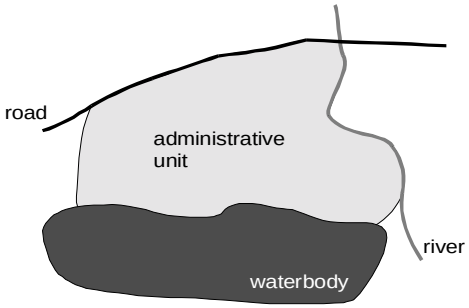
Geometric representation and minimum size

The administrative unit is represented as a point feature if its surface area is less than or equal to 1 square kilometre and as an area feature if it is larger than 1 square kilometre.

Notes

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

The following is a sample database dictionary for distribution of a census GIS database of localities for the hypothetical country of Poplandia. This example 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 GIS data layer of localities for the entire country. The GIS database is distributed in ArcView 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 ArcView shape file version. The geographic attributes 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, 1995.
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 ArcView shape files or MapInfo Interchange format files. Minimum systems configuration depends on the software used to access the data. Generally, a 486 MHz 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
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. The National Statistical Office should be contacted for more information.
References	National Statistical Office (1995). Technical report on the census mapping activities for the National Population and Housing Census of Poplandia, 1995, Census Branch, Cartography Section. National Statistical Office (1995). Methodological and administrative report for the National Population and Housing Census of Poplandia, 1995, Census Branch. National Statistical Office (1996). Results of the National Population and Housing Census of Poplandia, 1995, 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 Web: www.census.gov.xx

Geographic data files

LOC.SHP – GIS database of locality boundaries		
File name:	LOC.SHP	
File type:	ESRI ArcView 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
AREA	Area of the locality in square kilometres.	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 square kilometres.	Real, 10.3	Positive value	None	-999
AREA_LAND	Area of locality covered by land in square kilometres.	Real, 10.3	Positive value	None	-999
AREA_WATER	Area of locality covered by water bodies in square kilometres.	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 square kilometre (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

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 is 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.

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, in 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

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 has to 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.

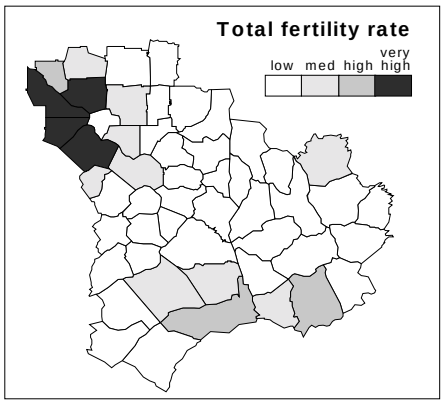
The production of presentation maps is a design process in which the cartographer communicates an idea or concept to the reader (Monmonier, 1993). This is similar to other forms of communication of qualitative or quantitative information in graphical form using charts, pictures or other visuals. The same design principles that guide graphic design therefore also apply to cartography.

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 maximization of the data to ink ratio (Tufte, 1983): 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 with "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 limits, of course. Some map elements, such as the legend itself, a concise title and source information, are clearly required for the understanding of the map.

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.

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 grey shade. These are the hot spots that should catch the attention of the viewer immediately. In figure A.V.1, for instance, the low to high classes are deliberately shaded in light grey 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 is discussed in more detail below.

Figure A.V.1. Establishing visual hierarchy through choice of colours or grey shades



The cartographer can also use other tricks 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.

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.

A final map design comment relates to the responsibility of cartographers 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.

1. Elements of a thematic map

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, for instance, the country's borders, and possibly some reference features such as major rivers or cities. These provide orientation for the reader who wants 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.

In addition to the actual map information, a publication-quality map contains additional elements. These may include:

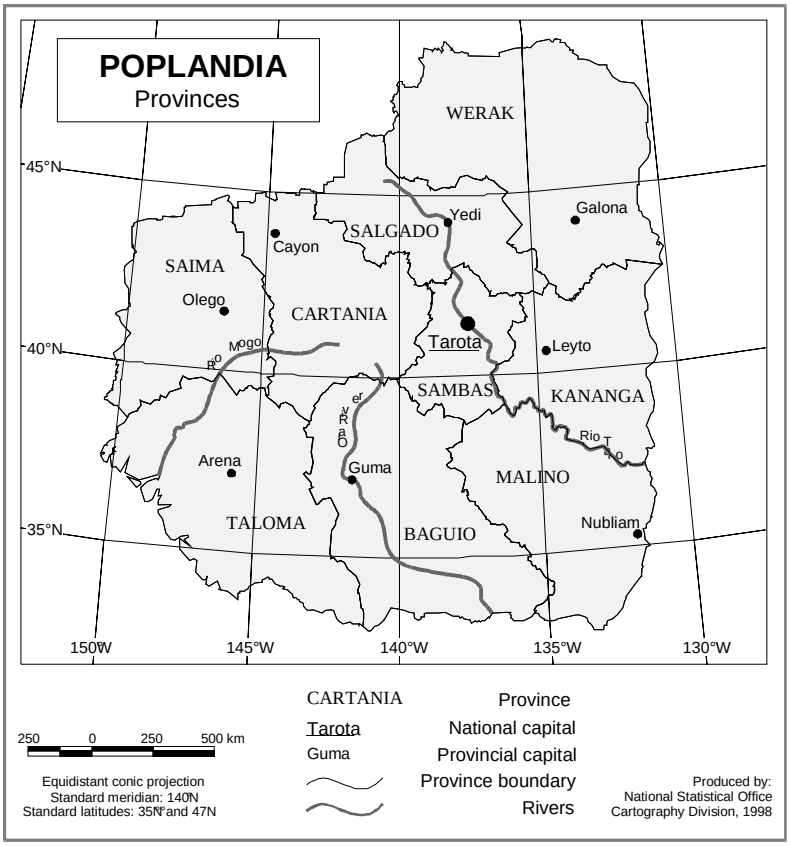
- *Titles and subtitles* should be short and highly descriptive. Titles including phrases such as "Map of …" should be avoided;
- *Data source, credits and production date* give the user information about reliability and credibility of the map. Some agencies that regularly produce maps also add reference 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 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, for example, "persons per square kilometre";
- 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-sized 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, in 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. So, 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;
- *Graticule*, 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 may 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, which 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.

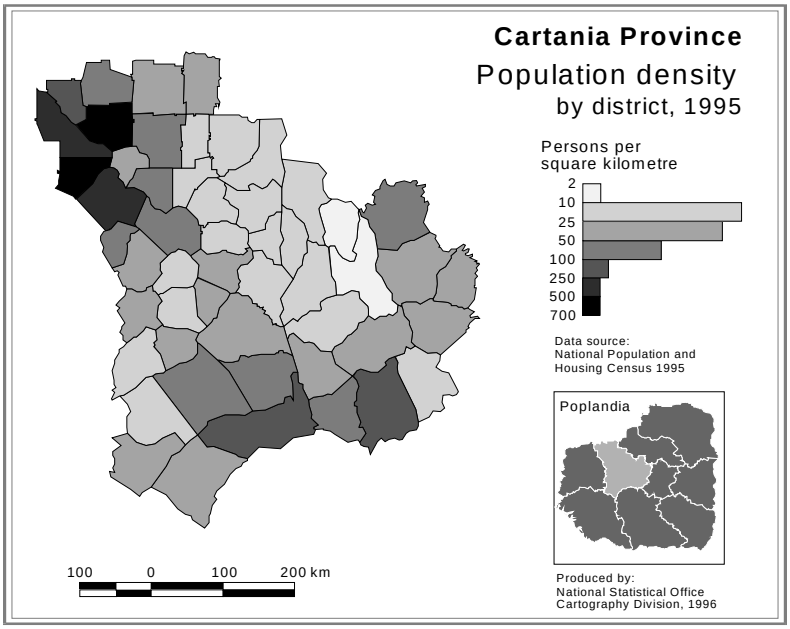
Figures A.V.2 and A.V.3 show two examples of maps that incorporate many of the thematic map elements. Figure A.V.2 is a map of first-level administrative units in the hypothetical country of Poplandia. Draped over the map is a grid of latitudes and longitudes that provides the geographic reference. The 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.

Figure A.V.2. A Sample map of administrative units and major urban centres



The thematic map in figure A.V.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. The design of the map 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 purpose of a traditional legend—relating values to shade 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, this 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 by the readers.

Figure A.V.3. Example of a thematic map of population density



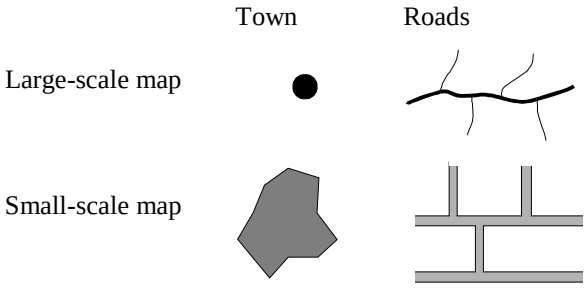
Measurement levels and graphic variables

(a) Spatial dimensions

Thematic maps do not just show the location of a feature, they 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 those elements. This means that in designing a thematic map we have to consider the spatial dimension of the geographic features, and we need to be aware of the level of measurement of the variable. Both will determine cartographic options that are available for producing a map that is visually appealing, easy to interpret and accurate.

Geographic features are represented in a GIS 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 geometric 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 A.V.4). A road might be shown as a line on a province map, but as a double line—that is, an area feature—on a city map.

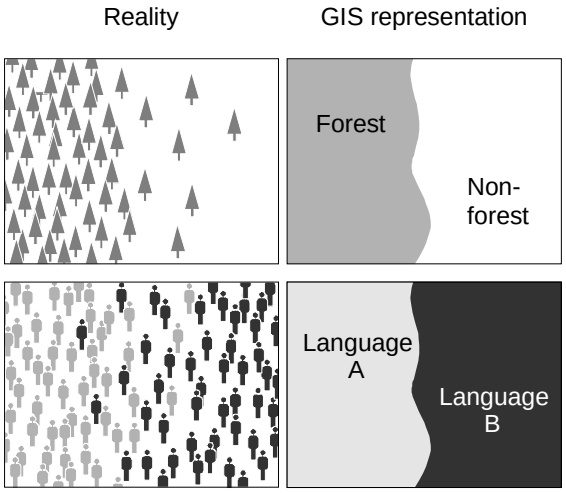
Figure A.V.4. Effect of generalization on the display of spatial features



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 GIS database. Complex real-world features often need to be generalized, simplified or abstracted to represent them 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 A.V.5).

Examples in the socio-economic realm where ill-defined boundaries are encountered are the distributions of ethnic or language groups. Despite the sometimes very distinct distribution patterns of such groups, there are likely 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.

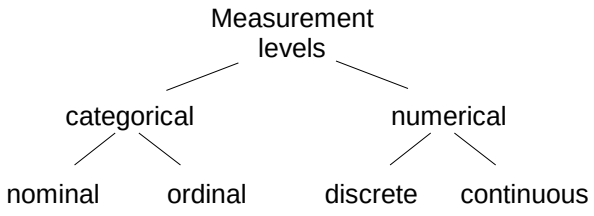
Figure A.V.5. Real-world complexity sometimes needs to be simplified for GIS representation



(b) Measurement levels

Equally important is how the variable we want to map is measured. The main distinction is between categorical and numerical information (see figure A.V.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 is 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 A.V.6. Measurement of variables



If we can quantify the difference between the categories, we have numerical data. Discrete data are counts, for example the number of bedrooms in each household, as well as 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

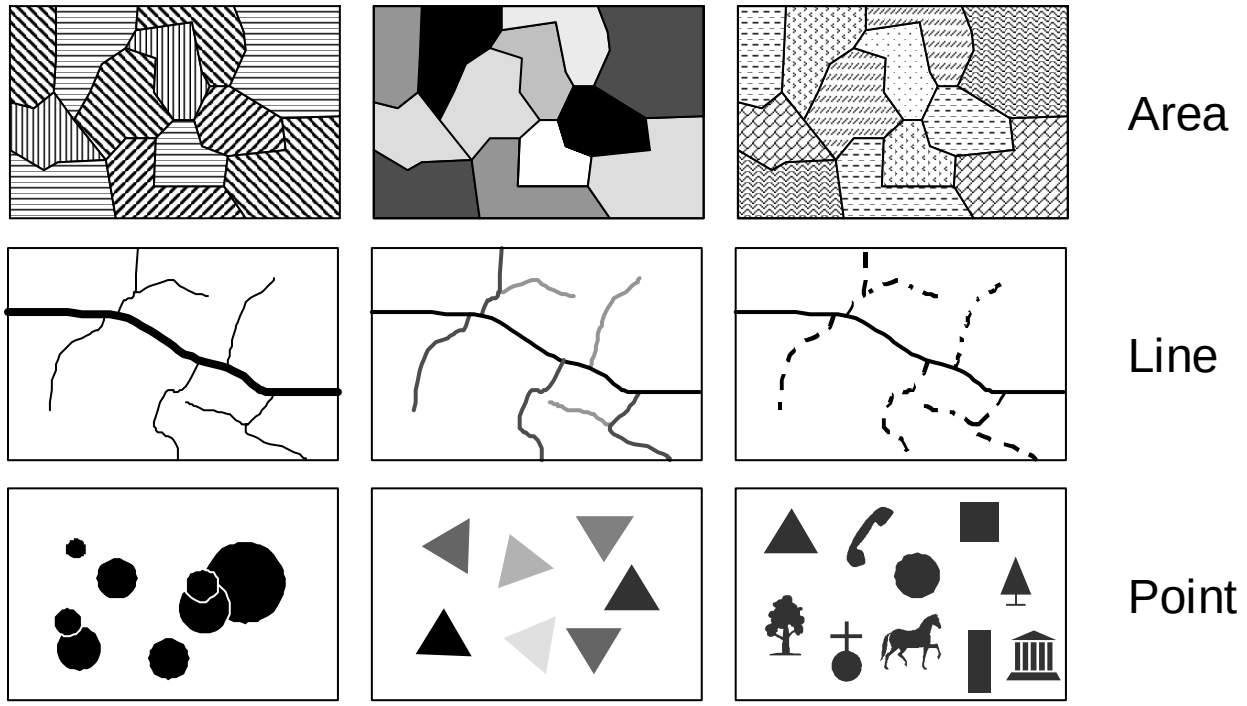
On a thematic map, graphic symbols reveal the differences in values or categories among geographic features to the viewer. The concepts of symbolization that are applied in cartography are similar to those developed for graphic design applications by Bertin (Bertin, 1983; see, also, MacEachren, 1995). Bertin distinguishes between the following graphic variables:

- Size is an indicator of ordinal or numeric differences. Size is most important for point or line features, for instance, 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 is therefore useful to show 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 shortcut if output devices have limited capabilities in printing colours or grey 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 is therefore discussed in more detail below.

In principle, each of these dimensions is applicable to every geographic feature type—that is, 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 A.V.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 to represent numeric values. Shapes of point symbols or texture of area features represent different nominal values.

Figure A.V.7. Graphic variables for polygons, lines and points



3. Types of thematic maps

(a) Mapping discrete features

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 *choros* (place) and *pleth* (value). Choropleth maps show data for discrete reporting units that are often established independently from the actual spatial distribution of the data (e.g., administrative boundaries). The symbol—that is, 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.

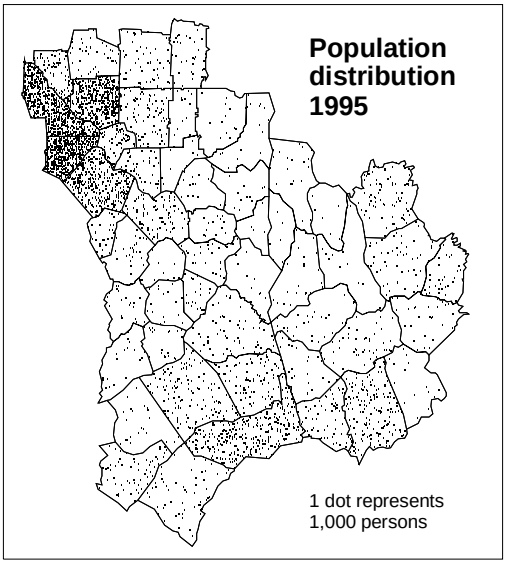
An example of a choropleth map has already been shown in figure A.V.3. Choropleth maps are constructed by first dividing the complete range of data values for the reporting units into a set of categories. Each category is then assigned a colour or shade pattern. Since enumeration data have a natural 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 is therefore discussed in detail in the following section.

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.

The values used for producing choropleth maps are almost always ratios. 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, for example, the crude birth rate as the number of births per 1,000 persons. Most often, when we map socio-economic 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*.

An alternative method for displaying count data is *dot maps*. Dot maps were been 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 A.V.8 presents an example of a dot map showing population distribution.

Figure A.V.8. Dot density map



For the placement of the dots, two approaches are possible. The cartographer could 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 could 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.

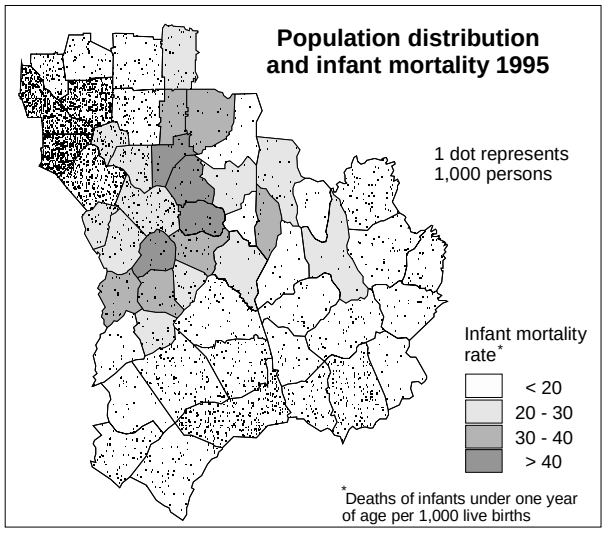
The alternative is to place the dots randomly within each district. In this case, dot density reflects overall value density. GIS and desktop mapping packages that provide dot density mapping functions usually use random placement of dots. The user has control only of the size of each dot and the symbol used for the dots. This symbol could be chosen to reflect the variable mapped, although a simple dot usually provides the clearest display.

Some specialized programs have been written at universities that allow dot placement assisted by some other data layers, but these have so far not been incorporated in commercial software packages. Manual placement of dots that could incorporate the cartographers knowledge of variable distribution is, of course, very tedious.

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 if 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 A.V.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.

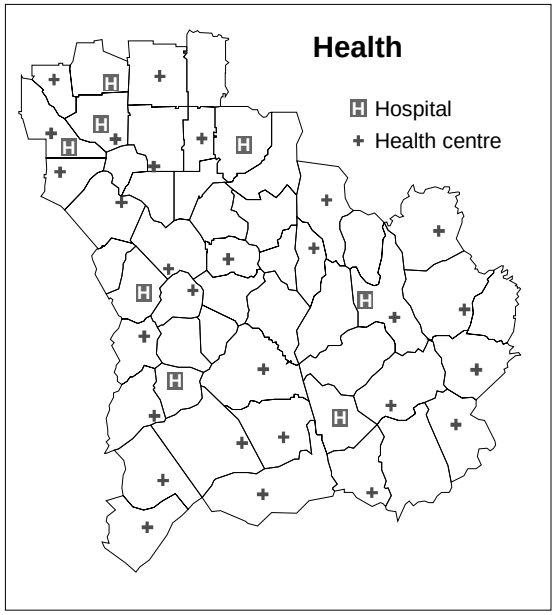
Figure A.V.9. Combination of dot density and choropleth maps



(b) Nominal point data

The simplest case of a dot map is where each dot represents one discrete element such as a farm or a 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 A.V.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 A.V.10 shows the distribution of two types of health facilities with easy-to-interpret symbols. The symbols used are typically text font characters or bitmaps. Most packages have their own font sets, which 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 bitmap symbols.

Figure A.V.10. Mapping discrete point objects

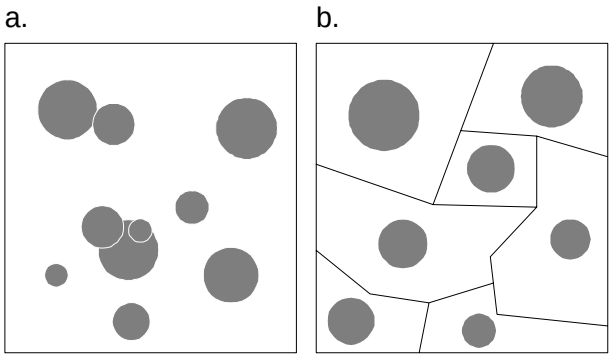


(c) Proportional point symbols

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.

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 A.V.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 A.V.11b.). Note that most systems draw a halo around each circle so that circles that are very close to each other can still be distinguished. The system draws the largest circles first to prevent smaller ones from being covered.

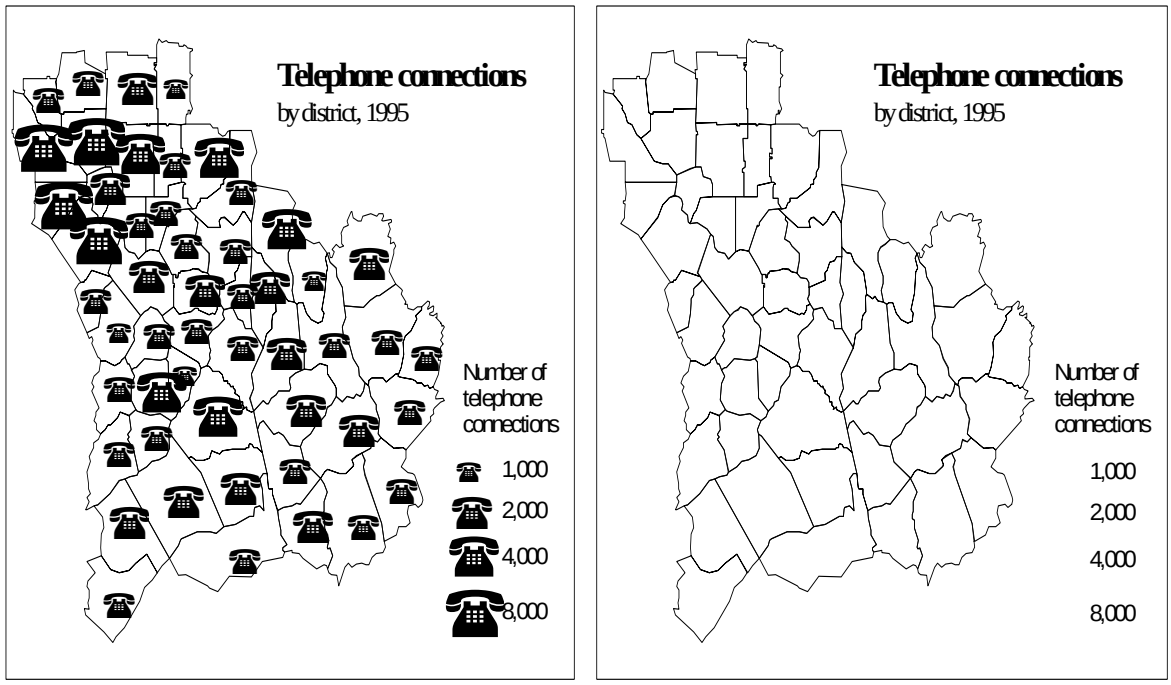
Figure A.V.11. Proportional symbols for point and area features



As before, a computer mapping package will allow us to select 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 A.V.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 in the simpler map on the right. The cartographer has to 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.

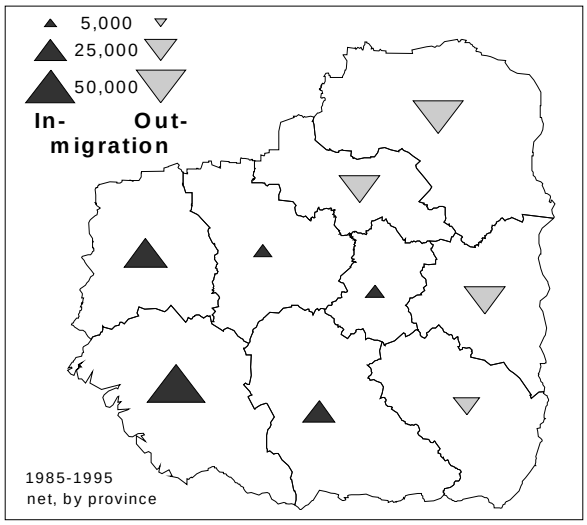
Proportional symbols can also be used to present two variables at the same time. For instance, the size of the circles could represent the number of households in a reporting unit, while the colour or grey shade of each circle indicates the percentage of the households that have a telephone connection. 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.

Figure A.V.12. Pictograms versus simple graphical symbols



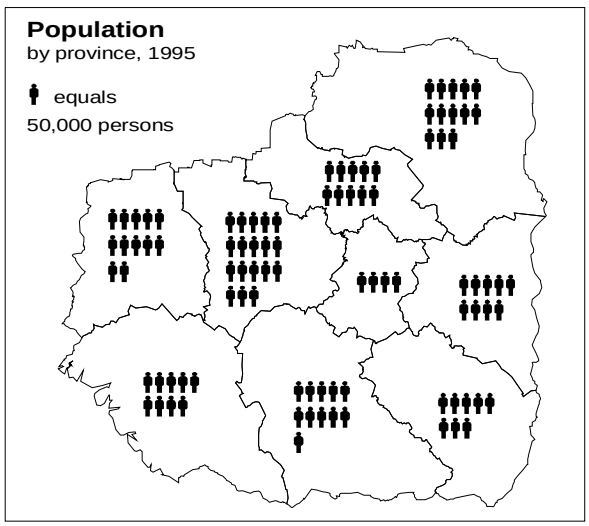
Besides circles, other commonly used geometric symbols include squares and triangles. By varying the orientation of triangles, we can show the magnitude of diverging variables such as in- and out-migration from each reporting unit (see figure A.V.13). Different grey shades or colours ease interpretation further.

Figure A.V.13. Showing magnitude and direction of flows, using simple graphical symbols



Related to graduated symbol maps are maps in which differences in value are represented by the number of a standardized symbol that are drawn for each geographic unit. For instance, total population can be represented as in figure A.V.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 A.V.14. Representation of data values by varying the number of map symbols for each feature



(d) Chart or diagram maps

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 in which it is hard or impossible to extract useful information.

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 A.V.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 A.V.15. Pie chart map

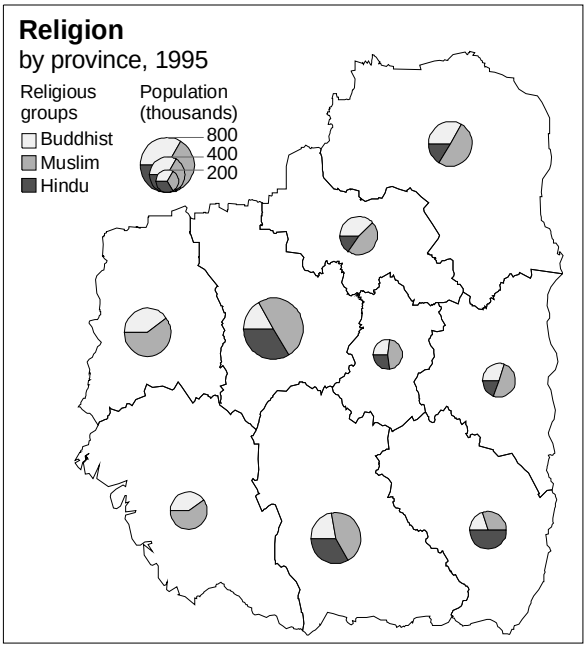


Figure A.V.16. Combination of choropleth and pie chart maps

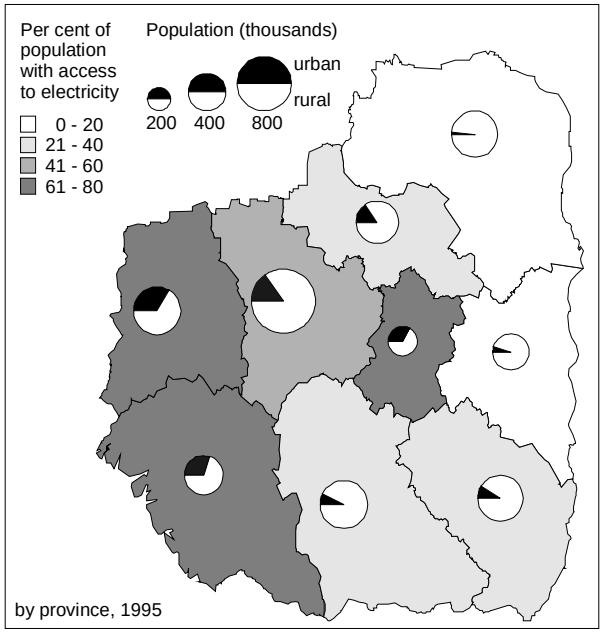
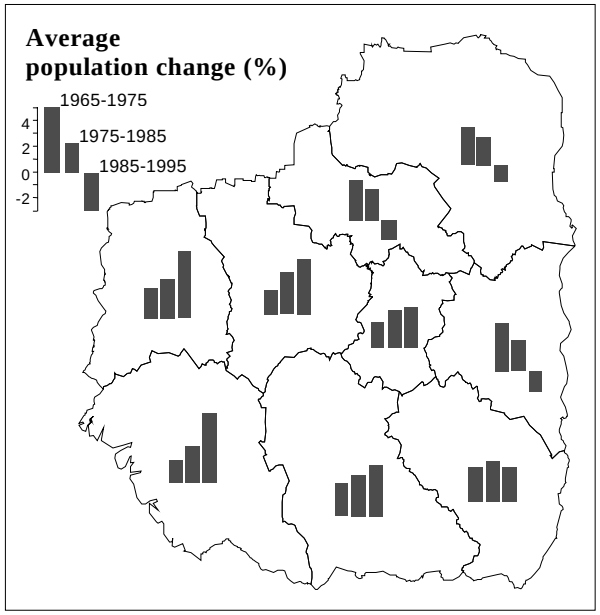


Diagram maps work best if there are relatively few geographic observations and very few groups represented. For example, a pie chart map that has only two categories can be very effective in combination with a simple choropleth map to show several variables at once (see figure A.V.16): the spatial distribution 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 A.V.17. Map showing changes over time, using histograms



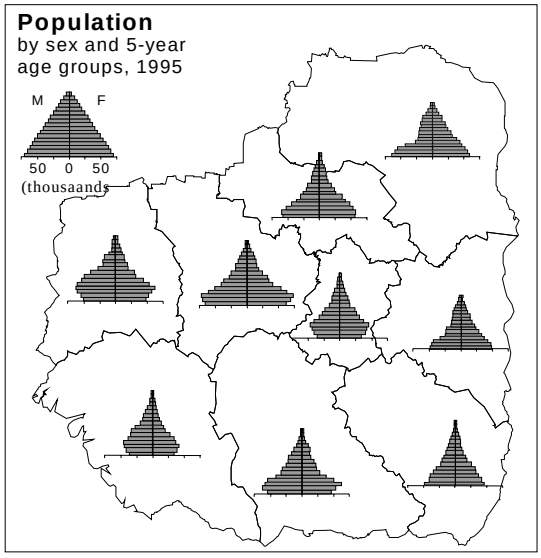
A further use of chart or diagram maps is to show trends over time. The map in Figure A.V.17, for instance, shows the average annual percentage change in the population of each province between the past three censuses. The bar charts are very simple, without a border and without a base line, since for these data it is apparent which bars represent an increase or a decrease of population. As before, what we want to convey is relative changes over time, not the exact values, which are better presented in a table.

One type of chart that is of great relevance to population census data is, of course, the population pyramid. The pyramids 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 A.V.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 that 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.

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

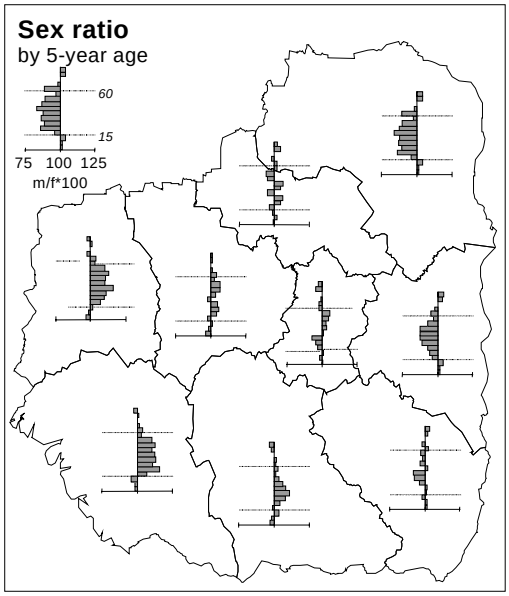
much insight. In figure A.V.18, there is some indication that provinces in the south-east have been experiencing a fertility decline over the past 15 years, while provinces in the north have not. Furthermore, it appears that the provinces in the north-east show a skewed sex ratio distribution. There 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 A.V.18. Combination of maps and population pyramids



Variations in the sex ratio can be highlighted using a different type of bar chart, as shown in figure A.V.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 discussed at the end of the present annex.

Figure A.V.19. Display of sex ratios on a map



(e) Flow maps

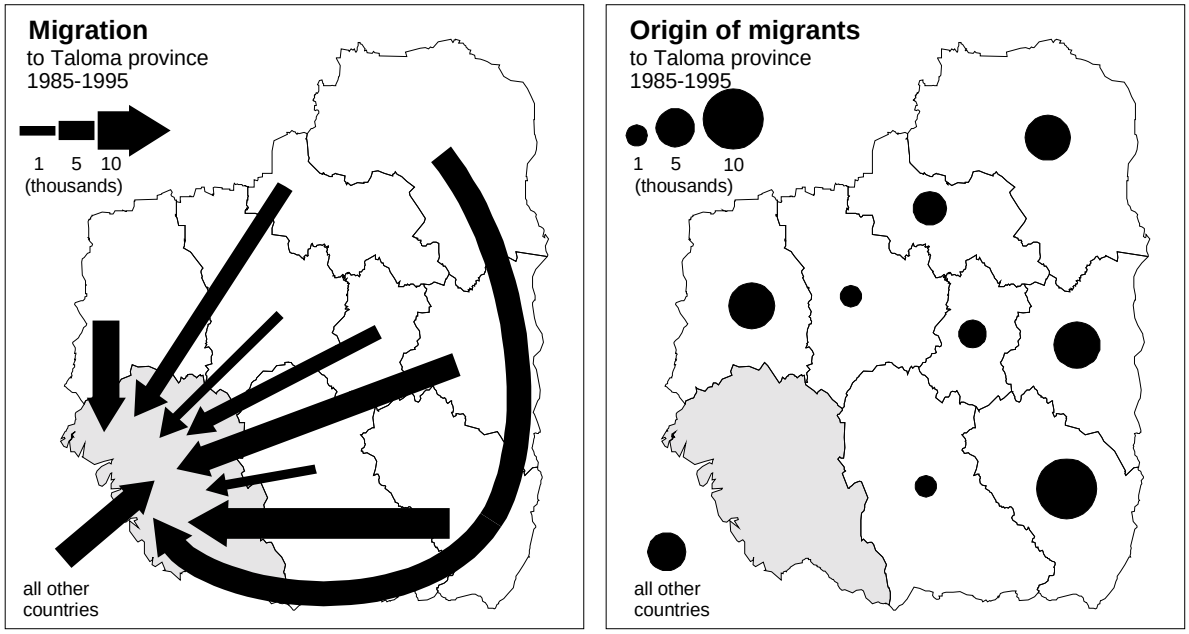
Migration is a demographic variable that represents movement of people from one part of the country to another (internal migration) or between the country and the rest of the world (international migration). Migration can be portrayed on maps in several ways. Migration rates are shown using choropleth 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 A.V.13 above). 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 and the direction (from-to), by using an arrow symbol, and the magnitude of the flow, by varying the thickness of the line.

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 out from the province (see figure A.V.20). For our sample provinces, this would result in a series of nine pairs of maps. 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 regions are far apart.

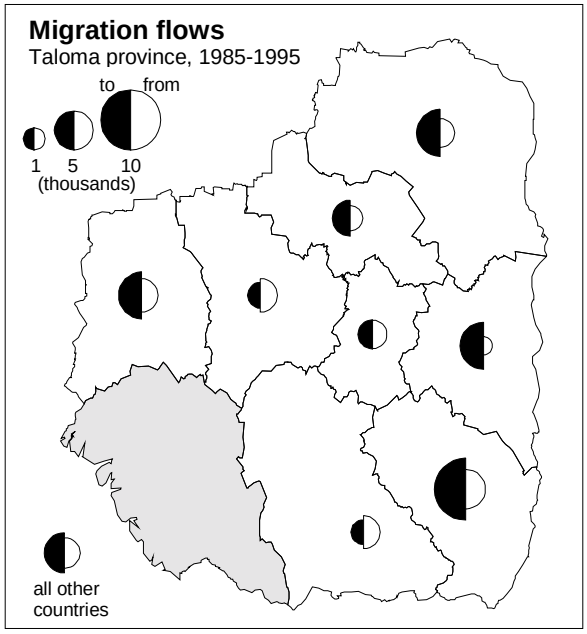
In flow maps that use arrow symbols, the visual impression is guided by the length and thickness of the arrow. A longer, narrow arrow may be visually more dominant than a shorter, thicker arrow owing to its larger surface area. Although, in some instances, a cartographer may want to use this fact to 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 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 A.V.20).

Figure A.V.20. Alternative ways of representing flows between regions



Using special types of graduated symbols, each map can show both migration into and out from the province, as shown in figure A.V.21. Here, half-circles of different colours or grey shades are used to distinguish between in- and out- migration.

Figure A.V.21. Representation of in- and out-migration



(f) Mapping continuous phenomena

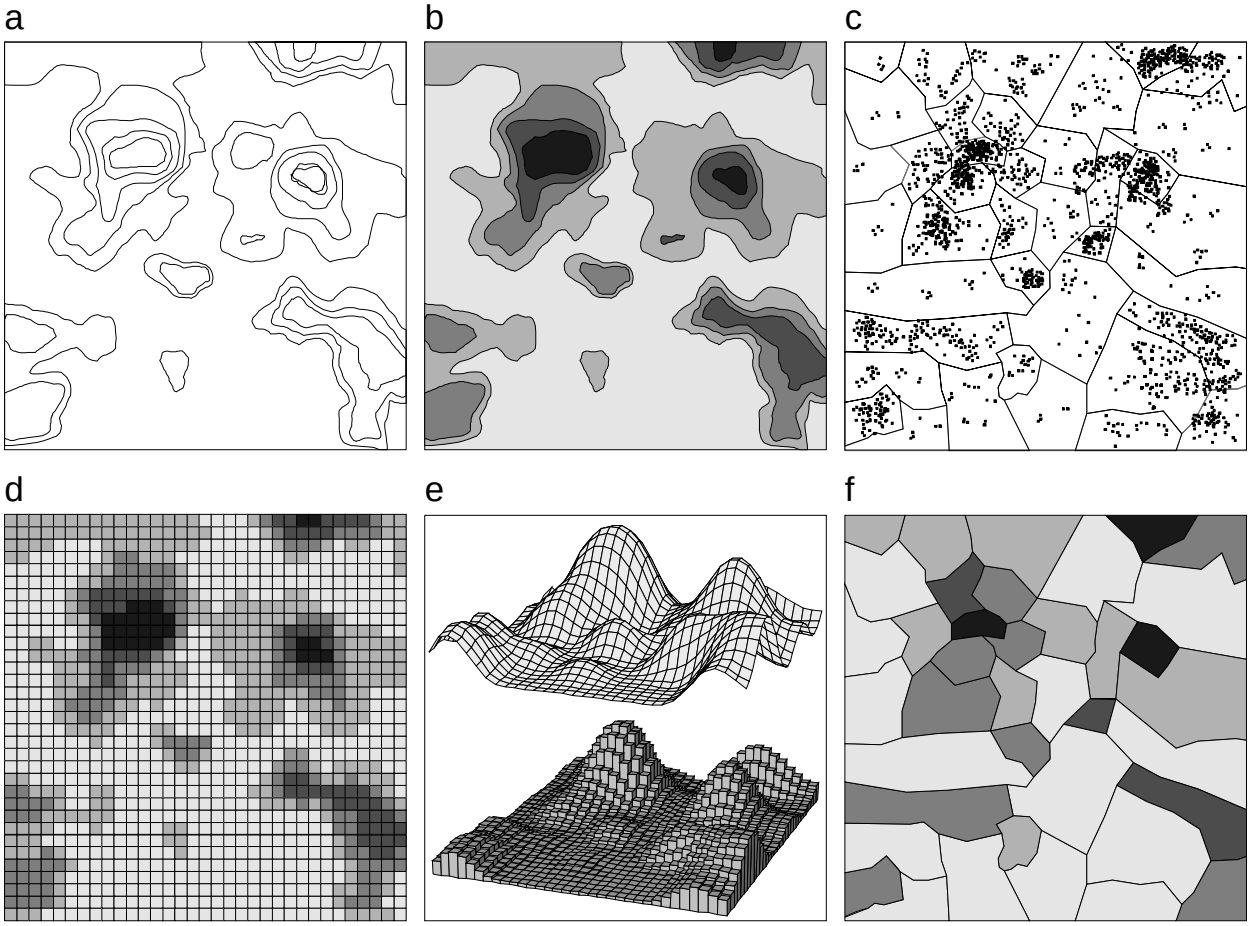
The map types presented in the previous sections are appropriate for data that are referenced for discrete geographic features such as point locations or areas. Some geographic phenomena, however, are continuous. Temperature or elevation, for example, vary smoothly across space. However, it is also possible 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.

True continuity cannot 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 discretize the data in some way for mapping purposes. Several ways are shown in figure A.V.22.

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 (figure A.V.22a). They are used on topographic maps that show elevation. Contour maps can also be shaded, which causes them to look more like choropleth maps (figure A.V.22b). 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 variable. 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 (figure A.V.22c).

For modelling and analysis in a GIS, continuous data are typically stored as regular raster grids (figure A.V.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 A.V.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 socio-economic 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. More standard mapping techniques are therefore generally more appropriate. For comparison, figure A.V.22f shows a choropleth map, where the reporting units are not determined by the data distribution.

Figure A.V.22. Alternative cartographic methods for displaying continuous data



C. Data classification

In the previous sections, the tools that the cartographer has available to display thematic information on maps were discussed. 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, we often 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.

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, which 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. These defaults may or may not be appropriate for the variable that is mapped—most often they are not. Automated classification tools often lead to inappropriate or even misleading map designs. The following paragraphs therefore discuss some of the classification options in more detail.

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—not the 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.

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—that is, one that has many low values and only a few very large values.

For the preparation of publication-quality maps, the data should therefore always be evaluated using statistical graphs. 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.

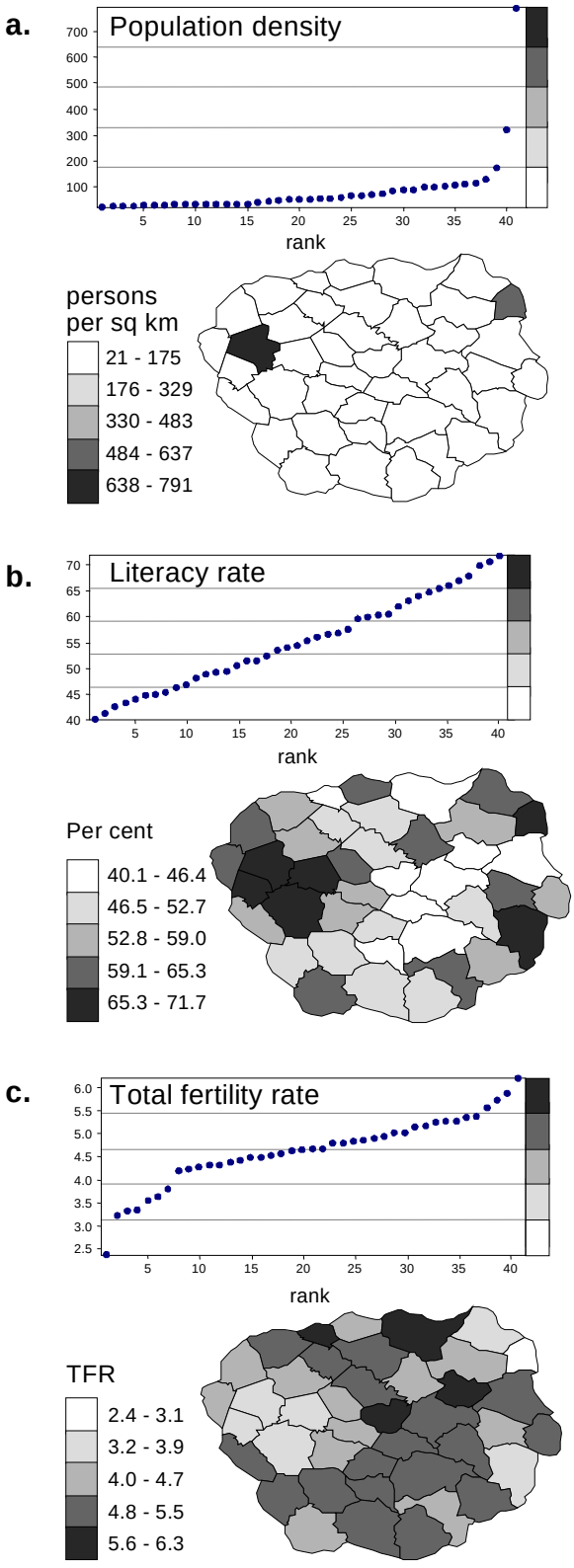
The most useful type of chart for determining class ranges is a rank-order plot. All data points are sorted according to their values from low to high. They are then plotted next to each other—the x axis shows the rank of each observation and the y axis shows the data value. Vertical gaps or *natural 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.

The following pages present examples 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 square kilometres, 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.

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 so-called 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.

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.

Figure A.V.23. Equal intervals



1. Serial data classification

One of the simplest methods of classification is to divide the range of data values into *equal intervals* (figure A.V.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.

For the population density variable, the lowest value is 21 and the highest is 791. The range is therefore 770. Since we want to use five categories, the common difference is thus 770 / 5, which is 154. So, the first class is from 21 to 175, the next is from 176 to 329, and so on.

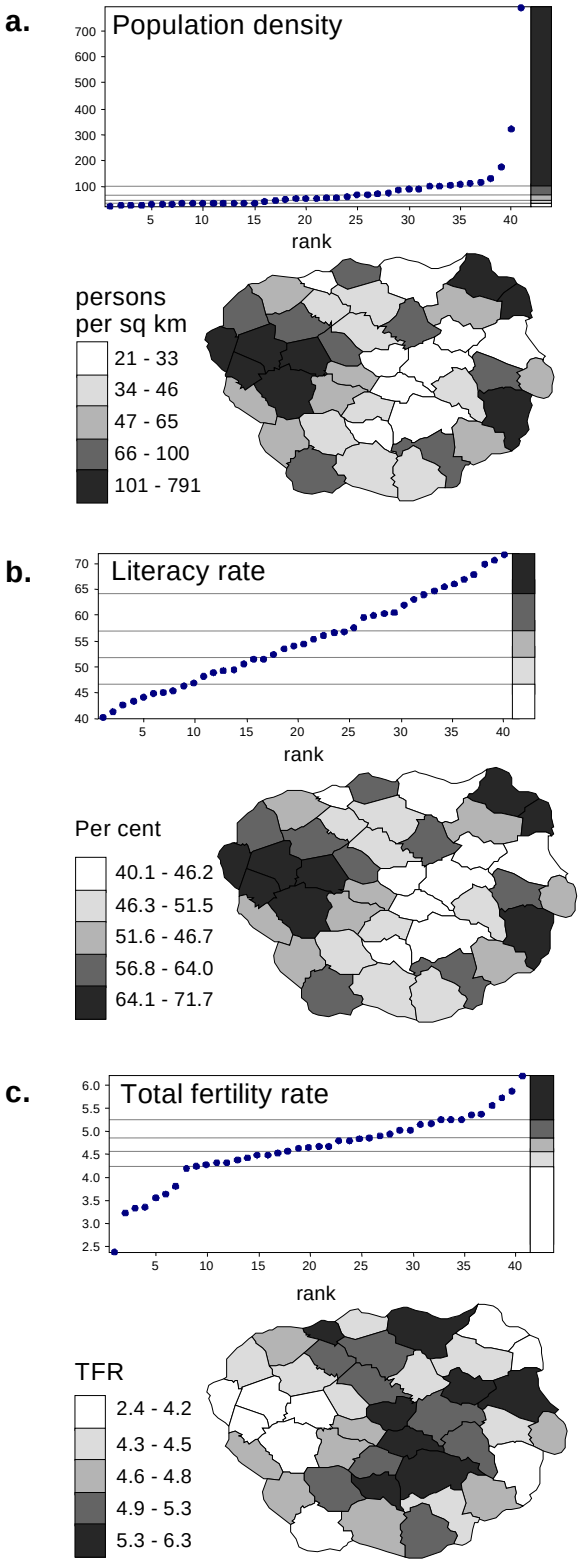
The map for population density illustrates why this can lead to problems. The range of the values is influenced by one very large value. The common difference is therefore so large that the first class range includes all but two of the observations. Clearly, the resulting map is not very informative.

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.

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.

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.

Figure A.V.24. Quantile (equal frequency) mapping



2. Statistical classification

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, which 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.

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.

Quantile mapping is implemented in many desktop mapping packages and this method has therefore become very popular for map production.

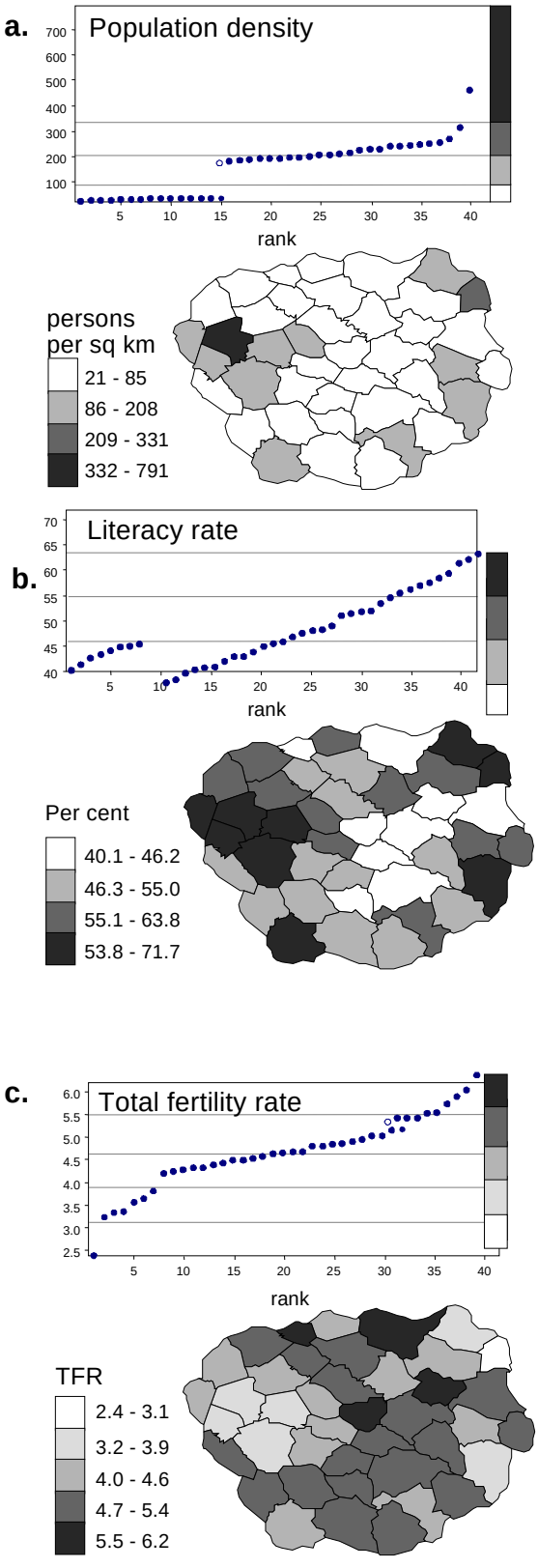
The three sample maps all look good. There is, by definition, a good distribution of observations across classes so that all maps make good use of the full grey scale range.

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.

However, in the population density and TFR maps, we see that the method groups 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).

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 misleading.

Figure A.V.25. Standard deviation



Another statistical classification technique is based on summary measures of the 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.

Map categories based on standard deviations therefore show how individual observations—for example, districts—compare to the average value for the entire province or country.

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.

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 we see in figure A.V.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.

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.

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.

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.

Standard deviations can be used to represent different types of trends in a data set (figure A.V.26; see Dent, 1999). In the examples given in figure A.V.25, a grey 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 A.V.26a. This is, in fact, the least common application of standard deviation classification.

The method is more commonly used to highlight diverging trends. For instance, to show income levels, we might want to highlight the poorest and the richest districts. In that case, we would assign strong colours or texture to the districts with values of more than one and two standard deviations from the mean and relatively muted shades to the ones located in the centre of the data distribution (figure A.V.26b).

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 can be used on both sides. If the interest is also on 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.

In other cases, we may wish to highlight the middle ranges (figure A.V.26c). McEachren (1994), for example, discusses a map of Northern Ireland 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).

Figure A.V.26. Assignment of shades for classes determined by standard deviations

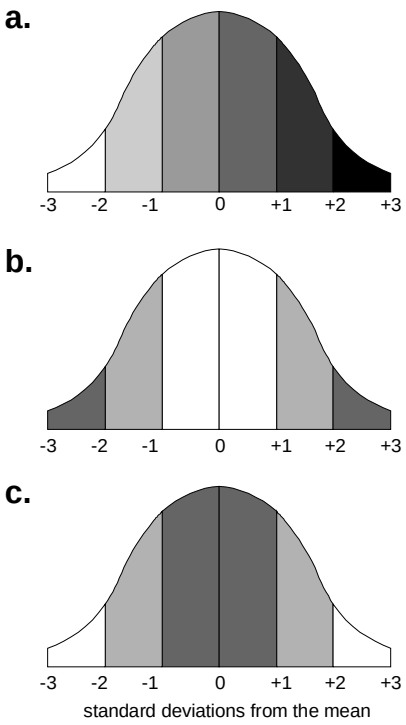
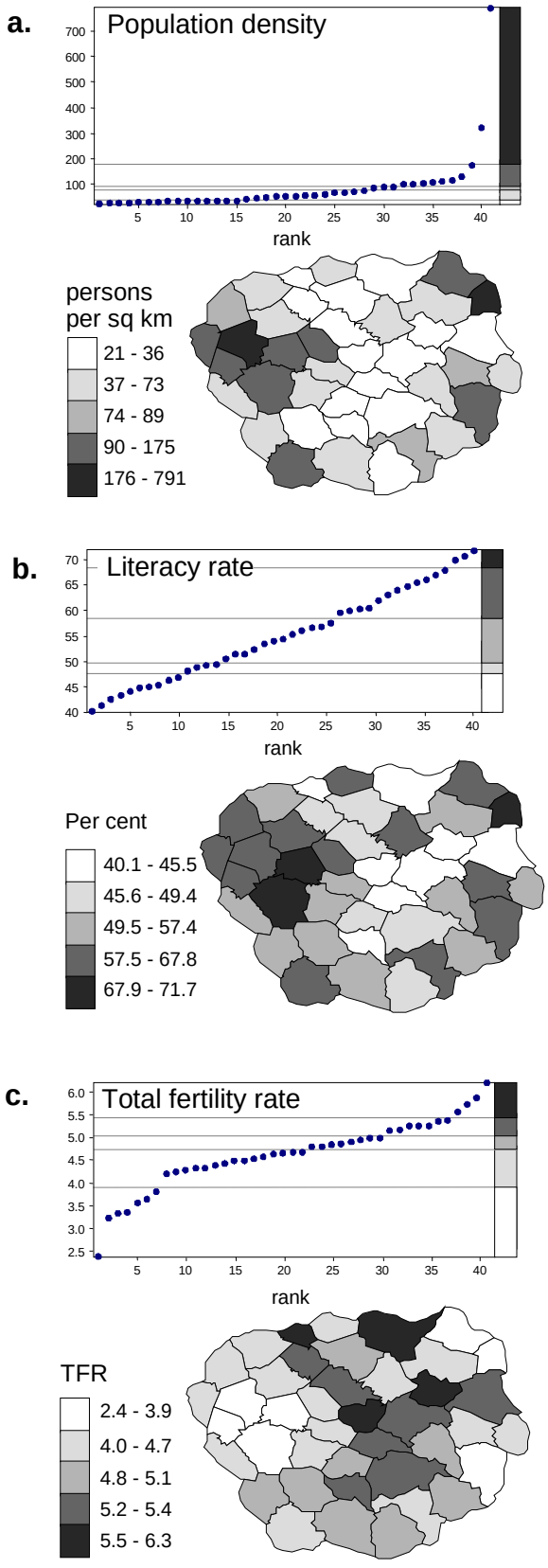


Figure A.V.27. Natural break points



3. Natural breaks

As we have seen in the previous examples, 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 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.

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 A.V.27. For the TFR variable, this is quite straightforward, since it shows several distinct break points in the distribution.

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.

Similarly, for the uniformly distributed literacy variable, the class breaks are not very distinct since the value difference between observations does not vary much.

Despite these difficulties, classification according to natural breaks, which explicitly considers the data distribution, usually results in accurate cartographic representations of the data and good visual contrast.

Rather than relying on somewhat subjective judgement, one can also let the computer determine natural or optimal break points. A few 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

So-called unclassed choropleth maps do not require any choice of classification method by the cartographer. Owing to improved display and print technology, computer screens and printers can produce a large range of different colour tones or grey shades. For an unclassed or *n*-class map, the data values would directly determine, for instance, the per cent grey level. For instance, for a percentage variable, we can select the corresponding grey level on a scale from 0 per cent grey (white) to 100 per cent grey (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.

In practice, this may not yield optimal results, however. One reason is that many variables do not range from 0 to 100, but instead have values concentrated in a smaller range. The map may thus end up with very light or very dark grey shades only. We can avoid this problem by "stretching" the data distribution: using the lightest colour for the lowest value and the darkest for the highest will produce maps that are easier to interpret for the viewer.

In general, however, there is a limit to the number of grey shades or colours that can be distinguished easily. While a continuous shading scheme is useful for analytical purposes, classification of data values into a small number of categories is generally preferable for presentation maps.

5. External data classification

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. To allow for accurate comparisons between maps, for example of the fertility rate in the provinces of a country, the classification needs to be identical.

6. General remarks

The present 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.

Each method has strength and weaknesses, which are highlighted in Table A.V.1. Which method 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. The optimal number of categories and the best break points will then often be quite obvious.

It should be noted that natural breaks are not appropriate, if several maps are presented together for comparison, for example, a time series of sex ratios by district or maps of access to safe drinking water for two provinces in the country. In this case, 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. Two quartile maps could, for example, highlight the 25 per cent of districts with the highest literacy rates as determined in the last and previous census.

Table A.V.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
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 judgement (visual determination) Does not support comparison of maps over time
Unclassed choropleth maps	No category break points need to be defined Grey 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 grey shades or colour tones Maps with subtle grey or colour differences do not reproduce (e.g., photocopy) well Not easily implemented in most GIS and mapping packages

Colour choice

The map examples discussed in the present annex all use grey scales for symbolization. Black and white publications are less expensive to produce and grey 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 continue to drop in price. Also, in the near future, many more maps will be presented electronically on Web sites or electronic publications. Here, colour can be used extensively in map design.

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 and 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, orange and 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 grey while one with a high saturation value will look more pure.

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 grey shades. The lowest levels of red, green and blue combined produce black, the highest values produce white.

Colour choice depends on the measurement level of a variable, the type of map used and the message that the cartographer wants to communicate. Humans are good at differentiating colour hues, which makes it very appropriate 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—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.

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 are, of course, not related in direct proportion to the values of the data categories. For the population density variable example used in the previous section, for instance, we would otherwise use many very light and barely distinguishable shades of light red 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.

If a classification consists of many categories, we may end up with more categories than can be clearly distinguished on a printed page. In this case, colour hues that are adjacent to each other can be combined—a so-called part-spectral colour range. To use the population density example again, we could 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 confusing to the viewer.

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 that 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 light blues to a bright blue for the highest in-migration.

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 where, for example, differences in literacy are indicated by adjacent hues in a partial spectral scheme, and differences in fertility rates are shown 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. In general, multivariate maps should be avoided. Section F contains some alternative approaches to presenting multivariate information geographically.

Returning to the types of measurement levels that were discussed earlier, Table A.V.2 summarizes guidelines concerning the use of grey shades and colours (see, also, Brewer 1994).

Table A.V.2. Choice of grey 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 grey versus dark grey.	Strong contrasting colours of different hues such as blue and red or yellow and green.
	categorical	Dominant language (English, French, 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 grey shades, with relatively strong differences between the grey 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 grey shades are acceptable.	Similar to ordinal data, but smaller differences between grey shades are acceptable.
Continuous	sequential	Literacy rate (any value between 0 and 100 per cent)	Continuous range of grey shades. Level of grey may or may not be proportional to data values. Subtle but distinguishable differences in grey 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 grey) 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.

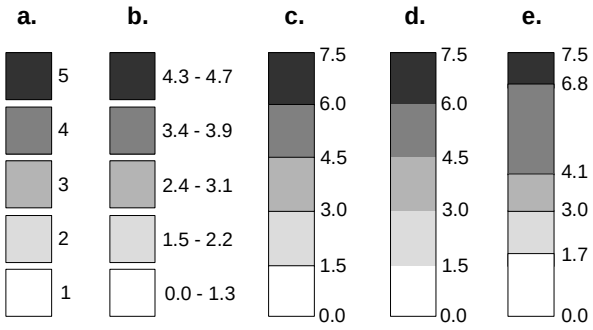
Map legend design

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.

Figure A.V.28 shows some examples. For categorical data, individual legend boxes should be kept separate (legend a). Similarly, class ranges that are not contiguous—for example, there is a gap between the upper limit of one class and the lower limit of the next one—can be emphasized this way (legend b). 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 (legend c). Continuity of data values is emphasized even more if the individual category boxes are not enclosed by an outline (legend d). A legend for a classification of a continuous variable with irregular class ranges, finally, is shown in legend e.

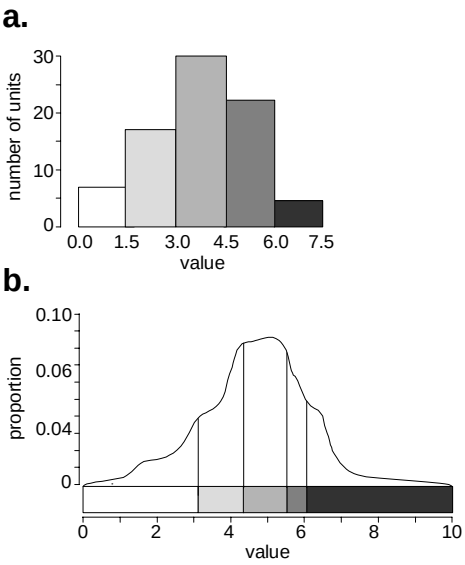
Figure A.V.28. Different types of legends for shaded maps



The last three legends in figure A.V.28 show breakpoints rather than data ranges. When using data ranges for a continuous distribution, we encounter the problem of showing a data value for two categories: for example, 0 – 10, 10 – 20, 20 – 30. This problem can be overcome by using the *less than* symbol to assign each value to only one category: for example, 0 – <10, 10 – < 20, 20 – 30. With open-ended classes, the *greater than or equal* symbol can be used: <10, 10 - < 20, ≥20.

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 A.V. 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 in 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 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 in the accompanying text.

Figure A.V. 29. Legends showing statistical data distribution



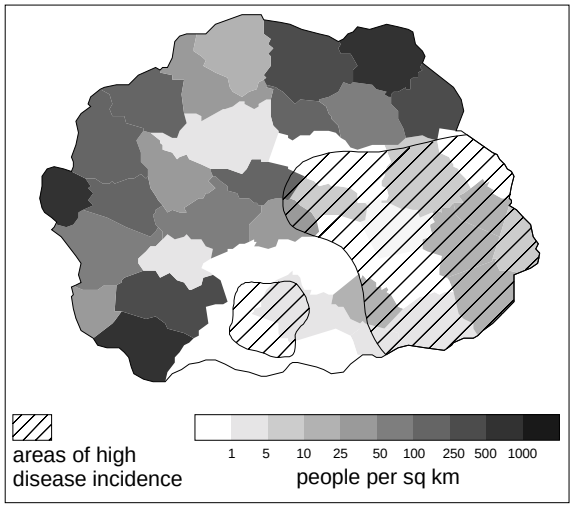
Statistical software also allows the computation of density plots showing the data distribution in a more continuous form compared to a histogram (see figure A.V.29b). The surface under the density curve sums to one so that the approximate frequency of any individual data value can be read from the graph. Legends of this kind have been used, for example, in the Atlas of United States Mortality (NCHS, 1997).

F. Maps that tell stories

1. Multivariate maps

With few exceptions, the previous examples 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, we sometimes would like to display more than one variable at a time. In the section on colour choice, we observed 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 has been mentioned earlier, 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 A.V.30).

Figure A.V.30. Combinations of solid and hatched shade symbols to display two variables on the same map

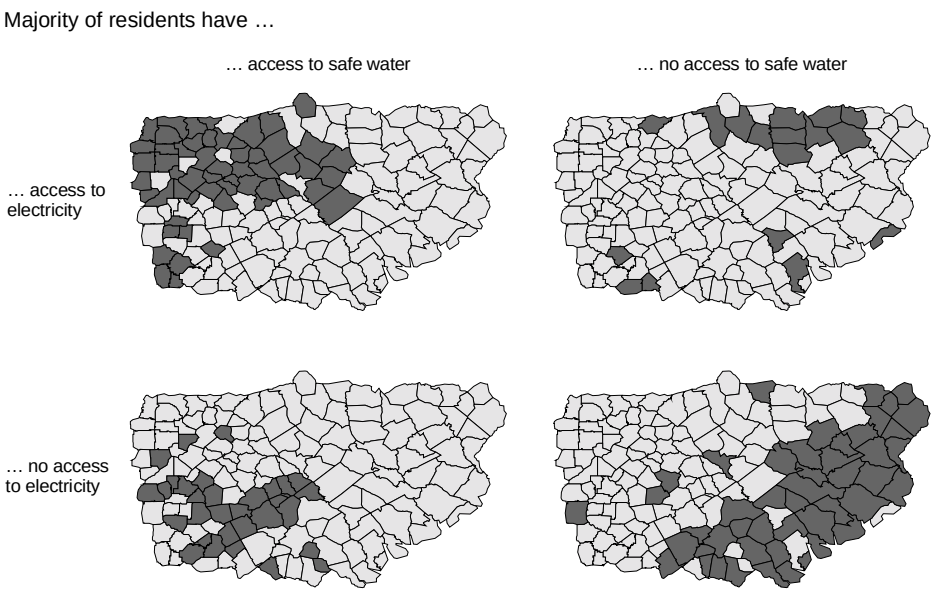


In statistical data analysis, two categorical variables that take on only a small number of values are analysed using cross-tabulations. Such tables are also called contingency tables. The rows and columns of a two-way table show the categories of the two variables and the cells show the number of observations 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

If we want to present this information geographically, we could produce a map with four classes: one for each of the cells in the cross-tabulation. Yet, 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 A.V.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 A.V.31. The map equivalent of a two-way table



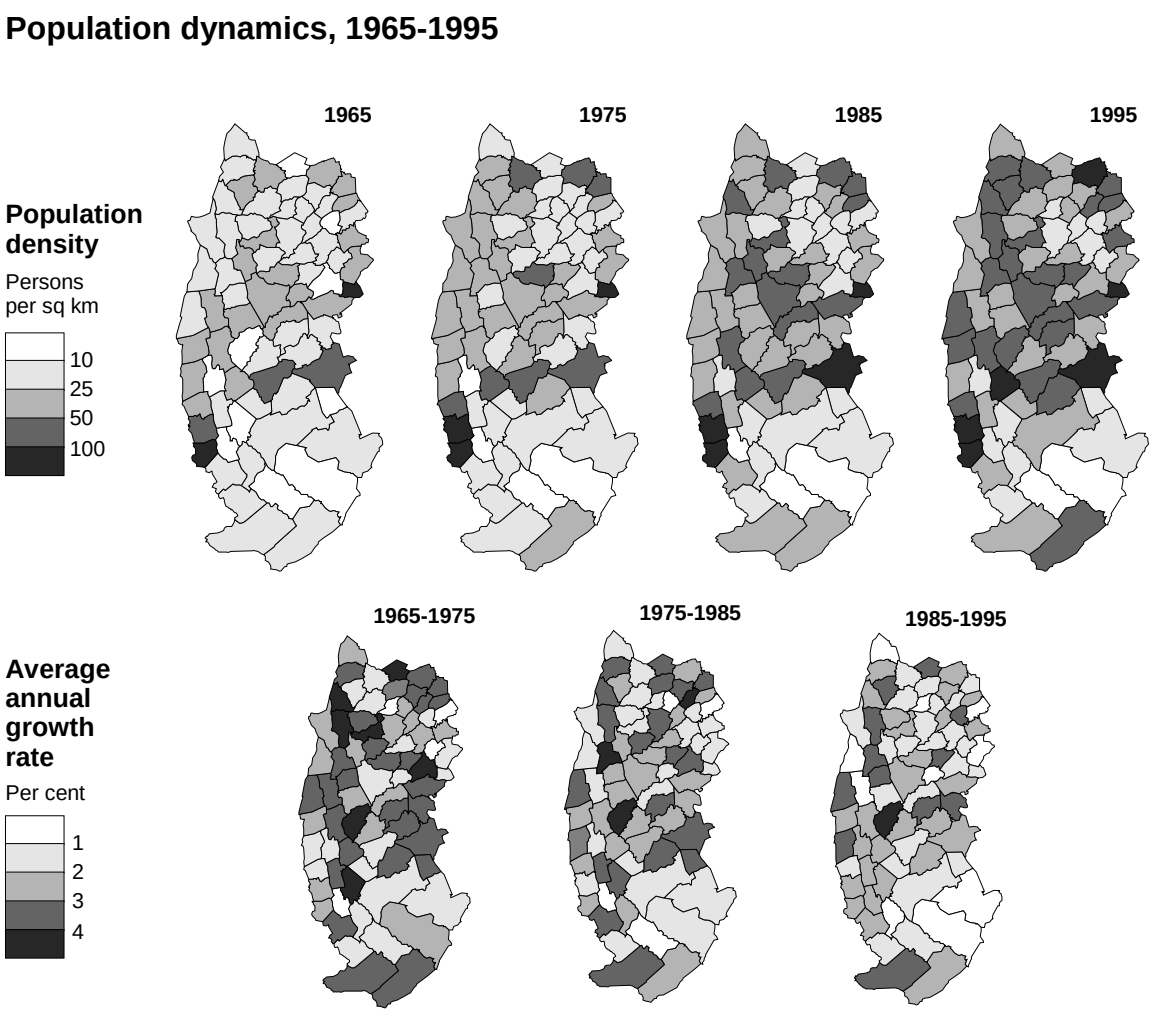
Patterns are immediately apparent even on a small map that covers only about one third of the page. Most districts in the north-west have access both to safe water and electricity, while the majority of households in districts in the south-east have neither. In cross-tabulations, the off-diagonal cells 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.

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 grey shades are required.

2. Small multiples

Arranging data in several maps can also effectively present dynamic information. Figure A.V.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.

Figure A.V.32. Small multiples – depicting changes over time



Displays such as the ones in figure A.V.32 are termed *small multiples* (Bertin, 1983; and 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 possibly complicated legend design.

Another example of a map that uses the concept of small multiples is shown in United Nations (1997a), figure 4.8. It shows sex ratios for five-year age groups for 75 districts in Nepal. The graphic shows 17 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 grey 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 1,275 (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 that can be attributed to migration over the life cycle of men in a number of districts.

Annex VI. Glossary

Accuracy — freedom from error. The degree by 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 its 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. They are usually defined and established by legal action.

Aerial photography — the techniques for taking photographs from an aerial platform, usually a low-flying plane. 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 interchange of alphanumeric data and special characters between computers and across operating systems. Each character is assigned a one byte code, that is, 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.

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 include land cover zones or enumeration areas.

Attribute — a characteristic of a geographic feature. For example, a numeric or text field that is stored in a relational database table that 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 square kilometres, 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, and area. General attributes are typically stored in separate data tables that can be linked to the geographic attributes 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.

Arc second — one second of latitude or longitude, or 1/3,600th of a degree.

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.

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 that 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 GIS 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 mapping), or they may represent concepts and more abstract characteristics (thematic mapping).

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 the cyan, magenta and yellow (CMY) model.

Color 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 a 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 — if two or more geographical features are neighbours or adjacent.

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 or 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) — a 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 of paper maps or air photos.

Data conversion — the transfer of data from one format to 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 to another.

Data dictionary — a data catalog 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 the database entities and their relations to one another.

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 that 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 that 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 sometimes 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 an ortho-image, 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. (Sausalito, California). Initially developed for CAD applications, it has also become a standard for GIS data exchange.

Edge-match — a manual or automated editing technique in a GIS that matches shared features that were digitized from adjacent map sheets. Edge-matching 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 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 — 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 that 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.

Gazetteer — a list of place names and their geographic location (usually latitude/longitude).

Generalization — see cartographic generalization.

Geocoding — (a) a GIS function that determines a point location based on an address. See, also, address matching; (b) the process of assigning geographic codes to features in a digital database.

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 attributes 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 geography.

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 — when two or more geographic features 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 Orbiting 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 GPS but is not subject to selective availability. Some receivers combine GPS and GLONASS signals to improve coordinate accuracy.

Global Positioning System (GPS) — a system of 24 satellites orbiting the earth that broadcast signals that 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 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, and canals.

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 news groups. 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. Sometimes 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 or end point of a line feature, or the point at which two or more lines connect.

Normalization — 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 classes quintiles, and 10 classes 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., a 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 a 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, for example, that one 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 a 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 is introduced on purpose by the United States Department of Defense. Selective availability is scheduled to be phased out over the next few years.

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 World Wide Web.

Sketch map — a map (often hand-drawn) that shows main features of a given area, but that 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, as well as 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 metadate 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. 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 — term often used to describe GIS raster or image data that describe a continuous, smoothly varying phenomenon 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.

Tagged and image file format (TIFF) — a standard image or raster file format that can store black and white, grey scale or colour images in compressed or uncompressed form. Scanners and other devices that create image data often provide output in TIFF. In GIS, Geo-TIFF 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 Bureau of the Census 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 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 to 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 GPS 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 the Defense Mapping Agency) intended to become a universally accepted vector data exchange format.

Vertex — one of a series of x,y coordinates 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 composed 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 (1999). Online resources include the following:

Canada Centre for Remote Sensing	www.ccrs.nrcan.gc.ca/ccrs/eduref/ref/glosndxe.html
Geographer's Craft Project (University of Texas)	www.utexas.edu/depts/grg/gcraft/gloss/glossary.html
GPS World magazine	www.gpsworld.com/resources/glossary.htm
Perry-Castañeda Library, University of Texas	www.lib.utexas.edu/Libs/PCL/Map_collection/glossary.html
United States Bureau of the Census	www.census.gov/dmd/www/glossary.html
United States Geological Survey	edcwww.cr.usgs.gov/glis/hyper/glossary/index

Annex VII. Useful addresses, URLs

GIS packages

Autodesk Inc.	San Rafael, Calif., CA	AutoCAD	www.autodesk.com
Bentley Systems Inc.	Huntsville, AL	MicroStation	www.bentley.com
ESRI, Inc.	Redlands, CA	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 geo-spatial data distribution)	www.coresw.com

Remote sensing image processing systems

ERDAS Inc.	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

Space Imaging	Thornton, CO	Carterra and Ikonos satellites	www.spaceimaging.com
Earthwatch Inc	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

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, UK		
GeoInfosystems	Advanstar Pub., Eugene, OR		
Journal of the Urban and Regional Information Systems Association	URISA, Park Ridge, IL		http://www.urisa.org/

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	Netherlands		www.eurogi.org

Geographic Information (EUROGI)		
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/

Odyssey	GIS papers
ESRI GIS jump station	links to GIS applications around the world
GeoWorld business links	