NCGIA

National Center for
Geographic Information and Analysis

A Review of Spatial Population Database Design and Modeling

by

Uwe Deichmann
University of California, Santa Barbara

Technical Report 96-3

March 1996

Simonett Center for Spatial Analysis
University of California
35 10 Phelps Hall
Santa Barbara, CA 93106-4060
Office (805) 893-8224
Fax (805) 893-8617
ncgia@ncgia.ucsb.edu

State University of New York
301 Wilkeson Quad, Box 610023
Buffalo NY 14261-0001
Office (716) 645-2545
Fax (716) 645-5957
ncgia@ubvms.cc.buffalo.edu

University of Maine
348 Boardman Hall
Orono ME 04469-5711
Office (207) 581-2149
Fax (207) 581-2206
ncgia@spatial.maine.edu

A Review of Spatial Population Database Design and Modeling

Prepared for the

United Nations Environment Programme,
Global Resource Information Database

and

Consultative Group for International Agricultural Research

Initiative on the

Use of Geographic Information Systems in Agricultural Research

Uwe Deichmann

National Center for Geographic Information and Analysis
Department of Geography
University of California, Santa Barbara, USA
Email: uwe@ncgia.ucsb.edu

March 1996

ACKNOWLDEGMENTS

This report benefited greatly from comments and suggestions provided by the following individuals: Jake Brunner (World Resources Institute), Bob Chen (Consortium for International Earth Science Information Network), Hy Dao (University of Geneva and UNEP/GRID Geneva), Griffith Feeney (East-West Center, Honolulu), Patrick Gerland (United Nations Statistics Division), Sindre Langaas (UNEP/GRID Arendal and Department of Systems Ecology, Stockholm University), Ron Witt (UNEP/GRID Geneva), Jaap van Woerden (National Institute for Public Health and Environmental Protection / RIVM, Utrecht), and Vu Duy Man (United Nations Statistics Division). All remaining errors and omissions are, of course, the author’s.

ACRONYMS

AVHRR	Advanced Very High Resolution Radiometer
CGIAR	Consultative Group for International Agricultural Research
CIAT	Centro Internacional de Agricultura Tropical
CIESIN	Consortium for International Earth Science Information Network
CIR	Center for International Research; recently renamed to International Studies Program, US Bureau of the Census
CITAS	China in Time and Space Project (University of Washington)
DCW	Digital Chart of the World
DHS	Demographic and Health Survey
DMSP	Defense Meteorological Satellite Program
EDC	Earth Resources Observation System Data Center
ESCAP	United Nations Economic Commission for Asia and the Pacific
EWC	East West Center
FAO	United Nations Food and Agricultural Organization
FEWS	Famine Early Warning System
GIS	Geographic Information System
GRID	Global Resource Information Database
ICRAF	International Center for Research in Agroforestry
ILRI	International Livestock Research Institute
IMF	International Monetary Fund
ISO	International Organization for Standardization
NCGIA	National Center for Geographic Information and Analysis
OECD	Organization for Economic Cooperation and Development
RIVM	Rijksinstituut voor Volksgezondheid en Milieuhygiene
SDTS	Spatial Data Transfer Standard
UNDP	United Nations Development Programme
UNEP	United Nations Environment Programme
UNFPA	United Nations Population Fund
UNICEF	United Nations Children Fund
UNSD	United Nations Statistics Division
USAID	United States Agency for International Development
WALTPS	West Africa Long Term Perspective Study
WHO	World Health Organization
WRI	World Resources Institute

1. Introduction

An integrated perspective to agricultural research involves the study of interactions between environmental, economic and social factors (e.g., Dvorak 1993). The major objective of such research in a developing country context is to devise agricultural technology that will enable a sustainable increase in food production to match the increasing demand by a growing population. Agricultural technology encompasses external inputs (such as improved varieties and fertilizers) as well as farming practices (soil conservation, integrated pest management). Sustainability in an agricultural context means to ensure an equitable distribution of agricultural products and rural income in the short run without compromising food production and income generation opportunities for future generations. Even more so than industrial technology, agricultural technology is site-specific due to its reliance of natural inputs in the form of sunlight, soils, precipitation, etc. The fact that centrally planned, universal solutions to rural development problems do not work has thus been long recognized in agricultural research (e.g., Pingali *et al.* 1987, MacIntire *et al.* 1992).

Although cross-sectional analysis with countries as the basic unit of analysis are still popular in some areas of development research, the need for detailed, spatially disaggregated data on important variables is therefore generally acknowledged. For agricultural research and targeting of rural technology this implies that disaggregated information on agroecological endowment, physical infrastructure and socioeconomic factors is required at various levels of resolution. This includes high resolution data that may relate to individual plots or parcels as well as medium resolution data at, for example, the level of subnational administrative units. Medium resolution data for large areas are required for cross-national, macro-geographic analysis in a range of applications including

- sampling design/survey planning,
- agricultural targeting (matching technologies),
- cross-national survey analysis, and
- studies of population-environment interactions.

Much emphasis has been put on the generation of generic data sets on physical factors such as climatic variables, vegetation cover and soil properties. Some of these can be captured using remote sensing techniques, while others are based on standardized measurement programs. The development of generic socioeconomic databases at a subnational level is, in contrast, much less advanced. This is due to several reasons:

- the (until recently) predominant emphasis on physical factors in agricultural and environmental studies,
- the limited standardization in the collection of such data - in contrast to, for example, climate data -, and
- the lack of indirect interpolation techniques to estimate the spatial distribution of socioeconomic variables.

These difficulties are also summarized in the report on population data for global change applications prepared by Clarke and Rhind (1992, see also National Research Council 1994). Table 1 lists the major difficulties and complexities associated with the compilation and use of socioeconomic data.

Table 1: The nature of population data (from Clarke and Rhind 1992)

1. Population data are mainly available for political units.
2. Populations are generally conceived as areal units.
3. States and their administrative subdivisions vary greatly in areal and population size.
4. The political division is extremely uneven.
5. Population and scale [*variability of population characteristics changes with aggregation level - i.e., ecological fallacy; u.d.*].
6. States vary greatly in shape.
7. The effectiveness of boundaries as demographic divides varies greatly.
8. States have evolved as political spaces rather than as environmental regions.
9. Population data are rarely related to physical environments.
10. Population concentration is growing nationally and globally.
11. National population data are inadequate for simple analysis of population concentration on a global scale.
12. Population data vary greatly in comprehensiveness and reliability.
13. Almost all population data are residence based.
14. Geocoded data are rare.
15. Coordinated data collection systems are desirable but rare.

The interest in the social and demographic aspects of environmental change and agricultural transformation has been growing steadily. At the same time, geographic information systems (GIS) have been embraced by many demographers and population geographers *as “one of the most important enabling technologies in population geography*” (Jones 1990). This has led to a number of studies and initiatives at various scales that explicitly focus on population dynamics in a spatial context. These include, among many others,

- at a conceptual level, the aforementioned study by Clarke and Rhind (1992) on population data and global environmental change;
- at the national level, Michigan State University's Rwanda Environment and Society project (e.g., Olson 1994), the China in Time and Space project (Hartwell and Hartwell 1993), various data sets developed by the POPMAP project of the UN Statistics Division (United Nations 1994a), as well as many initiatives conducted by national census bureaus;
- at a macro-regional level, a recent study on population dynamics and economic geography in West Africa (Snrech 1995, Brunner *et al.* 1995), or the construction of population distribution map for the Baltic region (Sweitzer and Langaas 1994);
- at the continental level the generation of population distribution data sets for Africa for UNEP's World Atlas of Desertification (UNEP 1992), a data set for Europe by RIVM (Veldhuizen *et al.*

1995), or the Asian spatial data infrastructure envisioned by several Australian universities (Crissman 1993);

- and at a global level, a first attempt at generating a consistent global population data set at NCGIA (Tobler *et al.* 1995), and the ongoing work to generate standardized raster data sets by country for the whole world at the US Bureau of the Census' International Programs Center (formerly Center for International Research, CIR; see e.g., Leddy 1995). The CIR data are distributed by CIESIN.

The material presented in the following sections is aimed at providing an overview of issues and options concerning the development of population related databases that will be of use to the agricultural research community as well as in other population/environment applications. It is hoped that this paper will contribute to ongoing discussions regarding standards and guidelines for the development of spatial population databases which have been stimulated by the work of Clarke and Rhind (1992) and initiatives by CIESIN, UNEP/GRID, UNSD, the U.S. Census Bureau, NCGIA, and the WRI, among others.

The remainder of this paper is divided into three parts. The following section presents a brief summary of important concepts in population geography and spatial demography. This includes a discussion of critical demographic variables required for integrated spatial analysis as well as data sources, accuracy and copyright issues. Section 3 deals with particular issues concerning the development of spatially referenced population databases. Finally, Section 4 presents various spatial modeling approaches aimed at reconciling population data with other geographically referenced databases.

2. Population geography / Spatial demography

Spatial aspects of population dynamics have been relatively neglected within the field of demography. A standard reference work (Pressat and Wilson 1985), for example, does not include an entry for 'geography' or 'space', even though demographers routinely analyze variations in population parameters in a cross-national or cross-regional context. Population geography, on the other hand, is a relatively recent field of specialization within geography and has mostly concentrated on the spatial distribution of people, their characteristics and their relationship to the nature of places (Jones 1990). The main area where both demography and geography take an explicitly spatial perspective, however, is migration. Apart from fertility and mortality, migration is the principal determinant of regional population dynamics. This relationship is succinctly summarized in the fundamental demographic accounting equation:

$$P_1 = P_0 + B - D + IM - OM \ .$$

That is, the population in a given time period (P_1) is the population in the previous time period (P_0) plus the natural increase - births (B) minus deaths (D) - plus net migration which is the difference between inmigration (IM) and outmigration (OM). The time period for which the components in this equation can be compiled may vary depending on data availability from, e.g., months in countries with official registration systems to decades in countries with infrequent census activities. Also, the spatial unit of observation for which data are compiled may vary between high resolution level of enumeration areas or census blocks, on the one hand, and the national level on the other hand.

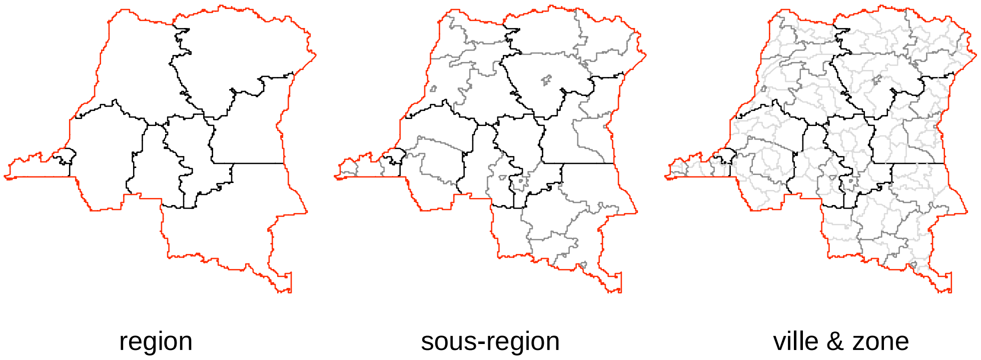
This review paper focuses on population databases used in a wide range of environmental and socioeconomic applications, rather than on demographic analysis specifically. In many applications the demographic mechanisms are not of primary interest, but only their outcome - the particular population distribution. The following section lists the main variables that would be of interest for integrated analysis in the context of sustainable development. The most important sources for these types of data are subsequently described, and the section closes with a brief discussion of copyright and accuracy issues.

2.1. Population Indicators

In developing generic databases that are aimed at serving a heterogeneous user community, the choice of variables necessarily needs to be small. A minimum data set of versatile indicators should consist of those variables that are useful for a wide range of applications, are consistently available across space, and whose characteristics are clearly defined. This last requirement is important when data are used to develop indicators for monitoring purposes (Rossi and Gilmartin 1980). If data definitions vary between data collection periods, no useful conclusions can be drawn from temporal comparisons. A simpler, but clearly defined variable may thus be preferable to a comprehensive, but more complex variable whose definition is subject to interpretation.

In the following paragraphs only a small set of variables that are seen as most important are briefly discussed. Specific issues are elaborated in the boxes which also include standard definitions of demographic and geographic variables that are of use in developing and analyzing generic, spatially referenced population databases. Excellent reference volumes on demographic definitions that provide very comprehensive treatments of demographic indicators are Pressat and Wilson (1985), United Nations (1989) and Bogue *et al.* (1993).

Administrative reference system. Geographic population databases consist of two basic components: the spatial reference system and the attribute information. The former will most often consist of a boundary data set describing the hierarchically structured administrative levels in a country (see e.g., Figure 1).


region

sous-region

ville & zone

Figure 1: The first three subnational administrative levels in Zaire.
Boundary data source: Université Catholique de Louvain and University of Maryland, College Park.

Administrative Hierarchy. The way in which the national territory is partitioned into administrative units varies from country to country. In many cases, the major divisions are historically determined, while the lower levels are designed to facilitate administration and are usually associated with a regional urban center. As an example, the administrative divisions in Zaire are defined as follows (Republique du Zaire 1988):

region: an administrative entity staffed with judicial personnel.
sous-region: an administrative level for coordination, supervision and inspection, but without judicial personnel.
ville: the major city in the region, or an agglomeration with more than 100,000 people, or one that is judged strategic and important by presidential order.
zone: a rural or urban administrative entity, staffed with judicial personnel.

Further administrative divisions in Zaire are termed *collectivité* and *cité*.

It is purely a matter of definition whether the international boundaries are defined as the first or the zero level administrative unit, although, to avoid confusion, it is best to refer, for example, to the regions in Zaire as the first *subnational* administrative region. Since the size of the administrative divisions at a given level in the hierarchy varies greatly from country to country, the rank in that hierarchy is a poor indicator of the spatial detail available in an administrative boundary data set. Many of the second subnational units in Zaire, for example, are larger than some African countries. Instead, an appropriate summary measure is the *average resolution* which is calculated as $\sqrt{country\_area / number\_of\_units}$. In cases, where population distribution is very uneven - e.g., in countries where large areas are uninhabitable - the number of people per administrative unit provides additional information about the available detail.

Distribution and size of the population by administrative unit and/or settlements. The most basic of population indicators is the size of the total population within a clearly defined geographic region. If the regions are small enough, a good representation of population distribution can be achieved.

Total population. A distinction in census enumeration is made between *de jure* and *de facto* population. The former is the population usually resident at a place, excluding visitors and including those residents that are temporarily absent. De facto population refers to the number of people actually present in an enumeration area at the time of the census. Problems occur in cross-national census comparisons of countries with strong economic interdependence (e.g., Sahelian and coastal countries in West Africa) when one country records the *de jure* population while the other uses the *de facto* definitions. Labor migrants may thus be counted, for example, in Burkina Faso as well as in Côte d'Ivoire. Interestingly, IDP *et al.* (1988) state that in Africa, the *de jure* concept is usually preferred in English and Portuguese speaking countries and the *de facto* concept is more prevalent in French speaking countries.

Population by gender and age groups, ideally by 5-year intervals, but at least including the number of children, working age population and elderly. From this information, along with the number of births in the year prior to the census, a number of indirect measures describing population dynamics can be estimated, such as birth and fertility rates. Sex ratios by age groups provide an indirect measure of population mobility (see for example, National Research Council 1993). Low sex ratios - indicating a surplus of women - typically indicate out-migration by economically active males who work in urban or mining areas. This has consequences for agriculture since labor shortages pose one of the major barriers to increased agricultural productivity. In Malawi in 1987, for example, the ratio is close to one until about age 15 (see Figure 2). In economically active age groups, the relative number of males increases in central Malawi as well as in Blantyre district in the South, while most districts in the South and far North show a surplus of women in those age groups (see also Deichmann 1994b, for a similar application using data for Nepal). Patterns in the last maps may be spurious since the absolute numbers of old people are relatively small due to the low life expectancy in the country.

Age composition. A complete census or registration system yields tabulations by one-year age groups. In some countries, age heaping is common when respondents report certain ages (e.g., ending with 0 or 5) more frequently than would be expected. More compactly, a country's or region's age distribution is summarized using five-year age groups. However, for the development of additional indices it is usually desirable to also record the number of children born in the year prior to the census. Information about the age structure also allows the calculation of other important indicators such as the *dependency ratio*: the number of persons under age 15 and over 64 per 1000 persons aged 15-64.

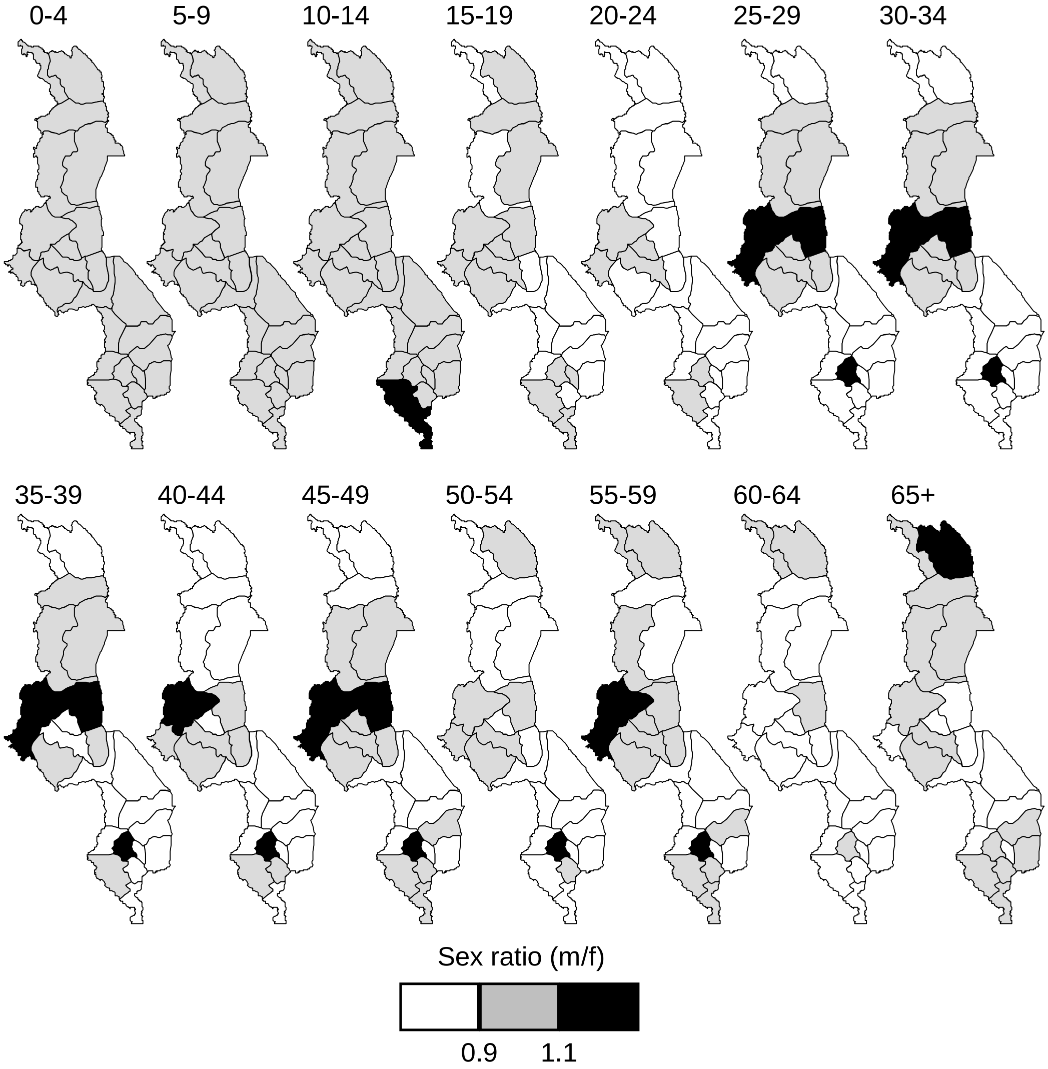


Figure 2: Sex-ratio by five year age groups for Malawi. Data source: Census of 1987.

Sex ratio - the number of males divided by the number of females (often multiplied by 100), or correspondingly the number of events occurring to males divided by the number of events occurring to females. Graphically, the sex and age distribution of a population is best summarized in a population pyramid. Heilig (1994) developed a PC based demographic visualization software that features animated national population pyramids illustrating changes over time.

Rural versus urban population. Place of residency is frequently used as a crude proxy for economic activity. The total population within an administrative unit is therefore often broken down into rural and urban population. Urban dwellers are assumed to be more likely employed in the secondary or tertiary sector, while rural population, especially in developing countries, are most often active in the agricultural sector.

Rural population - although very important for environmental and agricultural research - is an indicator that is often frowned upon by demographers. The reason is that definitions of what constitutes an urban settlement vary widely between countries; e.g., from towns with at least 500 people to those of 20,000 or more. Sometimes urban population only refers to the capital, or a functional definition is followed such as "major administrative centers." Section IV.A. in United Nations (1993a) contains a comprehensive list of definitions by country. It is clearly necessary to design methods for indirect estimation of consistent rural/urban figures for cross-national analysis. This will be discussed further in a subsequent section.

Population movements. Migration is one of the fields in demography that is most difficult to study due to the problem of obtaining direct measurements. This may explain the relatively low representation of migration issues in the otherwise comprehensive compendium published by Bogue *et al*. (1993). Migration data are usually collected using a recall element in the census, but are often considered unreliable. Migration occurs within and between countries. A specific aspect of migration are refugee flows due to economic reasons, conflict or environmental crises (Segal 1993). These can trigger very sudden movements of large numbers of people, and often result in long-term displacement. Such abrupt occurrences put a large strain on the resource base of the receiving area and consequently are very important from an agricultural management perspective.

Population dynamics over space. Though inherently spatial, migration rates are generic demographic indicators that are frequently reported in census publications and yearbooks. The three most important indicators are the *in-migration rate* - or *immigration rate* when international migration is analyzed - which is defined as the number of migrants into an area divided by the area's total population (typically multiplied by 1000), the *out-migration rate* which is similarly defined as the number of out-migrants per 1000 population, and the *net-migration* defined as in-migrants minus out-migrants per 1000 persons in the area.

Temporal dynamics are important in explaining current patterns and to make informed planning decisions affecting future developments. Monitoring and predicting future population size and distribution depends crucially on the availability of reliable variables that reflect temporal population dynamics. These can be summarized by demographic indicators that provide a measure of current population change, such as births, deaths, fertility and mortality rates. Or, if sufficient information is available, actual population totals within a region can be compiled, estimated or projected for several time periods. From these time series, growth rates can be calculated and used in analysis or in the production of cartographic output.

Birth, death, fertility and mortality rates. Birth and death rates are defined as the number of live births (or the number of deaths) divided by the mid-year population in that year (*crude birth/death rate*), and are usually expressed per thousand. The crude birth rate is obviously influenced by the number of women in reproductive age, the number of married women, etc. Similarly, the crude death rate is affected by the age distribution within the study area. A complementary indicator is the *life expectancy at birth*, which, however, is often unavailable at the subnational level. Complementing crude birth rates are the *general* and *total fertility rates*. The former is the number of live births divided by the number of women in child-bearing age (usually 15-49). The total fertility rate (TFR) is the sum of the age specific fertility rates which are the number of births within a given one-year age group. The TFR is the most widely used measure of fertility and due to the difficulty of obtaining the necessary detailed data in many countries, several indirect estimation methods exist. For example, Rele's method allows one to estimate TFR for the census year and for five years prior to the census based on the five-year age distribution from the census. An interesting application is reported in United Nations (1988) which presents maps of recent TFR change at the subnational level in several South-East Asian countries.

Population dynamics over time. Population growth is determined by the fertility, mortality and migration rates within a country or region and is thus indicated by the variables described above. A measure of population change is the growth rate which can be calculated on the basis of two population estimates (e.g., intercensal growth rate) as

$$r = \frac{\log_e\left(\frac{P_2}{P_1}\right)}{t} \qquad ,$$

where r is the average rate of growth, P_1 and P_2 are the population totals, for example, in the first and second census, and t is the number of years between the two enumerations. To obtain the average annual *percentage* growth rate, r is multiplied by 100. To estimate the population for a specific year using the base population and a specific growth rate, the following formula is used:

$$P_2 = P_1 \cdot e^{rt} \qquad .$$

A negative rate yields an estimate for a point in time prior to the base year. Since census years vary by country, these equations are useful for standardizing population figures across in the development of a multi-country database.

Settlement size and location. The acceleration of urbanization in many regions in the developing world has been one of the most important demographic processes in the last decades. Concentration of population due to rural-urban migration and higher urban fertility rates in developing countries have resulted in the rapid growth of many cities. One implication of this development is the increase of environmental and social problems within large urban areas as unbalanced growth has created hard-to-manage mega-cities. On the other hand, this development has considerable implications for agricultural research, since the growing number of urban consumers has the potential to provide an incentive for agricultural surplus production and to increase rural incomes (i.e., urban-rural linkages; see, for instance, Snrech 1995). Information about settlement patterns is also required in planning the provision of services such as extension, health, and transport infrastructure (e.g., Rondinelli 1985).

Urban databases: Although it is relatively straightforward to obtain latitude/longitude coordinates for urban centers from gazetteers which are increasingly available in digital form, the number of geographically referenced urban databases is relatively small. The United Nations regularly publishes the population sizes for the largest cities in the World (e.g., cities with more than 100,000 inhabitants). The UN Centre for Human Settlements (Habitat) has started the compilation of comprehensive urban databases for large cities in preparation for the Habitat II conference in Istanbul in 1996 (United Nations 1995b). At the opposite, higher-resolution end, digital maps in which population totals are tied to the location of towns with a population of 5000 or more have been produced by CIESIN for Mexico. This data set includes an urban space, time-series spreadsheet with population data back to 1921 for about 700 of the largest cities in Mexico (http://sedac.ciesin.org/home-page/mexico.html). For villages in Burkina Faso and Niger the USAID Famine Early Warning System (FEWS) linked population figures to village locations. In many cases, it proved to be difficult to match census data and administrative names to village names recorded on often outdated maps. As reported by Brunner, Dalsted and Arimi (1995), a considerable number of villages in Niger could not be located:

"*Village names and locations were digitized from IGN maps. Village populations and other demographic variables were acquired from the 1988 census. Because of changes in name and location between the publication of the IGN maps and the census, many settlements identified on the maps are not found in the census database and vice versa. As a result, the original village database only included 67% of the total 1988 population of 7,244,000. Settlements identified in the census database were processed and aggregated by the Service National d'Information Sanitaire (SNIS) to produce the Fichier National de Localité (FNL). By joining the village and FNL databases, 83% of the population was accounted for.*"

In countries where rural population is not largely organized in villages, the percentage population accounted for will be much smaller, since smallholders living in isolated settlements will be more difficult to account for. The WHO/UNICEF HealthMap project (1995) is, similarly, aimed at providing similar village level data on population, health infrastructure and disease incidents for all tropical countries.

Specific indicators. Basic demographic variables are the most important component of a population database as they allow for the development of derived indicators that provide more immediate information about social processes of interest. Special purpose applications, for example, may require counts of total population to calculate rates of incidence or population at risk. An example is work in progress to produce an atlas of schistosomiasis infection rates by administrative unit for all African countries (Nuttall 1995). A multi-purpose database needs to provide these basic data. As the next step, derived generic indicators are required that allow for the monitoring of social, health and educational conditions in a city, region or country.

Agricultural policy, for instance, is aimed at improvements in nutritional standards (i.e., food security) and rural income generation. In order to target needy population or to identify the effectiveness of agricultural intervention, the outcome of policies needs to be evaluated on the basis of standardized measures of well-being.

Related to this are plans to create a generic set of spatially referenced variables characterizing farming systems (Carter *et al.* 1992). A set of "general-purpose" questions could be included in field data collection efforts that are carried out by various groups within a region. Over time, a comprehensive database would become available that provides wider spatial coverage than any single project could achieve. The problem in developing generic databases is that the choice of indicators is difficult. General-purpose means that many interests will be involved, each having specific requirements. Compromises are necessary in order to keep the number of variables at a manageable level while addressing the broadest possible requirements. More than one group of researchers has encountered problems when trying to define a "minimum data set."

Social Indicators: A comprehensive list of candidate variables and indicators is contained in the POPMAP manual (United Nations 1994). A working group representing several UN agencies (UNSD, UNFPA, UNDP and UNICEF) has defined a set of priority indicators for monitoring the achievement of social goals (United Nations 1993). These include basic health and education statistics like infant, child and maternal mortality, access to safe drinking water, adult literacy rates, as well as factors measuring the status of women. Appendix A lists the full set of thirty-four indicators suggested by this group. While the objectives of this initiative were aimed at the national level, most of the indicators would be even more relevant at the subnational level to improve targeting and monitoring.

2.2. Geographic Indicators

Geographic indicators are reviewed in Woods (1982), Jones (1990), Bähr (1992), and Plane and Rogerson (1994), among many others. Explicitly spatial summary measures of population data relate to the distribution of people over the land area of a country or region, or to specific features of the land area. The *population density*, or number of people per unit of land area (e.g., people per square km) is the most common summary measure of population distribution:

$$D_j = P_j \ / \ A_j \quad ,$$

where D_j is the density in areal unit *j* and P_j and A_j are the corresponding population and land area of the unit. While conceptually straightforward, population density can be very misleading when population distribution within the region for which the measure is calculated is very uneven. This is best illustrated by looking at the population density of countries. China, for example. has a very high overall population density, but some regions in the Western parts are virtually uninhabited.

Also, the total area of an administrative unit that is used in the density calculation sometimes includes parts of large water bodies. In Rwanda, for example, the actual population densities of some administrative units that contain large areas of one of the Rift Valley lakes were underestimated in the official census publication by up to fifty per cent. It is therefore preferable in most cases to relate the population present in a given areal unit to a more precise measure of land area, or even to a land use category. This is termed *dasymetric mapping* in cartographic applications (e.g., Plane and Rogerson 1994). For example, lake areas can easily be subtracted from the total area using a GIS. For agricultural applications, (rural) population per unit of arable land (agricultural population density) is a more precise measure of pressure on the resource base than overall density. With larger spatial detail, population distribution within each areal unit is likely to be more homogenous, thus leading to more meaningful population density figures.

Complementing population density, the inverse relationship, the land area per person in a given area is a useful measure to describe resources available to individuals. As nicely illustrated in the recent *National Atlas of Sweden* (Statistics Sweden 1993), if the population in region *j* were completely evenly distributed, each inhabitant would have an area *a* available which has the shape of a hexagon and whose size is $a = A_j / P_j$. The *average* distance to each of the nearest six neighbors can then be calculated as

$$\bar{d}_n = \sqrt{2a / \sqrt{3}} \quad .$$

If instead of areal totals, observations on individuals or settlements are available, one can also calculate the *actual* average distance, d_a, to the nearest neighbors. This observed mean distance to the nearest neighbor can be compared to the theoretically expected value that can be compactly expressed as

$$\bar{d}_t = \frac{1}{2\sqrt{\frac{n}{A}}} \quad ,$$

where *n* is the number of observations (e.g., settlements) and *A* is the area of the region studied. The ratio $\bar{d}_a / \bar{d}_t$ approaches one, if the observations are randomly distributed, and zero if population concentration is large. A regular distribution will lead to a ratio that approaches a limiting value of 2.149 (Bähr 1992).

In comparing population figures for a set of areal units (e.g., districts in a country, or enumeration areas within a district), several other summary statistics can be derived that describe population distribution. A measure of concentration that is based on areal units rather than individual observations is Hoover's index of concentration (or simply *dissimilarity index*). This measure indicates how evenly population is distributed within the overall study area. It is calculated as

$$I = 50 \sum_{j=1}^{n} \left| p_j - a_j \right|$$

where p_j is the proportion of the entire region's population resident in areal unit *j* and a_j is the proportion of that unit's land area. The multiplication by 50 produces a percentage figure which can be interpreted as the percent of the total population that would have to move in order to create an even distribution of population in the country. The index is useful for comparing the level of population concentration within different regions, or to analyze concentration effects over time. For example, based on estimated population figures derived using census data, the index of concentration for Malawi at the district level decreased between 1960 and 1990 from 32.2 to 26.4, while it increased during the same period for the sous-regions in Zaire from 37.3 to 45.5. Thus, while the population distribution in Malawi appears to have become more even at the district level, population is now more concentrated at the level of sous-regions in Zaire. The Hoover index of concentration is closely related to the graphical representation of concentration levels by means of the Lorenz curve and its associated summary statistic, the Gini coefficient.

Many other indexes of concentration, nearest-neighbor measures, centro-graphic parameters (e.g., the mean center of population in a region), and potential measures have been developed. From an operational standpoint, potential measures may be the most useful as they relate closely to the concept of accessibility - an important factor in socio-economic and man-environment studies. Population potential measures will be discussed in more detail in the modeling section of this paper.

3. Data Source and Data Issues

3.1. Sources of Population and Boundary Data

Population data sources have been reviewed with significant detail by Bogue et al. (1993, Chapter 3), Clarke and Rhind (1992), whose report includes a survey of materials that were available in British libraries, and Tobler *et al.* (1995). Rhind (1991) and United Nations (1994a) also discuss issues relating to statistical demographic accounting. Benzine and Gerland (1995) compiled information on population resources available on the Internet. In light of these recent overviews, the following paragraphs provide only a brief description of major data sources.

3.1.1. Primary data

- *Population, housing and agricultural censuses*. These are conducted more or less regularly by national statistical offices, often supported by the United Nations or bilateral donor agencies (United Nations 1992). The UN Statistics Office and the US Census Bureau regularly compile lists of the types of censuses that have taken place in each country of the World. The detail at which census data are made public varies widely by country - both in terms of geographic resolution and the scope of population characteristics. Often, only the most important indicators are published in highly aggregate form and with significant delays. In those cases, more detailed information can only be obtained at the source: the national statistical office. Frequently, research organizations compile handbooks and evaluations of census activities in particular regions, see, for example, Goyer and Domschke (1983), Cho and Hearn (1984), Domschke and Goyer (1986), ESCAP (1988) and IDP *et al.* (1988).
- *Civil registration systems*. Population registers provide continual information about the distribution of a country's population. However, few developing countries maintain an up-to-date registration system, and even in a number of industrialized countries population registers are either absent (e.g., United States, Britain), or frequently of limited accuracy (Rhind 1991).
- *Administrative data (governmental or non-governmental)*. This includes electoral, vehicle registration, tax or insurance records which are not generally publicly available, but are the basis for aggregate statistics.

- *Household sample surveys*. Many relevant socioeconomic indicators are collected using sample surveys, several of which are conducted at regular intervals. Examples are the USAID sponsored Demographic and Health Surveys (DHS), the UN National Household Survey Capability Program, the World Fertility Survey, and the World Bank's Living Standard Measurement Surveys and Social Dimensions of Adjustment Program. Few of these have so far been systematically referenced to spatial coordinates, although census boundaries are often used to design the initial survey sampling frames. A project undertaken jointly by the US Census Bureau and USAID/REDSO Abidjan is currently testing the feasibility of spatially referencing the DHS surveys and linking these data to other surveys. Clearly, much could be gained through better utilization of existing survey data.

3.1.2. Secondary data

If access to primary census and administrative records is not possible, secondary data generally need to be used. These are compilations based on primary census data published in fairly aggregate form.

- *National statistical yearbooks*. These are published by most national statistical offices. Most follow a standardized scheme in which data are published in fairly aggregate form covering climatic, demographic, health and economic conditions in the country. Many countries now publish at least total population figures by administrative units and often a map is included as well. Frequently, historic data are provided to highlight temporal trends.
- *International organizations' compilations*. United Nations and similar organizations compile statistical tables on particular topics at regular intervals. Rarely, however, are subnational break-downs provided because the focus is generally on world-wide cross-national comparison.
- *Gazetteers, commercial yearbooks, atlases*. Several private publishers produce yearbooks or fact books regularly. The annual yearbooks produced by *Europa Publications* (London), for example, not only provide updated essays on geography, history, politics and economics for nearly all developing countries, but also include a data appendix for each country which sometimes includes subnational figures. Similar information can be found in Munro (1990) and many other yearbooks. In collaboration with official census publications, several population atlases are also available which provide maps and data by administrative units. Some excellent examples are the *Population Atlas of China* (Population Census Office 1987), the *Population Atlas of India* (Roy 1988), or the *Atlas du Vietnam* (Vu and Taillard 1994).
- *Commercial digital data*. The spread of GIS technology in Europe and North America is to a large extent driven by marketing applications. The great demand for high resolution demographic information that can be easily mapped is witnessed by the increasing number of commercial companies offering census information and simple mapping packages. For a few developing countries with promising economic potential commercial vendors are also offering data - for example, for India at the district level and for Mexico at the municipio level (Klein 1995).

- *Satellite estimates and rooftop surveys*. Indirect techniques for counting people using remotely sensed imagery or aerial photography have been tested extensively (e.g., Clayton and Estes 1980, Lo 1986, Stern 1986, Paulsson 1992). The essence of these approaches is to identify land use categories by classifying satellite images, and to associate different land uses with particular population densities. In more local studies, in particular in urban applications, individual houses are identified and counted. Using information on average household occupancy rates, estimates of urban population are derived. This obviously requires high resolution imagery or the use of aerial photography. A recent agricultural survey in West Africa and Ethiopia, for example, used a rooftop survey to derive coarse resolution estimates of rural population densities (Wint and Bourn 1994; see Figure 3). The difficulties associated with image classification and the large resources required to cover large areas with aerial photography have so far prevented the use of these techniques for the compilation of high resolution data by administrative units for whole countries. For the time being, the development of population estimates based on remotely sensed imagery will thus be limited to particular, small-area applications.

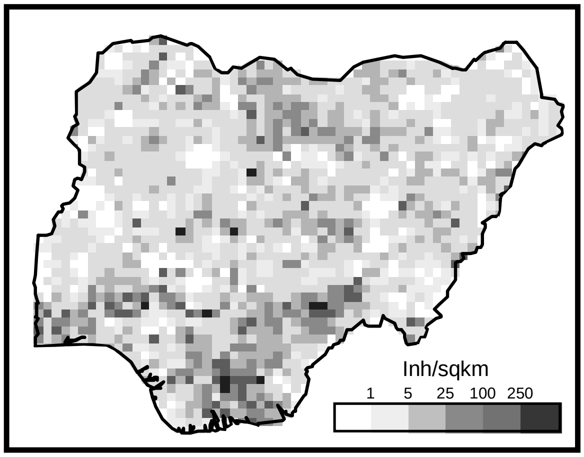


Figure 3: Rural population densities in Nigeria derived from an aerial rooftop survey. Data source: Environmental Research Group Oxford (see Wint and Bourn 1994).

3.1.3. Coordinate data

For use in geographic analysis, it is essential that population data are linked to a sufficiently accurate boundary data set. As population data are typically compiled by administrative units, the vector data model is the most appropriate for the compilation of population databases. Increasingly, standardized digital census boundary data sets are produced by national statistical offices, sometimes as part of an elaborate database and software design. For example, the US Census Bureau developed the TIGER system that includes digital boundary data down to the block level as well as retrieval software (Marx 1990, Davis *et al.* 1992). The approximately seven million blocks (in 1990) are the lowest areal unit in the hierarchy of the US census geography.

In developing countries, administrative data sets are often produced as part of the planning process for aid projects. For example, the USAID FEWS project produced administrative data layers for most countries in the Sahelian region and Southern Africa. The UN Food and Agricultural Organization

(FAO) regularly develops administrative boundary data sets for use with population data and agricultural production statistics. Similarly, administrative units are digitized by many of the CGIAR centers - notably CIAT, ILRI and ICRAF - either at high resolution for project specific applications or as generic base layers for use at small geographic scales.

3.2. Accuracy Issues

It is useful to distinguish between the accuracy of the attribute data (e.g., demographic variables) and that of the geographic data. The accuracy of population figures, as determined through post-enumeration surveys, varies greatly by country. The two main types of error are *coverage error* (under-enumeration or over-enumeration) and *error of content* (e.g., incorrect responses or coding errors). In cases where registration is mandatory, the coverage error for population totals has been found to be as low as 0.3 percent in the case of Sweden (Clarke and Rhind 1992). On the other hand, the estimated error determined through post-census enumeration in some countries exceeds ten percent (IDP *et al.* 1988). Even in the United States significant census undercount is assumed to occur. This prompted several city governments to demand an adjustment of census figures (especially in low-income neighborhoods) from the federal government because federal grants depend on population totals. In cases where the distribution of national funds is based on population figures it is tempting for local governments to exaggerate the number of inhabitants within their constituencies. The 1963 census in Nigeria, for example, is widely believed to be of limited accuracy because regional figures have likely been inflated. Due to the population-based system used to divide national revenues among the federal regions there is a clear incentive for regional authorities to inflate the numbers of their inhabitants. Even at a national level, considerable debate exists regarding population totals.

Furthermore, population counts are often outdated shortly after they have been released. Generally, demographic processes are operating with slow speeds and population dynamics are quite predictable over short time periods. However, political events can force radical changes in the distribution and composition of regional populations. For example, events such as the turmoil in Rwanda and Bosnia can severely change population distribution practically overnight, rendering published census figures more or less obsolete for all but historical studies. Major shifts of population occur due to political and even environmental displacement, and, in the short run, it is difficult to adjust regional figures accordingly. In fact, even at a national level and over longer time horizons, multilateral statistical agencies have problems accounting for sudden changes due to external shocks (e.g., Uvin 1994). Also, demographic fundamentals do change over time - witness the drop in fertility rates in Asia in recent years (see Feeney 1994), making it necessary to adjust forecasts constantly; Stycos (1994) refers to this as "*the short life expectancy of population projections.*"

Accuracy issues in the context of spatial databases have received significant attention in GIS research (Goodchild and Gopal 1989). At the most basic level, this refers to *positional accuracy*: features are recorded in the digital database in accordance with their true geographic location in the real world. The problem with administrative boundaries is that they are artificial constructs that - for the most part - cannot be surveyed on the ground. The precision with which boundaries are defined varies by country and also depends on the level within the administrative hierarchy. Typically, at higher levels (e.g., provinces), administrative units are well defined and often follow natural features, such as rivers. At lower levels, however, boundaries may not be defined at all. For example, in Ghana approximately half of all districts are not precisely defined cartographically (Jake Brunner, pers. comm.). Thus, we may know that two neighboring villages are located in different districts but the exact boundary between them has not been defined. The problem is aggravated by the absence of proper cadastrial and land ownership records. Territorial disputes also introduce uncertainty. There are numerous examples including the Peruvian/Ecuadorian border, the island groups in the South China Sea, or parts of Jammu and Kashmir which are claimed by India and Pakistan. According to Indian census geography, several districts cross the international boundary that is defined by Pakistan.

Sometimes more important than positional accuracy is *logical consistency*. For example, if two districts share a common boundary in the real world, they should also share a boundary in the GIS database. Whether the length and geographic position of the boundary is absolutely correct may be less important (of course, in an ideal world of perfect positional accuracy, logical consistency is assured automatically). Especially where the primary objective is visualization of spatial patterns of census indicators, or, if spatial analysis based on neighborhood relations is the primary objective, so-called "*cartoon maps*" are often quite sufficient. For these maps the spatial reference system (i.e., as defined by the map's projection and associated parameters) is typically unknown, they are digitized or scanned from maps at small cartographic scales, and consequently show little detail. Problems occur when such coverages are the only information available such that *ad hoc* approaches such as rubber-sheeting are required to bring them into a consistent reference system in order to integrate the information with other data. The loss of accuracy involved in such operations is typically rather large which renders such maps inappropriate for high resolution applications.

Another aspect pertinent to logical consistency relates to the way in which, for example, digital spatial data on administrative units matches related information in the same or in other thematic coverages (e.g., Prévost 1995). An example for this topological integrity or "cross-layer consistency" is where district boundaries are known to be defined by the course of a river (though it may still be unknown whether the center or one of the river banks represent the boundary). In a comprehensive, general purpose database it is desirable to achieve a perfect match between the district boundary and the digital lines representing the corresponding river in a hydrological data layer. Similarly, points or polygons representing particular urban areas should fall into the administrative units that they belong to, village or city locations should be connected to the roads that pass through them, and forest boundaries should not overlap with urban areas. In contrast to positional accuracy, the absolute position of a geographic feature or the intersection of two or more features is less important than the correct representation of the relationships. This concept can also be used to simplify maps in order to communicate the most important information. Examples are the highly generalized London subway map, or the cartographic model of the transport infrastructure of Vietnam in Vu and Taillard (1995; see also several examples in Tufte 1990).

Logical consistency is required for integrated spatial modeling where the results depend on the characteristics and spatial location of several sets of spatial features. Furthermore, this topological integrity is a prerequisite for what is now fashionably called *interoperable* GIS, where interoperability can be defined as "*the capability to access transparently remote data and processes in an open* *environment*" (Tryfona and Sharma 1995). This goes beyond the definition of common data structures and translation routines and is aimed at making spatial data manipulation and analysis functions independent from specific software and hardware systems.

The wide distribution of data sets has increased significantly since global networks allow the speedy exchange of large data sets. While the exchange of information via the Internet means that availability of data has improved tremendously, it also encourages an informal approach to data documentation. Thus, many data sets come with insufficient information for the potential user to assess their quality and lineage (e.g., the sources and prior processing steps). The lack of adequate meta-data may be the single most important impediment to proper use of secondary data sets.

3.3. Copyright Issues

Producers of analog or digital information on administrative and census data are most often national mapping agencies (e.g., survey departments) or census bureaus. With decreasing public funds, these agencies are under increasing pressure to recover a share of their operating expenses. Consequently, data sets are frequently sold at a price closer to the proportional cost of producing the data rather than at reproduction cost. This implies that official agencies restrict further dissemination - for example, through the Internet. This development is well reflected in the size of the legal paperwork involved in obtaining official data sets. The same is true for nearly all industrialized, and increasingly also for developing countries. Official data - required, for instance, in planning or academic research applications - thus become a commercial product, and the price of purchase make it increasingly difficult for less well-endowed organizations to get access to critical information. Unfortunately, the information policy in the United States, where public data are not copyrighted (with the argument that citizens should not have to pay for services that were financed through taxes), is becoming increasingly rare.

The consequence is that many users will decide to replicate the work done by the official agency by digitizing source maps possibly with lower accuracy. This will also lead to the proliferation of multiple, inconsistent versions of the same data set. However, in some countries, even products derived from an official paper map will still fall under the copyright of the mapping agency. Clearly, there is a need to define the balance between a restrictive data policy aimed at cost recovery with an approach that guarantees the wide dissemination of essential data upon which good analysis and informed decision making so crucially depend.

4. Population Database Considerations

Demographic mapping and other population related applications - e.g., marketing, service provision planning - have arguably been the main driving forces behind the explosive growth of the desktop mapping sector particularly in the United States. The reasons for this growth have not so much been the development of suitable, easy-to-use software packages, but rather the availability of detailed census data at fairly low cost. Increasingly, census agencies are publishing the results of censuses and other enumerations in digital form. These are often packaged with suitable software by private vendors and are applied extensively by users with varied backgrounds and objectives. In contrast to many other types of spatial databases, the coordinate data (e.g., boundaries) of a geodemographic database typically represents only a small part of the overall data volume. For each census district, several tens or hundreds of variables may be available that summarize socioeconomic characteristics of the population within that district.

Building a complete census database down to the block level is a tremendous task. The U.S. TIGER system includes data for about 2 million census blocks. For France, a digital census database is available for about 32,000 communes. The logistics involved and the specific software and hardware issues relevant to the development of such databases are beyond the scope of this paper. The focus is instead on building moderately sized population databases for use in analysis and modeling in combination with other socioeconomic and environmental variables.

In any database development, choices will have to be made regarding design and overall structure. Most importantly, there is a trade-off between an elegant and efficient data model on the one hand, and one that is robust across hardware and software platforms on the other. In generic database design, system independent data modeling approaches exist (e.g., Batini et al 1992) which are designed to work on any particular software system. The attribute (i.e., demographic data) part of a spatially referenced population database could be handled in this way. In dealing with the spatial data component, however, standards have not been developed to the same extent despite considerable discussions about, for example, a national spatial data transfer standard (SDTS) in the United States, and similar efforts in Canada, Europe and at the International Organization for Standardization (ISO).

In developing a database that is likely to be used by a heterogeneous user group, it is therefore preferable to choose a simple and robust design that is likely to be transferable to a wide range of software packages, rather than exploiting the sometimes elegant and unique features of a particular system. This is the rationale behind the suggestions for a simple database design in the following paragraphs (see also United Nations, 1994a).

4.1. Elements of a Small Area Population Database

The types of population data sets used in integrated analysis and modeling as discussed in Section 4 fall into three categories. The central component consists of the administrative boundaries - typically a vector data set - and a link to a database that includes identifiers and demographic data. A complementary view is to focus on specific settlements, which are most compactly represented by the coordinate pair defining the town or village center (although, larger urban areas could be represented by polygons). A settlements database is thus best designed as a point coverage with a link to a database that includes town identifiers and any available data. Finally, for most environmental applications, it is useful to render population data compatible with data on physio-geographic variables such as climate, vegetation or soils. These are most often stored as regular raster grids, where each grid cell contains the value of the variable that has been recorded or estimated for the corresponding location. The resolution (cell size) of the raster may vary from half degree in many global data sets down to a few meters in high resolution applications. Section 4 focuses on techniques to convert population data from a polygon and point data set into a regular raster grid.

It is important to point out that any location in each of the three representations - areal units, points and raster grids - is linked to the other representations by a real world coordinate system. Thus the particular administrative unit into which a settlement falls can be determined through a standard point-in-polygon operation, and the administrative identifier (e.g., `AdminId`) is added to the towns attribute database. Likewise each grid cell in the raster surface is assigned to a particular administrative unit, and similarly, a settlement's center point falls into a particular grid cell. In this case, the grid cell value is the corresponding administrative identifier. Figure 4 illustrates these links.

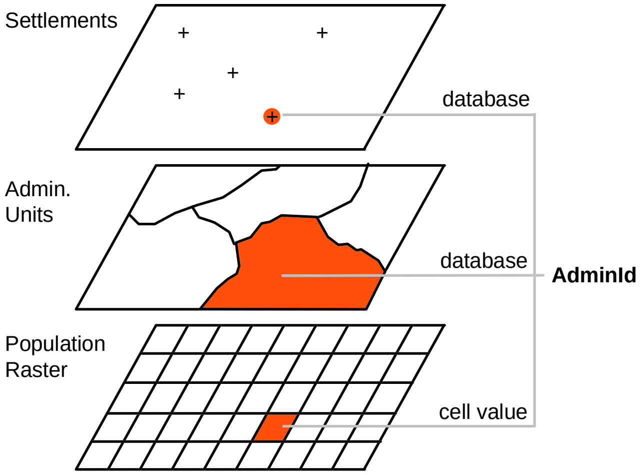


Figure 4: Links between different representations of a population database.

4.1.1. Coding scheme

Each administrative unit in the population database needs to obtain a unique identifier. This code is used throughout to associate each polygon or settlement with a census district, and serves as the main key for relating the spatial features to external tables containing various attribute information. Many countries have identification systems in which, for example, all administrative units belonging to a given level in the hierarchy are numbered sequentially. A concatenated identifier will then uniquely label each polygon in such a way that identifiers for individual administrative levels can still be reconstructed.

For medium resolution international databases, however, it is preferable to design an identification scheme that is comparable across countries. This can be achieved by using the same coding scheme for all countries, while storing country-specific administrative codes (if available) as a separate attribute. An international code can be added to the generic identifiers whereby the nation becomes the root level in the administrative hierarchy. An identifier for a second level administrative unit such as a district could thus be composed of:

country code + code of 1st level unit + code of 2nd level unit

where the country is here defined as the level zero in the administrative hierarchy, a province may be the first subnational level, and the district the second subnational level. This coding scheme is implemented, for example, in the POPMAP software which is a desktop mapping program for population applications produced by the UN Statistics Division (United Nations 1994a)

In order to achieve maximum compatibility, it makes sense to use standard codes whenever possible. The trouble, however, is that at least at the country level there are several coding schemes currently in use (confirming popular wisdom that "*the nice thing about standards is that there are so many to choose from*"). The most commonly used codes at the national level are the three digit country codes assigned by the United Nations Statistics Division (UNSD), and the two-character and three-character codes established by the International Organization for Standardization (e.g., the two and three letter *ISO Alpha-2* and *Alpha-3* codes; see United Nations 1994c). The ISO character codes have intuitive appeal because they are typically based on an abbreviation of the English language country name. However, country names change occasionally (Upper Volta became Burkina Faso, Burma became Myanmar), and the English basis of the two or three letter abbreviation is not always followed (e.g., KHM for Cambodia or ESP for Spain). Two additional arguments against a character-based coding scheme are that a long character string takes up more disk storage than an equally long integer, and that a numeric identifier can be more easily manipulated in database operations (e.g., selection or cross-tabulation).

The numeric UN codes (listed in United Nations 1994c) are aimed at identification although for some applications they are combined with additional codes (prefixes or suffixes) to create a socioeconomic classification scheme. The three-digit country code was originally based on an alphabetical listing of country names. With changes in country names, however, the numeric codes do not necessarily relate to official country names anymore (e.g., 854 for Burkina Faso but 108 for Burundi). Nevertheless, the three digit UN code represents a widely used coding scheme. Generally, codes have been assigned to nations. However, ISO and/or UN codes also exist for some territories belonging to a specific country (e.g., the British Indian Ocean Territory). Several other coding schemes exist, including those used by the European Union or the International Monetary Fund (see IMF 1994). Yet, it is here suggested to use the UN three digit country codes as a component in a unique coding scheme for the administrative hierarchy. If required in a global database, a higher level code could be used to determine, for example, the major regions used in UN publications (e.g., South-Eastern Asia, Western Asia).

Attached to the three digit UN code can be numeric codes for each level in the administrative hierarchy. For a medium resolution database, three levels in the subnational hierarchy will usually be sufficient for medium resolution databases. For national applications, more levels may be required dependent on the available detail. Administrative and naming conventions vary between countries. For example, many nations are partitioned into a set of only a few large regions (e.g., South, West) which are not usually used for administrative purposes, but represent an additional level that can be considered in a database and that may necessitate an expansion of the administrative code. Also, the width of each sub-element of the complete code needs to be determined. In most cases, there will be no more than 99 administrative units for each higher level. However, exceptions exist, such as several provinces in China which consist of up to 150 counties.

To sum up, a unique identification scheme covering three subnational levels may result in an administrative unit identifier of 180080403, where 180 is the three-digit UNSTAT code for Zaire, 08 is the two-digit code for Kasai-Oriental *region*, 04 represents the Kabinda *sous-region*, and 03 identifies Ngandagika *zone*. Note that the concatenated code can still be used to select or manipulate individual subsets in the administrative hierarchy. For example, to select all administrative units belonging to Kabinda sous-region one could select ID >= 180080400 and ID < 180080500. Also, using algebraic expression and the modulo operator the codes can be split into their component parts if separate fields are preferred. Some database systems allow a redefinition of a sequence of attributes such that each component of the identifier could be stored in a separate field, while a redefined (but not duplicated) field represents the concatenated unique identifier used in relations or other database operations.

Although the unique identifier contains a complete basic description of the administrative unit and additional information could be stored externally, it is generally preferable to include easier to interpret information in the database table as well. This may include the name of the country or its 3-letter ISO code as well as the names of each administrative unit in the hierarchy. Since disk space is not usually an issue, it is suggested to store this information in the basic attribute table to allow for easy identification, selection or labeling.

4.1.2. Land area

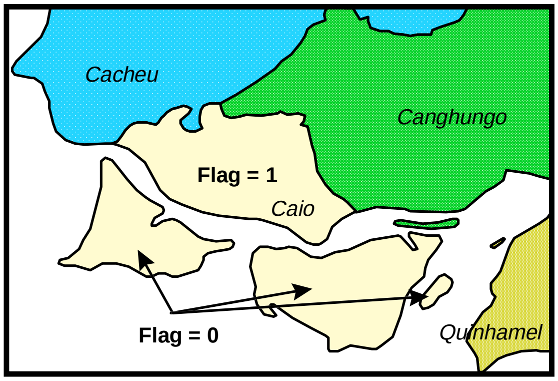
A generic variable that should be stored within the main attribute table is the areal extent of the administrative unit. While this sounds straightforward, there are several issues to consider. The areas of a set of polygons are easily computed within most GIS systems. However, in order to obtain correct areas, the coverage needs to be converted into an equal area projection first. Often, an administrative unit will consist of several individual polygons. For example, a coastal district may consist of a mainland polygon and several islands (as further discussed in the next section), the areas of which need to be summed before densities are computed. Due to digitizing error, generalization and other sources of error, computed areas will generally not be exact. If available, published areas of administrative units are therefore often preferable to GIS calculated figures.

In both cases, using computed or official figures, the question is whether the total area of the polygon/administrative unit is necessarily the correct basis for calculating density measures. In countries that contain large water bodies, for example, population or agricultural cultivation densities would be significantly underestimated if a large part of an administrative unit is, in fact, covered by a lake. Other land cover features that preclude human habitation could also be considered, although data on inland water bodies are usually easiest to obtain. These areas can be subtracted from the total administrative unit areas. Although in some databases the lake polygons are kept within the administrative unit coverage, it is generally preferable to keep different themes in separate layers.

4.1.3. Disjoint units

The problem of an administrative unit consisting of several polygons also has repercussions for other database operations. For example, it raises the question whether the complete data record of an administrative unit should be replicated for each polygon (in the polygon attribute table or after a relational operation). In this case, certain operations like summing the total population for a set of administrative areas after interactive selection would result in double counting for units consisting of more than one polygon. A way around this problem is to introduce a database field that identifies the largest or most important of the polygons belonging to the administrative unit. This "flag" item takes on value *one* for only one polygon belonging to a district, and *zero* for all other (minor) polygons (Figure 5). Any operation that involves summing or averaging database fields can then be performed after reselecting all polygons with a flag value of one.

Figure 5:
Use of a flag item as demonstrated for Caio district in Western Guinea Bissau.



The use of duplicate records and an indicator for the major polygon is arguably the most compatible way to address the problem of disjoint administrative units. An alternative method is available on some GIS systems such as workstation Arc/Info (7.0 and higher; see GIS World 1995 for all product references) which allows for a data type called *regions*. Regions allow several separate or even overlapping polygons to use the same record in a separate data table. While this is a more elegant way of designing a complex database, this capability is so far restricted to specific GIS software and is thus not compatible across platforms or programs.

Example database structure for WALTPS data set.

A comprehensive GIS database has been constructed with data compiled by the OECD/Club du Sahel *West Africa Long Term Perspective Study* (WALTPS, see Snrech 1995, Ninnin 1994) as described in Brunner *et al.* (1995). This database contains population estimates for about 2200 administrative units in 19 West African countries, as well as rural production statistics for about 400 agricultural census regions. To allow broad access to these data, a very simple and generic database scheme was chosen and implemented as an *Arc/Info* database. A requirement was that the data should be easily transferable into another GIS or desktop mapping system such as *Mapinfo* or *Atlas GIS*.

The structure of the data table associated with the administrative boundaries is displayed below. Each polygon in the digital map is linked to one record in the database. The first four variables in the data set contain system generated area, perimeter and identifier information. The first two of these are of limited use, since the database is referenced in latitude/longitude coordinates (the most generic of cartographic reference systems) and the units of measurement are therefore decimal degrees.

Following are areal figures calculated by temporarily projecting the GIS database into an equal area projection. SQKM is the area in square kilometers of the particular polygon. ADMSQKM is the sum of the areas of all polygons that belong to the particular administrative unit. Similarly, AGSQKM is the total area of all polygons belonging to the same agricultural census unit. IWSQKM is the area of any waterbodies that may cover the administrative unit and has been derived by temporarily overlaying a data set of inland waterbodies. This allows a user to subtract an administrative unit's surface area that is covered by a lake and thus to derive more accurate density estimates.

In addition to the administrative identifier, ADMINID, which is constructed as explained in the text, a character field providing the type of polygon (land or island), CODE, and the official three letter country abbreviation are also included. For the West African database, the use of two levels in the administrative hierarchy was seen as sufficient; in some cases, a higher level administrative level was ignored (e.g., the regions in Côte d'Ivoire). Following is the flag item used to identify the major polygon of an administrative unit - here termed DEMOFLAG because in the database an AGFLAG identifying the major polygon for each agricultural census unit exists also. Finally, the subsequent fields contain the actual demographic data, in this case total, urban and rural population estimates for four years which had been estimated by the Club du Sahel, as well as agricultural production statistics.

```
DATAFILE NAME: WAF.PAT

 COL  ITEM NAME      WDTH OPUT TYP N.DEC  DEFINITION
    1  AREA             4   12  F    3  system generated
    5  PERIMETER        4   12  F    3  system generated
    9  WAF#             4    5  B    -  system generated
   13  WAF-ID           4    5  B    -  system generated
   17  SQKM             4    8  F    1area of polygon in sqkm
   21  ADMSQKM          4    8  F    1area of admin unit in sqkm
   25  AGSQKM           4    8  F    1area of agric. unit in sqkm
   29  IWSQKM           4    8  F    1area of major inland water
                                        bodies located in admin
                                        unit (sqkm) - mostly 0.0
   33  CODE             3    3  C    -L-Land, IS-Island
   36  ADMINID          4   11  B    -unique id for admin unit
   40  COUNTRY          3    3  C    -three letter country code
   43  NAME1           25   25  C    -name of 1st level unit
   68  NAME2           25   25  C    -name of 2nd level unit
   93  DEMOFLAG         2    2  I    -flag for each admin unit
   95  P60              4    5  B    -estimated total pop 1960
   97  U60              4    5  B    -estimated urban pop 1960
  103  R60              4    5  B    -estimated rural pop 1960
  …    …                             -same for 1970, 1980 and 1990
  …    …                             - agricultural production data
```

4.1.4. Metadata

For an international database, summary information that is country specific can be kept in a separate database table or a spreadsheet that also serves as documentation. In a relational database, this country-specific information can easily be linked to the GIS coverage's attribute table via the country code. Useful information includes the full name of the country, summary statistics such as total population and growth rates, or the types of administrative units at each level (e.g., district, arrondissement, municipio).

More importantly, any supporting information available for the data set needs to be reported. This metadata should include information about the projection and associated parameters, data sources including source map scale, accuracy information, and any processing steps that have been performed on the data set (such as overlays, replacement of data components, generalization, etc.). Prior processing steps and input data are often jointly termed *lineage information*. Unfortunately, such information is often absent when data are distributed via the Internet or other channels that facilitate informal data exchange thus rendering potentially useful data sets more or less useless.

4.1.5. Further considerations

Unless the number of data items that are compiled is very small, all demographic and socioeconomic indicators are best stored in separate data files that can be linked to the GIS coverage by means of the unique administrative unit identifier (`ADMINID`). Separate databases can be maintained for different sets of variables such as time series information, age distribution at different censuses, or agricultural census information in separate files. It will make sense to store only data for the most disaggregated administrative level, since figures at higher aggregation levels can be compiled when needed. A separate database in a generic database management system may in the ideal case represent a small-area accounting system (e.g., Rees 1994). Such a system allows for great flexibility in performing consistency checks (e.g., accounting constraints), in constructing demographic cross-tabulations (i.e., those that do not involve geographic overlays), or in producing population projections.

4.2. Spatio-Temporal Databases

Temporal dynamics and historical patterns of demographic phenomena are of great interest in explaining contemporary land use, settlement or agricultural patterns. Research in agricultural economics, for example, is often concerned with long-term transformations of farming systems as suggested by population-driven intensification hypotheses. In order to study such processes and incorporate information about demographic dynamics into models of agricultural change, it is necessary to compile time series of population-related variables. Such spatio-temporal databases ideally describe the geographic distribution of people and their characteristics over time at fairly high temporal resolution. The time horizon can be rather long, as for example in the study by English *et al.* (1994) who studied land resource management in the face of rapid population increase in the Machakos region of Kenya between 1930 and 1990. The collection of papers in Fetter (1990) discuss methodological issues in the investigation of demographic processes in colonial Central Africa. On the other hand, recent patterns are often the most important in explaining current conditions and in making educated guesses about likely future trends. An example is the WALTPS project discussed earlier which used information from 1960-90 - a period in which most African countries had at least two or three censuses - to develop scenarios about demographic and economic developments over the next 30 years.

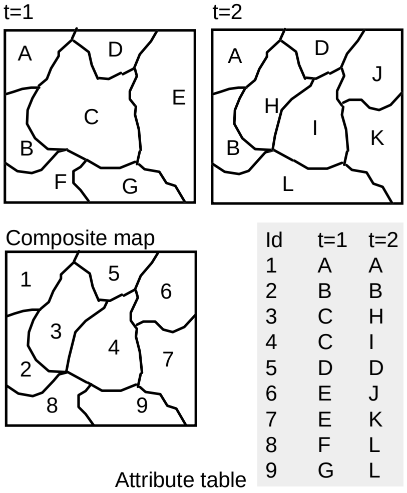
The compilation of time series information, however, is not always straightforward (see the comprehensive review by Langran 1992). Demographic and socioeconomic data are usually collected for relatively arbitrary political units. Censuses are typically conducted every ten years, and in some countries registration systems allow for the compilation of population statistics at any given point in time. The boundaries of the administrative units for which data are aggregated and published, however, often change over time. The lower the aggregation level, the more frequent and widespread are the changes (e.g., district and enumeration area boundaries change more often than provinces). Generally, modifications of administrative boundaries are aimed at maintaining a fairly homogenous distribution of population by areal unit - i.e., to keep the number of people per political unit relatively constant. In some cases, however, purely political reasons appear to be responsible for boundary changes.

There are three strategies to deal with changes in administrative 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, or
- integrating information about the complete time series in the data base.

In the *first* of these, the administrative boundaries for each census are stored in a separate GIS data layer. This approach was taken, for instance, for the *Indiamap* product (Klein 1994) which covers the two most recent censuses in India at the district level. 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 example, significant additional manipulation would thus be necessary.

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. This was done by the Club du Sahel for their work in West Africa (Snrech 1995). In cases where smaller districts for a previous census were merged to form a new larger district, 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 is used. This is similar to problems of estimating, for example, population data for physically defined areal units such as agro-ecological zones. Areal interpolation methods will be discussed in detail in Section 4.



Id	t=1	t=2
1	A	A
2	B	B
3	C	H
4	C	I
5	D	D
6	E	J
7	E	K
8	F	L
9	G	L

Figure 6:
Simple spatio-temporal database.

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 6). An ambitious study that implements this strategy is the China in Time and Space (CITAS) project which aims at developing time series of demographic and other data at the county level for China covering several decades (Lavely 1994, Tang 1994). This project employs a system in which 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* - i.e., an overlay of all county 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 particular county. 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.

This data model maintains a log of the boundary changes over time. However, this approach - as employed by the CITAS project - does not solve the problem of creating consistent data time series. Since the data tables for each census are in separate, often incompatible data tables, some form of areal interpolation is still required to compare data over time. Also, in many applications the historical boundary information may not be of interest, such that a compact representation of temporal dynamics may be preferred over a more complex data model. For historical and political analysis, or in the development of inventory-type applications such as cadastrial systems the maintenance of such a system is, however, certainly desirable.

5. Modeling

In the following sections of this paper several aspects concerning the modeling of population distribution will be discussed. The focus is on developing countries which are usually relatively data-poor in contrast to many industrialized nations where high resolution, up-to-date information make data enhancement efforts a less pressing issue. National censuses generate information that are used in planning and policy-making by collecting data from the lowest level up. For example, in a census, data collected by enumerators on individual households is aggregated first to the enumeration area or block level, and subsequently further to the district or state level. The most detailed information is generally not released. At the individual level, data privacy issues are the obvious reasons, and the data volume involved for large areas also prohibits use of high resolution information in many cases. Thus, only data at higher aggregation levels is typically available. The task for a researcher who needs detailed information is thus to reverse the aggregation process so tediously accomplished by the census agency. In doing so, various strategies can be chosen depending on the study’s objectives and available data.

The following section describes approaches to model population distribution within areal units. Firstly, direct interpolation from one set of areal units to another will be discussed, while the rest of the section deals with generating population density surfaces on a raster grid based on population totals available for areal units. Following is a discussion of techniques for modeling the size and shape of urban areas. Population figures are, of course, only the most basic of socioeconomic indicators.

Additional demographic variables are typically available from censuses, while more comprehensive characteristics, like economic status or poverty indicators, are more often derived from special purpose surveys. The final section will thus briefly describe approaches to the construction of small area estimates of socioeconomic variables based on census and survey data.

5.1. Modeling Population Distribution

In developing countries, demographic and socioeconomic data are often available only at fairly aggregate levels such as by district, *arrondissement* or *municipio*. For example, in a recent medium resolution continental database for Africa (Deichmann 1994a), the average resolution of administrative units was 123 km with a mean population per unit of 485,000. While these continent-wide averages hide large variation in national means (15 - 479 km resolution, and 41,000 - 3.1m people per unit), it is clear that the level of available detail is often not satisfactory.

Approaches to the problem of estimating population distribution within administrative units can be divided into two categories. In the first case, population totals are directly estimated for an alternative, non-nesting set of areal units. For example, in cases where boundaries changed between censuses, **areal interpolation** can be used to create time series of population parameters. In the second case, the distribution within the administrative unit is modeled explicitly by creating a more or less continuous surface of population density which is represented by a fairly high resolution raster grid. **Surface modeling** can also be used as an intermediate step in addressing the problem of incompatible zonal systems. It can thus be seen as a special case of areal interpolation.

5.1.1. Areal interpolation

Areal interpolation is the process of transferring data - e.g., population totals - from one set of areal units to another, incompatible set of units. Extensive descriptions of relevant approaches are given by Flowerdew *et al.* (1991), Goodchild *et al.* (1993) and most recently by 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*. For example, in Figure 7, census data are available for the 75 districts in Nepal (i.e., the source zones), while estimates of total population are required for the five major physiographic divisions (i.e., the target zones). Depending on the assumptions regarding the homogeneity of population distribution in either source or target zones, different areal interpolation methods are most appropriate. Additionally, auxiliary information can be incorporated in the interpolation process if population densities can be assumed constant in a third, auxiliary set of zones, termed *control zones*. These three cases are now considered in turn.

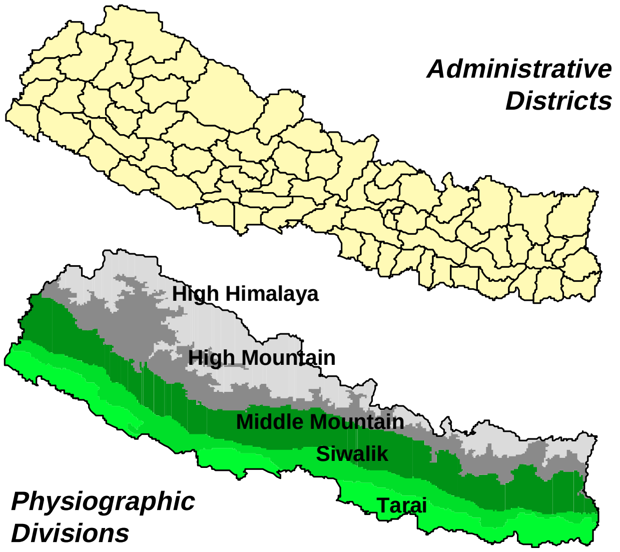


Figure 7: Areal interpolation example - Nepal. Population data are available by administrative districts, but are required for physiographic divisions. Boundary data sources: Geographic Decisions International, Topographic Survey of Nepal.

5.1.1.1. Source zones homogeneous

The simplest case is where we can reasonably assume that the source zones have a relatively homogenous population distribution; that is, population densities are fairly constant within each source zone. This may be a reasonable assumption when enumeration areas are designed specifically to represent homogenous conditions. In that case, total population for the target zones can be derived by overlaying the target and source zones and allocating source zone population to target zones in proportion to the areas of overlap. If the areas of overlap between source zone *s* and target zone *t* are denoted A_{st}, and the population of source zone *s* is P_s, the target zone population, P_t, can be estimated as:

$$P_t = \sum_s P_s \left(A_{st} / \sum_t A_{st} \right)$$

This approach is usually referred to as *areal weighting*. In practice, the easiest way of implementing this approach is by computing source zone densities (P_s / A_s) first and saving the results in a new data field. After overlaying target zones and re-calculating the correct surface areas for each new polygon, the total population for each area of overlap is derived by multiplying its area with the source zone density field. The resulting totals can then easily be aggregated for target zones.

Obviously, in most cases, densities vary within areal units. 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 areas or dense forests might be available from additional GIS layers which can be used to improve the target zone population estimates significantly. As mentioned earlier, in cartography this technique is called *dasymetric* mapping and involves, for example, the masking of irrelevant areas for the representation of densities using choropleth maps (e.g., Plane and Rogerson 1994).

5.1.1.2.Target zones homogeneous

Conversely, one might assume that the target zones show a relatively uniform density. This may be reasonable when the target zones are land use or land cover areas, where different land use classes are likely to have a fairly uniform population distribution. If the number of target zones is smaller than the number of source zones ($n_t < n_s$), the source zone population, P_s, can be represented as the sum of the target zone densities, d_t, multiplied by the area of overlap between source and target zones, A_{st}:

$$P_s = \sum_t d_t \cdot A_{st}$$

If we add an error term, this representation is identical to a linear regression forced through the origin (i.e., without an intercept term) of the form $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$. We can therefore estimate target zone densities as the coefficients in a linear regression using, for example, ordinary least squares estimation. 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. After estimating target zone densities, the total target zone population can be derived by multiplication of densities by the corresponding areas of overlap and subsequent aggregation as before.

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 a GIS. If extremely low densities are present in the study area, however, it may happen that negative regression coefficients result, implying negative population densities. Goodchild *et al.* (1993) suggest several special estimation techniques that can be used to enforce positive coefficients. Alternatively, the densities of the specific target zones can be set to zero or to some other externally estimated value (i.e., using a constrained estimation).

5.1.1.3. Control zones homogenous

Finally, in cases where neither source nor target zone densities can be assumed homogenous, auxiliary available information can be incorporated to improve target zone estimates. Such auxiliary information may consist of a digitized land use map for the study region (Moxey and Allanson 1994), a classified remote sensing image (Langford *et al.* 1991), or even a heuristically produced coverage in which an analyst familiar with the study area indicates polygons that are assumed to have homogenous densities (Goodchild *et al.* 1993). These control zones do not have to match either source or target zones, which allows great flexibility in choosing a suitable data set.

Provided that the number of control zones is smaller than the number of source zones ($n_c < n_s$), the control zone densities can be estimated using linear regression through the origin as before:

$$P_s = \sum_c d_c \cdot A_{sc} \quad ,$$

where d_C is the density of control zone *c* and A_{SC} is the area of overlap between souce zone *s* and control zone *c*. Using the estimated densities, target zone populations can be estimated by using the areas of overlap between control zones and target zones as

$$P_t = \sum_c d_c \cdot A_{ct} \quad .$$

In this case, the number of target zones is not restricted by the number of source zones.

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 mentioned, 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.

5.1.2. Surface modeling

The second set of approaches to population distribution modeling is aimed at producing a regular raster grid in which each cell contains an estimate of total population or population density that is representative for that particular location. This involves the disaggregation of total population that is recorded for administrative units by distributing people over the grid cells that fall into that unit. Representing socioeconomic data in raster form has several advantages. Firstly, data recorded on a fairly high-resolution, regular grid can be easily re-aggregated to any areal arrangement required. Rasterizing socioeconomic data is thus also a solution to the areal interpolation problem.

Secondly, as Clarke and Rhind (1992) point out repeatedly, many environmental data sets are stored in raster format. Some information, such as remotely sensed images, is by technical necessity recorded on a regular raster. Others, for example climatic data or digital elevation data, are usually interpolated onto regular raster grids, although they may subsequently be converted to linear or triangular features such as contour lines or triangulated irregular networks (TINs). Given the increasing interest in the integrated analysis of environmental and socioeconomic data, producing population data in gridded form is one way of ensuring compatibility between heterogeneous data sets.

Finally, as Bracken and Martin (1989) point out, compiling population data for a set of relatively arbitrary areal units creates a number of problems. For instance, only a small portion of a census division may actually be populated. Or census boundaries may split regions of homogenous density. Areal aggregates thus hide significant information about the underlying distribution of the variable of interest. Converting socioeconomic data into raster form could thus provide a way of avoiding some of the problems imposed by artificial political boundaries.

Compiling population data in raster form is by no means a new approach. Several countries have been publishing gridded demographic data for decades (e.g., Japan and Sweden). Adams (1968) presents a computer generated raster map of population densities in West Africa, a gridded population map for Southern Africa has been published by the Africa Institute (1965) in South Africa, and fairly high-resolution gridded population data for several regions in China are presented in the Population Atlas of China (Population Census Office 1987). Unfortunately, in most cases the methodology for generating the gridded data is not reported.

In this section four approaches will be reviewed which have been applied to create gridded population data: the distribution of population based on census tract centroids, maximally smooth interpolation, so-called "smart interpolation” which employs auxiliary data sets to guide the distribution of population, and a cartographic approach.

5.1.2.1. Centroid based distribution

In some instances, the complete census geography - i.e., population data registered by polygons - is not available, or the data volume associated with thousands of polygons would be prohibitively large. Instead, some agencies distribute census data recorded for latitude/longitude coordinate pairs which represent either geometric or population-weighted centroids. Examples are the US Census Bureau's Master Area Reference File (MARF) or the Small Area Statistics files for enumeration district centroids in Great Britain. One option is to construct Thiessen polygons first to create polygons around the centroid locations. Thiessen polygons are created geometrically around each point such that each location in that polygon is closer to the centroid from which it was created than to any other centroid. Or, in other words, each location in the study area is assigned to the closest centroid. In the absence of additional information, this is a reasonable assumption to guide the creation of polygons.

Alternative approaches have aimed at generating regular raster surfaces from centroid data directly. Under the assumption that the centroids are in some way representative for the population distribution (e.g., population-weighted), the total population for the census unit is distributed onto a regular raster grid that is draped over the study area. Bracken and Martin (1989; see also Martin 1991a, b) developed a method of using a moving window (kernel) function to distribute population to grid cells (see also Silverman 1986). The size of the window is determined by the local density of centroids, while the share of the population assigned to each grid cell within the window is based on a suitable distance decay function (i.e., closer pixels receive a larger share of population than those further away). This method allows for cells not to receive any population at all. So, while the distribution around each centroid is fairly smooth, the method does allow for discontinuities of population densities. Bracken and Martin applied this approach to a whole series of socioeconomic indicators (population density, employment rates, income) for various regions in Great Britain. Some of their raster images of population density resemble nighttime satellite images with scattered areas of high population density appearing in bright colors.

An alternative approach was used by Honeycutt and Wojcik (1990), who produced a population density map of the conterminous United States for use in route planning in the transport of hazardous materials. Their model uses a circular Gaussian function to distribute population to cells surrounding the centroid location. The range of the distribution of population was estimated using empirical data on urban size distribution for the United States. These urban density/urban size models will be discussed in a subsequent section. Clearly, the approach used by Honeycutt and Wojcik is most suitable for heavily urbanized areas which were the focus of their application. In rural regions, population distribution will typically be too sparse.

5.1.2.2. Maximally smooth interpolation

In contrast to the previously outlined methods, Tobler's (1979) *pycnophylactic* (i.e., mass-preserving) interpolation takes population totals by areal units as the basis for creating a smooth raster surface of population density. Tobler suggests the following way to visualize the pycnophylactic approach: total population for a set of administrative units can be conceived as a three-dimensional histogram made of clay placed on top of the corresponding unit. Clay is subsequently moved within each administrative unit so as to make the surface of the entire study area as smooth as possible. The pycnophylactic constraint, however, implies that clay cannot be moved from one areal unit to another; that means, total population within the administrative unit has to remain constant. Thus, clay will be moved from areas which border polygons with lower population densities to those areas that are located close to administrative units with higher densities. Figure 8 illustrates this process. The rationale behind this method is that population tends to be similar in nearby locations, since people tend to live close together. Thus, population densities are likely to be higher in areas neighboring high density regions than in areas that are close to administrative units that have lower densities.

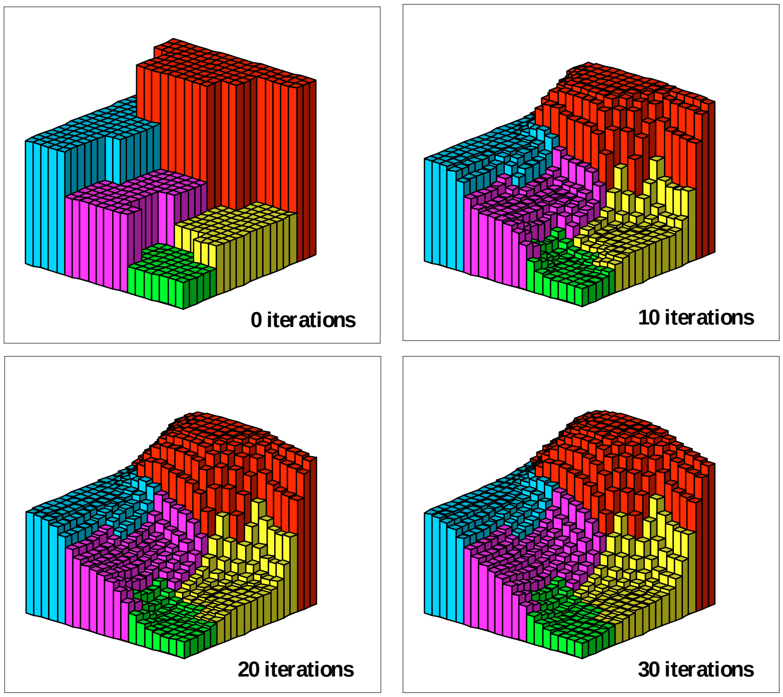


Figure 8: Pycnophylactic interpolation on a raster grid

Translated into a computer algorithm, the interpolation starts with draping a relatively high resolution, regular lattice or grid on top of the study area. Each grid cell falling into a particular administrative unit is initially assigned the same proportion of the total district population. Next, a suitably chosen moving filter operates on the grid which replaces each grid value with a weighted average of its neighboring values (e.g., using the four or nine nearest neighbors). This process is repeated iteratively. After each iteration, the pycnophylactic constraint is enforced by adjusting the new total population - which has been changed in the smoothing - to match the initial total. Since the filter will operate across district boundaries, the cell values will be modified from the district boundaries inwards. The process stops when any further adjustments would be smaller than a specified tolerance level and the surface is thus very smooth. Tobler's original paper (1979) contains the mathematical exposition of this technique.

Pycnophylactic interpolation has been used quite frequently for rasterizing polygon-based data and to address the areal interpolation problem (e.g., Stetzer 1990). The NCGIA Global Demography project (Tobler *et al.* 1995) used this method to create global raster images of population distribution from a set of about 15,000 administrative units with associated population totals. Generally, pycnophylactic interpolation is an appropriate technique if no additional information is available, and if the units are relatively homogenous. Steep gradients between, for example, highly urbanized areas and surrounding rural regions may lead to artifacts in the resulting output raster due to the enforcement of maximal smoothness in the population surface. A promising extension of the method could thus focus on the incorporation of additional information about population distribution that could be used to relax the maximal smoothness constraint where appropriate.

5.1.2.3. Smart interpolation

The proliferation of GIS technology has resulted in the increasing availability of standardized GIS data layers for many regions of the World. For example, the World Resources Institute has recently coordinated the compilation of the African Data Sampler, a CD-ROM that contains a basic set of GIS layers in a standardized GIS format for every country in Africa (WRI 1995). The data themes include roads, settlements, administrative units, rivers, elevation contours, protected areas, forest areas and wetlands. Some of this information is clearly relevant to predicting population densities, since people are not distributed randomly across space but prefer to settle in areas with certain characteristics. In analogy to somewhat similar interpolation methods used in creating surfaces of climatic variables - e.g., using elevation and exposure to interpolate temperature and precipitation fields - this approach will be called *smart interpolation* (Willmott and Matsuura 1995).

A heuristic approach to incorporating these types of information was suggested by Deichmann and Eklundh (1991) and implemented to create an African population density map for use in UNEP's global desertification atlas (UNEP 1992). The approach consists of three basic steps:

- First, a surface of weighting factors is created on a regular raster grid for the study area. These weights are based on the well-known concept of an interaction potential of population which is related to a whole family of gravity and spatial interaction models. The primary input to this model is the location and size of urban settlements.
- Secondly, using auxiliary data sources, the basic weights derived in the first step are heuristically adjusted. Areas which are known to be uninhabited or sparsely populated receive zero or very low weights. In contrast, the weights for grid cells which are close to physical features that typically attract settlements (such as roads or navigable rivers) are increased.
- Thirdly, total population recorded for the study area is distributed to the corresponding grid cells in proportion to the weights constructed in the previous steps. Thus, areas close to large settlements or to major transport infrastructure receive a proportionally larger share of population than areas further away, or areas located in wilderness areas, protected areas or otherwise uninviting regions.

The distributed population estimates for each grid cell can subsequently be converted into densities for mapping, or they can be reaggregated for alternative areal arrangements such as agroecological units or watershed areas. The three steps summarized above are now discussed in more detail in the following paragraphs.

Step 1: Calculation of Population Potential.

The primary information guiding the distribution of total population over the cells of a regular raster grid of the study area is a population potential surface. Population potential, or more specifically, interaction potential of population is a measure of average accessibility of a given location with respect to the size and location of other features - e.g., urban areas. In other words, it is a summary measure of the influence exerted by all features in a region - for instance, all settlements - upon the particular location. Note that *potential* here is derived from the physics interpretation of the term and is unrelated to population potential concepts developed in carrying capacity studies. Population potential measures were developed by the social physics school in the 1940s (see the summaries in Stewart and Warntz 1968, Plane and Rogerson 1994). In analogy to physical gravity models, the influence of, for example, a town upon a grid cell is assumed to be proportional to the town's size, weighted inversely by the distance of separation between the cell and the town. The inverse distance weighting implies that towns in the vicinity of the cell will yield proportionally more influence than those further away. The population potential is then an aggregate measure based on all towns in the area or within a given threshold distance:

$$V_{ij} = \sum_{k=1}^{n} \frac{P_k}{d^b_{(ij)k}} \qquad ,$$

where V_{ij} is the population potential for cell *ij*, P_k is the population of town *k*, and $d^b_{(ij)k}$ is the distance of separation between cell *ij* and town *k* weighted by an appropriate exponent *b* (see Figure 9).

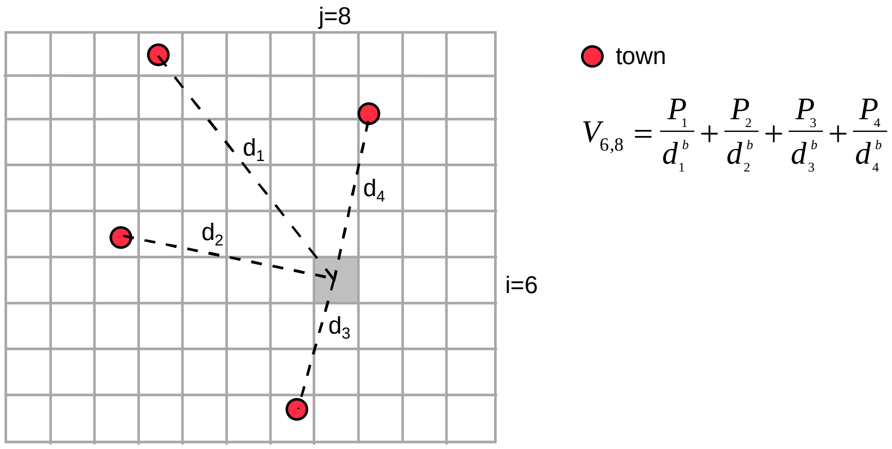

j=8
i=6
d1
d2
d3
d4
● town
$V_{6,8} = \frac{P_1}{d_1^b} + \frac{P_2}{d_2^b} + \frac{P_3}{d_3^b} + \frac{P_4}{d_4^b}$

Figure 9: Illustration of the calculation of the interaction potential of population.

In practice, for various reasons reviewed by Bröcker (1989), a better model of distance decay is a negative exponential function rather than simply weighted inverse distance, such that:

$$V_{ij} = \sum_{k=1}^{n} P_k \cdot e^{\left(-d^2_{(ij)k}/2\alpha^2\right)} \qquad ,$$

where α is the distance to the point of inflection in the distance decay function. To illustrate the difference in the choice of distance decay functions, Figure 10 shows inverse functions with exponents b=2 and b=0.5 as well as a negative exponential function with α=10.

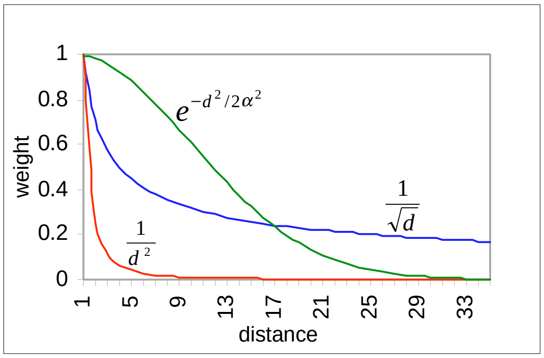


Figure 10: Comparison of different distance decay functions.

If the potential value is calculated for each cell in a regular raster grid using the straight line distance between the center of the cell and each of the towns, a relatively smooth surface of weights results. The significance of this surface in predicting population distribution lies in the tendency of people to live fairly closely together. Larger settlements or areas where several towns are located in close proximity are also likely to be surrounded by a larger number of smaller settlements. Thus population densities around major urban centers or agglomerations are likely to be higher than densities in more isolated areas.

A problem with the distance calculation is that straight-line distances may not be appropriate in areas where the transport network is not very dense. If a good database of roads, railroads and navigable waterways is available, this information can be used to create a more realistic potential surface. Instead of using the straight-line distance between a grid cell and each town, the actual distance along the network can be used instead (Geertman and van Eck 1995). If road quality information is also available, distances along road segments can be converted into travel times using average possible speed as a guideline. For example, on a dirt road, travel may only be possible at 40 km/h, while on a good tarmac road an average speed of 100 km/h may be realistic. Additionally, border crossings or other known obstacles can be explicitly incorporated by increasing the travel time by a suitable amount. The cumulative travel time may thus be a better weight in the population potential calculation than distance measures.

The most complex part in using the transport network explicitly is to find for each cell/town combination the shortest route through the transport network. While it is straightforward to compute straight-line Euclidean or great-circle distances, finding the fastest connection in a complex network is a computationally more demanding task. A standard method that has found many applications in transport modeling and which is also implemented in some GIS systems is the Dijkstra algorithm (Dijkstra 1959). Finally, it should be noted that the accessibility measure, which is used here as the first step in modeling population distribution, is also a useful summary variable in its own right.

Step 2. Adjustments

The previous step creates a fairly smooth population potential surface that can serve as the basis for a weighting scheme to distribute total population. However, in reality population distribution usually displays a certain degree of discontinuity due to the existence of physical or man-made features that are either associated with higher densities or indicate the absence of population. For example, people do not typically reside in protected areas and other wilderness areas, in very high mountain regions, and - most trivially - within waterbodies (the lagoon villages on stilts in Benin notwithstanding). Spatially referenced information about these areas is often available, and by overlaying a rasterized mask delimiting these features over the population potential surface, the weights in these uninhabited areas can be set to zero - or to a very low value if a small number of people is thought to live, for example, in mountain regions, dense forests or wilderness areas.

If the transport network has not been used explicitly in the population potential calculation, the proximity to roads, railroads or navigable waterways can, vice versa, justify the increase of the weights derived in step one. People tend to live where accessibility is good, and, conversely, roads tend to be built where people live. A good example, as pointed out by Noin (1979) is the role of the railroad in determining population distribution in Zambia. Again, a gradual distance decay from these major features using an appropriate inverse or negative exponential function can be used to derive an adjustment factor.

The appropriate magnitude of these adjustments is of course difficult to assess. Ideally, the adjustment parameters would be estimated statistically for each country using high resolution data on observed population distribution. However, such data are usually unavailable - making the modeling of population distribution necessary in the first place. Typically, the adjustment factors thus have to be chosen heuristically. An interesting approach to come up with a suitable set of weights was implemented by Sweitzer and Langaas (1994). In this study, adjustment factors chosen by a number of experts familiar with the study area, the Baltic states, were used in a multicriteria evaluation scheme (Eastman *et al.* 1993). Auxiliary variables were land cover, transport infrastructure and proximity to urban areas. In contrast, Deichmann and Eklundh (1991) and Veldhuizen *et al.* (1994) used "rules of thumb" and assessments of published population distribution maps to determine a set of suitable weights heuristically.

A good predictor of population densities used by both Sweitzer and Langaas (1994) and Veldhuizen *et al.* (1994) is land cover and land use. Particularly in studies covering relatively small areas, such data are often available from satellite images or extensive field surveys (see also Langford *et al.* 1991). For larger regions, however, no consistent land use or land cover data sets are available despite considerable efforts in the remote sensing community (see Meyer and Turner 1994).

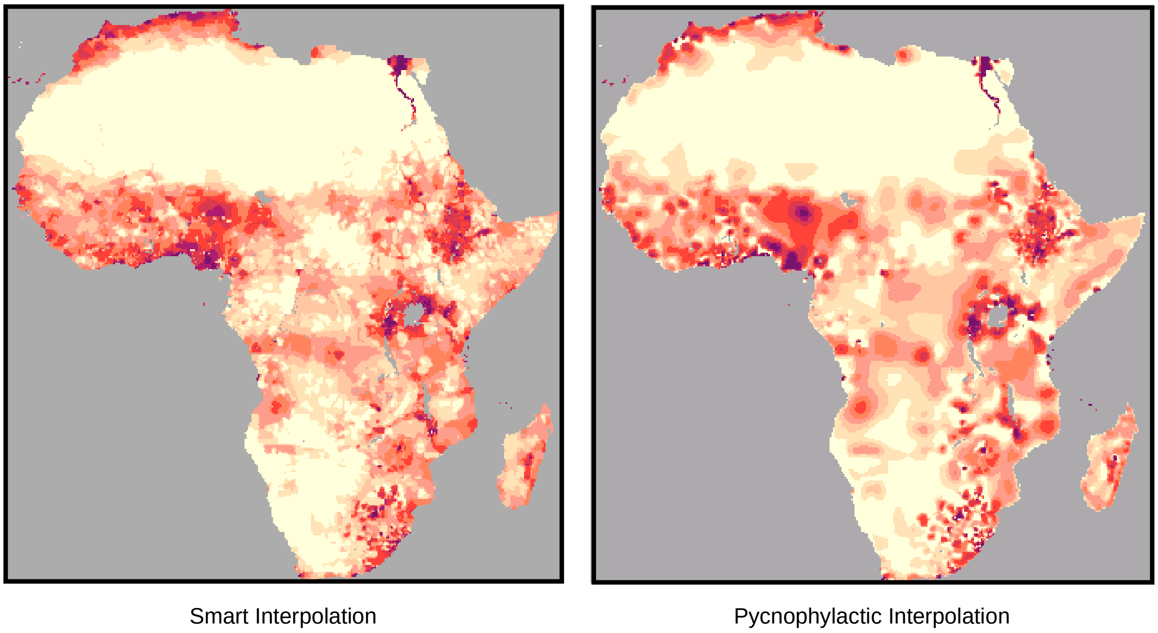
The process of deriving and heuristically adjusting the weights for the population distribution can be likened to a computer simulation of the thought process of a cartographer or geographer who, based on experience, external information and much intuition, draws boundaries delineating social or physical features on a map. This cartographic production process - although not quantifiable - is thus guided by all relevant observations and data available, but relies significantly on a theory and knowledge of the physical or social processes that are mapped.

The adjustment of the population potential is likely to proceed in an iterative fashion, where the resulting weights are evaluated and compared with distribution maps available for particular areas. In this way, information can be incorporated in an informal framework and combined with the "expert knowledge" of the analyst. The great advantage over traditional cartographic methods is that the procedure can be replicated, since the production process is implemented in a computer system.

Step 3. Distribution

Once a satisfactory raster surface of adjustment factors has been created, the weights within each administrative unit need to be standardized so they sum to one. Producing the population distribution map is then simply a matter of multiplying the total population recorded for the district by the standardized weight estimated for each of the grid cells in that unit.

Figure 11 compares a population density map for Africa which was generated using the approach just outlined with one that was derived using the pycnophylactic interpolation method. While both are based on the same set of population data by administrative units, the latter shows a much smoother distribution, especially in areas where the resolution of the census boundaries is fairly low (e.g., Central Africa, Nigeria).



Smart Interpolation

Pycnophylactic Interpolation

Figure 11: Comparison of two interpolation methods to derive estimated population densities for Africa from polygon data.

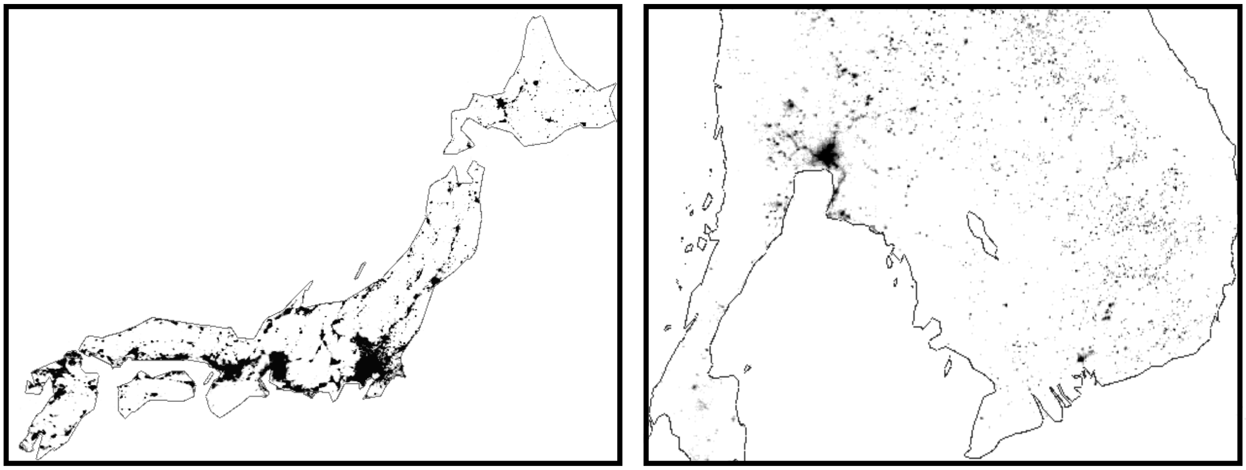
Modifications to the methodology

The approach outlined in the previous paragraphs could be modified in several ways. Climatic indicators, for example, have frequently been suggested as important determinants of population distribution. Although additional factors influence habitation-climate relationships -e.g., the importance of the Nile river in an otherwise hyper-arid region of North Africa -, climatic factors largely determine agricultural suitability which, in turn, has lead to early settlements which provided the nuclei for today's urban agglomerations. Likewise, the extremely large concentration of people in the Ganges plain and in the eastern parts of China are due to fertile soils which have historically sustained very high population densities.

On the other hand, population distribution maps for large areas are likely to be used extensively in the analysis of population and environment interactions (see United Nations 1994b, for a comprehensive literature review). Using climatic or similar agroecological indicators in the creation of population distribution, which in turn will be used for assessments of the effects of population dynamics on natural systems, creates a circularity which may lead to high correlation coefficients, but dubious validity. The same is true if population distribution maps incorporate land use/land cover information, and these data are subsequently used for the study of population/land use interactions. It will thus depend on the subsequent use of the data, whether it makes sense to use climatic and other environmental suitability measures in population distribution modeling. For general purpose databases, it may be preferable to rely on anthropogenic variables or observed land use/land cover patterns.

Improvement of the estimated surface of population distribution can also be achieved by incorporating information on the size and location of towns and cities explicitly. If reliable data on urban population totals are available, these can be distributed to the corresponding grid cells first such that only the - often much smaller - residual population needs to be allocated to the remaining grid cells in the administrative unit. The number of grid cells that receive the urban population can be determined using models of city size reviewed in a subsequent section. A problem that may be encountered, however, is that total population and city size estimates are often not from the same source. Definitions of what constitutes the city boundaries vary by country, and the urban area delimitation often does not coincide with administrative unit boundaries. Consequently, it is not unusual that negative numbers result when the accumulated urban population is subtracted from the total population of the corresponding administrative unit.

Finally, a data set that has been suggested as potentially useful for aiding the estimation of population distribution is based on the nighttime visible light emissions recorded by the Defense Meteorological Satellite Program (DMSP; see e.g., Foster 1983, Lo 1986). Figure 12 shows two sample images in reverse video mode. The challenge is to associate a given level of light emission with population density given the differences in energy utilization, economic wealth, and light sources that are not directly related to population activities such as gas flares.



Japan

South East Asia

Figure 12: Images derived from the visible band of the sensors of the Defense Meteorological Satellite Program (DMSP). Shown are nighttime visible light-emissions (in reverse video). The urban agglomerations in Japan as well as the Bangkok Metropolitan Area are clearly visible. Image data source: National Oceanographic and Atmospheric Administration, National Geophysical Data Center (NOAA/NGDC).

5.1.2.4.Cartographic approaches

A final example of a method to model population distribution at medium resolution levels is the *International Database* produced by the International Programs Center (formerly called Center for Internation Research or CIR) of the U.S. Bureau of the Census. This database has been reviewed in detail by Clarke and Rhind (1992) and will therefore be discussed only briefly here. For some time, CIR has produced digital maps of population distribution under contract for the Department of Defense (DoD). The data sets are developed country by country on the basis of official census statistics and published maps at different scales. The total population of a country is split into an urban and a rural component (see Figure 13). First, the population of all urban agglomerations of 25,000 or more people is distributed to a set of circles that are designed to represent built-up residential areas. The residual, "rural" population is then assigned to a regular, rectangular raster grid with a resolution of 20' by 30' ("bigcells"). For a number of countries (mostly NATO member states) a second grid with a resolution of 5' by 7.5' ("minicells") is also created.

Based on the initial census population, the population for each circle and grid cell is projected for the next 11 years using a simple ratio approach. The CIR national population forecasts that use the components of growth (births, death, migration) guide these projections. Leddy (1994) provides a detailed description of the method and examples using data for the United States and Spain.

A problem with the CIR small area population data in the past had been that the data were not generally available to researchers. Recently, however, many of the CIR country data sets have been released and are distributed by CIESIN. For a few countries, the data are quite out of date and are thus not based on the latest available census information. Also, the rectangular grid format chosen does not lend itself to manipulation in most raster or image processing packages. Despite these caveats, however, the CIR data represent a valuable source of population data. Hopefully, updated information, in particular for developing countries, will become available, and the cartographic techniques for distributing the population data will be better documented in terms of source maps and census publications.

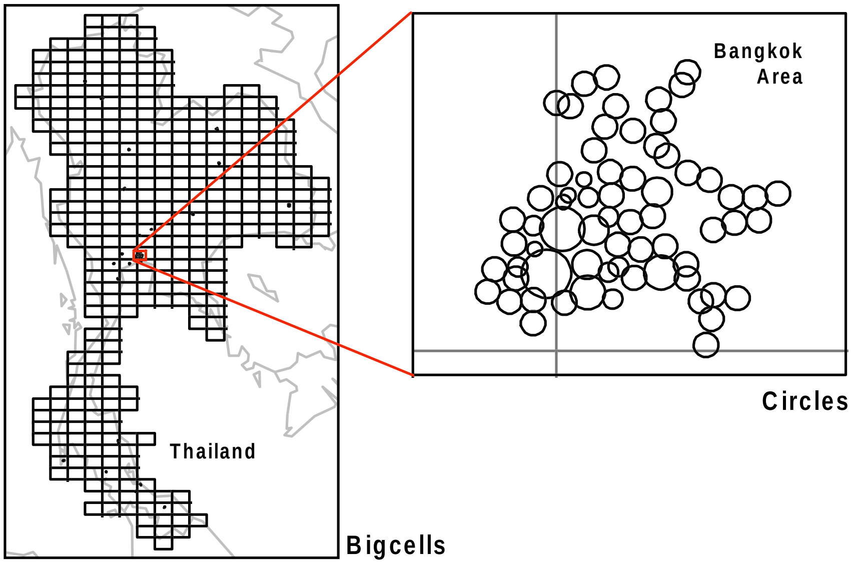


Figure 13: Population data by 20’ X 30’ grid cells (rural and total) and circles (urban areas). Source: International Programs Center, U.S. Bureau of the Census

5.1.3. Implementation Issues

Although commercial GIS packages are increasingly offering sophisticated modeling capabilities in raster as well as vector mode, more complex modeling schemes often require additional programming or the use of several complementary software systems. Almost all GIS systems allow the conversion of vector data into raster files. Once in raster format, data can be manipulated in a very flexible manner in *Arc/Info's GRID* system. *Idrisi*, which was used by Sweitzer and Langaas (1994) also has a large number of raster manipulation options. Finally, *Spans*, used by Veldhuizen *et al.* (1995), appears to be one of the few systems that has a built-in potential mapping module. Straight-line distances between grid cells and the closest of a specified set of features (points or lines) are calculated by most raster systems, while a shortest path algorithm is implemented in *Arc/Info's NETWORK* module as well as in some specialized transportation GIS packages.

A further issue pertains to the spatial referencing system used. Most generic databases such as DCW are referenced in geographic (latitude/longitude) coordinates. This is not an equal area system, which means that the cells in a regular raster grid referenced in geographic units will vary in size as a function of latitude. If the administrative units used are very large or extend narrowly in the North-South direction, the distortion can be significant. It may thus be necessary to adjust the weights using the actual area of the grid cell before applying the weights to the total population figures. Similarly, because of the Earth's curvature, calculating distances on a latitude/longitude grid or vector coverage using the simple Euclidean formula will not yield correct results. Instead, great circle distances need to be used. The formulae for calculating distances on the sphere and for converting between latitude/longitude degrees and metric measurements are given in most cartography books. An excellent reference is Snyder (1982).

5.2. Modeling Urban Areas

An increasing share of national populations are residing in urban settlements. While the process of urbanization has stabilized in most industrialized and some newly-industrialized countries, many developing nations continue to experience rapidly increasing urban populations. Furthermore, these increases, which are typically due to migration into the largest cities and higher urban fertility rates, are often highly unbalanced, with the consequence that a dominating core region is created that consists of a primate city and a surrounding cluster of settlements (MacKellar and Vining 1995).

In estimating the population distribution within a country, information about the location and size of urban settlements can greatly improve the results of modeling efforts. For example, as implemented in the CIR approach, the urban population of a district can be subtracted *a priori* before the residual population is distributed. Information about urban population is also of great significance for socioeconomic agricultural research, in that towns usually provide input and output markets and thus a stimulus for agricultural intensification. Similarly the urban network is important for health and educational facility planning and other forms of service provision.

As with regional population figures, however, information about the urban system in many developing countries is scarce. The location of towns can be derived from gazetteers or digitized maps. This is typically straightforward for larger settlements, except in cases where spelling is ambiguous or where source materials are outdated. For smaller settlements, a problem in many countries in Africa is that in slash-and-burn agriculture or nomadic cultures, settlements may actually “move” over time.

Information about the number of inhabitants in urban settlements is much harder to come by. In some countries, the village is the lowest administrative level, and figures by settlement are thus compiled for each census. In other countries, however, only regional urban totals are published - i.e., the total population of all towns in a district - while specific urban figures are compiled for only the largest cities. Indirect methods of estimating urban population figures are therefore required.

As was briefly mentioned earlier, even where urban population totals are available by district, a significant problem in cross-country comparisons is that - as was discussed in a previous section - the definition of *urban* varies widely by country. A comprehensive list of national definitions is given, for example, in United Nations (1993a). In some countries, a settlement is considered urban if its population is above 20,000 (e.g., Nigeria), while in others the threshold is set at 2000 (e.g., Gabon). Uganda goes much further in that all settlements and trading posts with a population of more than 100 inhabitants are considered urban. Yet others use a functional definition, for example by including all settlements whose inhabitants are predominantly active in the non-agricultural sector or all towns in which an administrative center is located (e.g., Pakistan), or they limit the definition of urban to the capital only (Burundi). Inter-country comparisons based on such figures are thus of little use. Instead, a consistent approach needs to be taken, where, for example, the number and total population of all towns greater than a standard population size are determined for each district. In cases where such data are unavailable, indirect estimation based on empirical regularities can provide a viable alternative.

5.2.1. Rank-size rule

Arguably one of the best studied aspects of urban systems is the so-called *rank-size rule* (e.g., Carroll 1982). This empirically observed relationship links the population size of a city to its rank in the national urban hierarchy. Mathematically, the rank-size rule states that

$$p_i r_i^q = k \quad ,$$

where r_i is the rank and p_i is the population of the *i*th city, and q and k are constant parameters. This relationship can be transformed into a linear, estimable relationship:

$$\log\ p_i = \log k - q \log r_i \quad .$$

The general validity of this relationship is easily confirmed using census data of population by city. Table 2, for example, presents summary information from the estimation of the rank-size relationship for a number of Asian countries. The urban population figures are taken from the respective national censuses and were compiled to create a spatially referenced Asian cities database. The available sample sizes vary significantly, so cross-country comparisons of these illustrative results should proceed with caution. By and large, however, the data confirm the rank-size relationship with R^2 values of 0.9 and higher. The estimated parameter *k* is essentially the estimated population of the largest city (e.g., the value at which the regression line crosses the y-axis of the logarithmic scale. For comparison the actual population of the largest city is also shown. This estimate is not reliable in cases where primacy is very high as exemplified by Thailand. The poor fit of the estimated model for this country as indicated by low R^2 is no doubt exacerbated by the small sample size.

Table 2: Rank-size relationships for selected Asian countries

Country	Sample size	Census Year	Intercept *k* (millions)	Actual primate city pop	Slope *q*	R^2	Est. num. of towns > 10k inh.
Bangladesh	11	91	3.08	3.60	-1.43	0.90	54
India	212	81	8.89	8.23	-0.84	1.00	3250
Indonesia	13	80	7.28	8.24	-1.12	0.96	361
Iran	38	86	3.80	6.00	-1.00	0.98	375
Israel	43	90	0.71	0.51	-0.91	0.97	106
Japan	77	91	5.13	8.02	-0.70	0.97	7076
Pakistan	120	81	3.69	5.84	-1.05	0.99	273
Philippines[a]	59	90	2.55	1.67	-0.93	0.93	376
South Korea	37	90	6.65	10.63	-1.15	0.94	280
Thailand	10	90	1.89	5.90	-1.32	0.77	52
Turkey	31	90	5.44	6.61	-1.12	0.98	278

a - using separate population figures for municipalities within Metro-Manila

The slope parameter, *q*, is an indicator of primacy. Countries with a dominant city that is many times larger than the second largest city display a steeper slope as compared to countries with a more even urban hierarchy. This can be seen graphically in Figure 14 which compares the rank-size relationships of Pakistan and Japan.

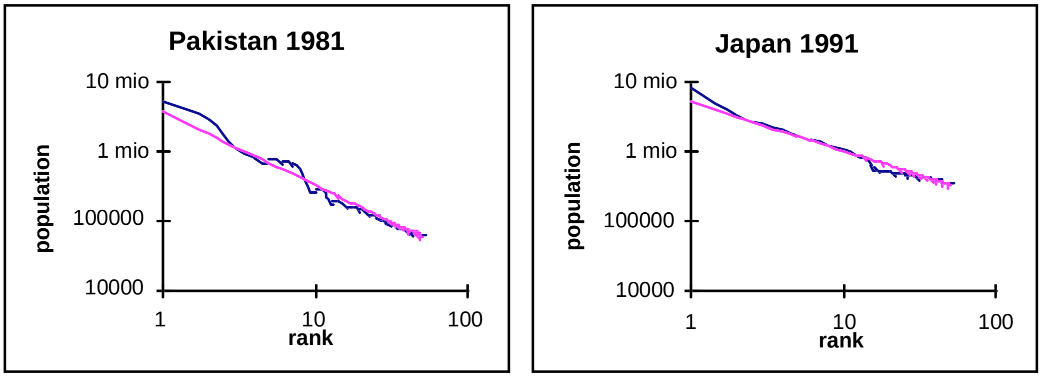


Figure 14: Rank-size relationships for Pakistan and Japan.
Shown are the fifty largest cities only. Straight lines indicate estimated relationship.

Using the estimated relationships, it is straightforward to solve for a particular population size in order to find the rank of a city of that size, and thus the number of towns in the country that have a population that is larger or equal to that figure. The last column in Table 2 gives the estimated number of towns in the country that have more than 10,000 inhabitants. In the absence of sufficiently detailed census data, this is a common way of estimating national urban and rural population figures that are standardized according to a consistent threshold value. For subnational population modeling, however, this approach is of limited use. The rank-size relationship is unlikely to hold for subnational administrative units, and an estimate of the number of towns larger than a given threshold level will not yield any information regarding where these towns are located. Only with additional information would it be possible to assign population figures to towns of a given rank within the districts.

5.2.2. Indirect estimation of urban/rural population

An interesting way of estimating urban population figures for subnational units was implemented for the West African Long Term Perspective Study (WALTPS; see Ninnin 1994, Snrech 1995). Here, rural population for a given district is estimated using the average of the rural population densities of the neighboring districts multiplied by the total area of the administrative unit. The published total population minus the estimated rural population is an estimate of urban population which is then allocated to the major towns within that district. These towns were chosen using published maps.

This procedure was assumed to produce more reliable figures than those derived from published sources, where inconsistent definitions were thought to make cross-national comparisons impossible. The approach could likely be improved further. Firstly, the rural population of the neighboring administrative units will likely have to be estimated as well. That means if the density of unit *A* is estimated using the densities of units *B*, *C*, and *D*, then a new estimate for *A* will change the estimates for

B, *C*, and *D*. The estimation will thus have to proceed in an iterative manner until the adjustments made are very small. Secondly, in cases where the administrative units are relatively small with respect to the size of the towns in that unit, the total area of that district might have to be divided into a rural and an urban part. Only the rural share would then be used to estimate rural densities and rural population. The urban area of the towns in the district could be estimated using empirical relationships relating a city's population with it's surface area as described in the following section. Again, this estimation would need to proceed in an iterative way.

5.2.3. City area

The relationship between a city's population and its surface area has been the subject of intensive study for much of this century. In the simplest form, this relationship is stated as

$$A = aP^b \quad ,$$

where *A* is the area of the city and *P* is its population. The parameters *a* and *b* can be estimated by transforming the relationship into linear form using logarithms as

$$\ln A = \ln a + b \ln P \quad .$$

This allometric growth function (see Nordbeck 1965) has been estimated for a number of countries at different time periods. However, with few exceptions reliable areal estimates that can be used to determine the parameters are available for industrialized countries only. Tobler (1969) used a Gemini V photograph to estimate the relationship between city land area and population in the Nile delta. Lo (1986) reviewed several studies that use other types of remotely sensed images including night-time images from the Defense Meteorological Satellite Program (DMSP) mentioned before. In general, one would assume that cultural and economic factors determine how densely people live within a city. In countries where mobility (car ownership and public transport) is limited, one would expect to find smaller city sizes for a given population compared to countries that have experienced significant suburbanization.

5.2.4. Models of urban shape and dynamics

In addition to the areal extent of a city, the distribution of people within a city can be approximated using empirical relationships. Assuming a circular city shape, urban population densities are typically modeled by a negative exponential function of the form $D = Ae^{-br}$, where *D* is the density and *r* is the distance from the city center (Clark 1951, McDonald 1989, Keersmaker 1990). Many other models have been suggested (e.g., as reviewed in Tobler *et al.* 1995) including specifications for polycentric cities. Alternative models attempt to estimate bidimensional density functions based on scattered observations as is common in statistical applications (Silverman 1986).

If urban areas are treated explicitly in population distribution modeling at higher resolution, levels, it may be appropriate to relax the assumptions of the circular city model assumed by the allometric growth model. Typically the shape of an urban area is modified by physical features: natural factors such as coastlines, rivers or mountains, as well as man-made features such as major transport routes. In a city situated at a river or at the coast, properties near the water may be at a premium and the built-up area can thus be expected to extend more along the water and less away from it. In an interesting book, Diakonis (1968), in outlining what he calls the science of *ekistics*, discusses the shape of urban settlements with respect to historical, economic and physical factors as well as across scales. With various degrees of abstraction, he shows how cities grow along major transport routes or other physical features (see also Tobler 1970). A real-world example is presented by Kirtland *et al.* (1994). Figure 15 shows three images from a continuous animation of urban growth in the San Francisco/Sacramento area, which are based on historical maps and contemporary satellite data. The linear growth pattern in the Silicon Valley along the San Francisco-San Jose axis is clearly visible.

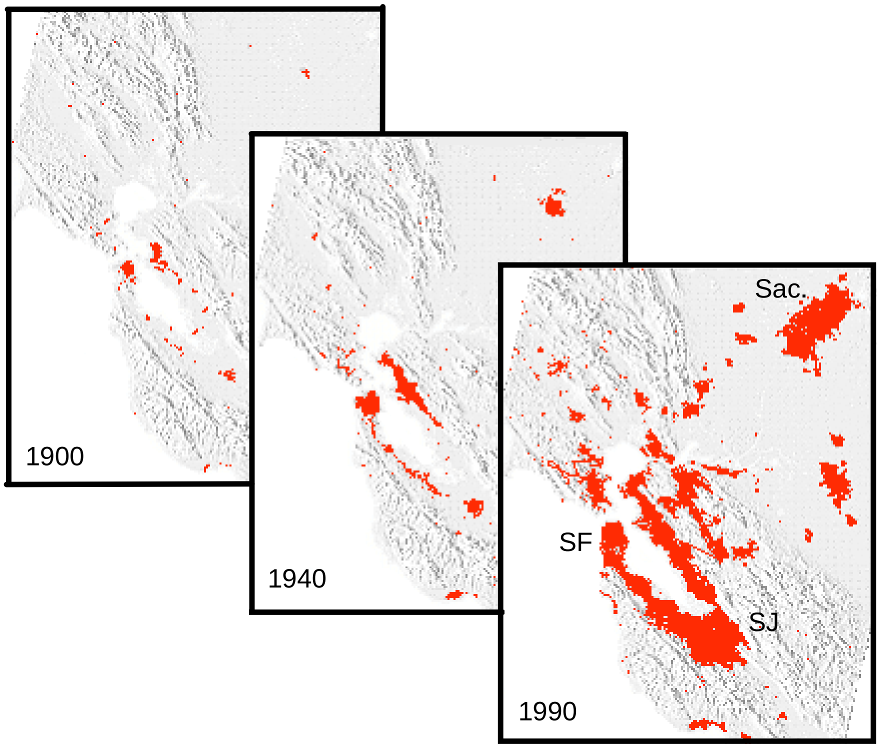


Figure 15: Urban growth in the San Francisco/Sacramento area.
Source: U.S. Geological Survey, Menlo Park (see Kirtland *et al*. 1994).
An on-line animation is located at http://geo.arc.nasa.gov/usgs/HILTStart.

As is clear from this example, cities are not always isolated, well-defined entities. In particular in industrialized areas, clusters of cities have developed that sometimes merge into one another. This point is also made convincingly by recent work on city dynamics (Krugman 1991, Batty and Longley 1994; Batty 1995; Makse, Havlin, and Stanley 1995). In these models, large cities are not considered to be monocentric entities that are the result of a central planning process. Instead, the fact that urban shape is the result of many individual, decentralized and generally non-coordinated decisions is considered explicitly. A city is thus composed of a hierarchy of clusters of varying size that possess their own dynamics. Using models from diverse fields such as fractal theory and physical concepts on the behavior of aggregates of particles, a great degree of realism is achieved in replicating growth and form of major cities. Clearly much can be learned from these models, not only in terms of aggregate human behavior but also in understanding and possibly predicting current and future spatial dynamics of settlement patterns at a local scale.

5.3. Small area estimation

Thus far, this review paper has dealt largely with the simplest demographic indicators only. Population figures and major components *per se* are, however, often a poor proxy for social phenomena and dynamics that are relevant for applications in policy planning, natural resource management or agricultural research. Available census information unfortunately does not usually include a broad range of indicators. Instead, the source for such information are special purpose surveys which are time-consuming, expensive and often cover only a small portion of a country or region. Consequently, the small sample sizes in a survey do not allow one to draw reliable conclusions about the value of a variable within a small area. For example, data from the demographic and health surveys (DHS) are often designed to be representative at the national or state level only. That is, the sample sizes are such that they would result in unacceptably large standard errors if the results were aggregated to small areal units. Since many of the variables in the DHS represent very small proportions, large sample sizes are need to assure significance even at high aggregation levels. The same is generally true for income and living standards measurement surveys that are conducted regularly in many developing countries. For targeting specific policies, this level is often too coarse. If, for example, information regarding poverty indicators or health statistics, were available at a higher resolution level, much leakage of funds to non-target groups could be avoided.

In recent years, statisticians have developed techniques to improve upon the information content in the available sample information by exploiting additional information that is available at higher spatial resolutions. These techniques are generally termed *small area estimation* (see Platek *et al.* 1987, Ghosh and Rao 1994). Most of this work has concentrated on developing appropriate statistical estimation techniques, which will not be reviewed at this point. Instead, the principle behind SAE is described briefly.

A rural socioeconomic survey may have resulted in an estimate of average household wealth for a set of areal units such as states, y_s. For planning purposes, however, estimates of this variable at a lower administrative level, y_d, are required. The task is now to establish a relationship at the state level between y_s and a matrix of explanatory variables, X_s, which are also available at the district level (i.e., X_d). In the simplest case, the relationship at the state level can be estimated, for example, using regression models that yield parameters, β, for a model $y_s = X_s\beta + \varepsilon$. Under the assumption that relationships at the larger level hold for the lower administrative level, the district level values can be used in combination with the estimated parameters to derive a district-level estimate of household wealth. The metaphor frequently used to describe this principle is that the prediction "*borrows strength*" from relationships at the aggregated level.

The choice of explanatory variables is obviously restricted to those that are available for the districts. Yet, in socioeconomic studies, indicators with strong potential explanatory power are often also only available for the survey sites. The choice of independent variables is thus limited to census data published for the small areas, data from other surveys that are significant at the higher resolution level, or external data sets that provide a value for every point in the study area. Such information may include satellite data, interpolated environmental information (e.g., climatic, soils or elevation data), or indicators based on readily available data such as an accessibility index based on roads and settlements.

6. Conclusions and Recommendations

Four years ago Clarke and Rhind (1992) made a convincing argument for the development of spatially referenced population databases for use in integrated global, continental and regional research on human-environment interactions. In their recommendations, they stressed the need for a family of demographic databases ranging from low-resolution country-level databases to high resolution raster population grids that are compatible to one kilometer satellite data. As reviewed in this paper, progress has been made on the development of raster-based population data sets for various regions of the world, and previously inaccessible data have become freely available. These data are used in a multitude of projects concerning, among many others, agricultural research, global change studies, and natural resources management.

Yet, we are still a long way from the high resolution, comprehensive and reliable data sets that are required to model human-environment interactions with a satisfactory degree of realism. The problems faced in compiling and reconciling international databases are still very large. Copyright issues - especially in developed countries - continue to make data exchange difficult, and the differences in demographic definitions, timing of censuses, data quality and census geography make the development of high resolution, high quality databases a very difficult task.

This review of database design and modeling, was meant to help stimulate future work. Provided that data are exchanged freely and that the methods used are well documented, work by different agencies with expertise in different geographic regions and areas of expertise can be pooled. Through gradual, incremental improvements, better databases will thus become available over time. This requires a certain degree of coordination, and the remaining paragraphs are meant to contribute to the discussions about a compatible framework for the development of spatially referenced demographic databases. These observations and suggestions are divided into three areas: technical issues related to base data development, technical issues related to modeling population distribution, and recommendations for the design of future projects

Technical - Base data

- DCW is currently the best available digital base map despite its shortcomings that were well documented by Tveite and Langaas (1995). Until a better data set becomes available, the use of DCW coastline and international boundaries is recommended even in cases where the internal boundaries are of lower resolution. This will facilitate the integration of the resulting databases with population data from neighboring regions as well as with data on other variables that are increasingly referenced to DCW as well.
- Coding schemes are rarely discussed in great detail. Although there are, admittedly, more stimulating topics, there is a clear need for consensus on this issue. The lack of a standardized way of assigning unique and consistent identifiers to geographic regions that lend themselves to standard database operations limits the comparability of databases produced by different agencies. The simple coding scheme suggested in this report is not the final answer, but -

based on the conventions used by the UNSD Software Development Project - presents an initial structure that is very flexible and adaptable. This system could be developed further to incorporate standard conventions to indicate boundary and name changes that will become an increasingly complex problem as data for various regions accumulate over time. The experience gained in the CITAS project will be extremely valuable.

- Data sets are of very limited use without sufficient information regarding content, conventions used, sources, data lineage, and technical parameters employed. There is no such thing as *bad* data (although there are many *wrong* data), since a data set that is unsuitable for one task may still be sufficiently accurate for another. Perfectionism is a noble attitude and high resolution, error free databases should be the ultimate goal. But for many who are working in data scarce regions of the world and who need to incorporate data which cannot be conveniently collected with remote sensors, any information is often better than none at all. The important point is that users need to be given the ability to judge the status of a data set for themselves. The most problematic aspect of disseminating imperfect data sets is that it reduces the urgency for investing in the development of improved data sets. By having *something* available, the need to improve data sets becomes a less pressing issue. Part of the problem is that digital databases are still too often regarded as static products - like a book or musical recording. The development of digital data sets should, in contrast, be considered a continuous process of updating, improving and augmenting that yields many intermediate versions but perhaps never a final product. The emergence of the Internet as a means of distributing data widely with little cost (as opposed to magnetic media or CD/ROMS) makes this argument even more valid.

- Meta-data are conveniently divided into two parts: information on the geographic component and on the attribute data. For the first, information on source maps, source scale, spatial reference system (projection plus all parameters), and file formats are required. Data lineage information should include all processing steps (e.g., generalization, replacement of boundaries, or merging with other data sets). An estimate of positional error inherent in the source map or introduced during digitizing or scanning is useful, but this information can often be based on subjective judgment only. Increasingly, GIS vendors are including software functions that let the user attach comprehensive meta information to the data set. This will, in the long run, hopefully lead to better documented data sets.

  For the attribute data, the following information on demographic indicators should be given: data sources, data definitions, descriptions, temporal reference and measurement units. The specific techniques used to make indicators compatible, such as conventions used in standardization or forecasting techniques, need to be made explicit.

- As more GIS databases on human population characteristics are developed, a natural question is where the emphasis should be placed. The current databases, which often include little more than total population, can be considered a starting point only. To allow for greater flexibility, more meaningful analysis, and more reliable forecasting, the choice of variables needs to be expanded. Depending on the specific application, priority should be given to temporal dynamics (compilation of historical population totals), and on the inclusion of additional population parameters (e.g., gender, age, educational, economic or cultural characteristics). Temporal databases of a limited number of factors are probably the logical next step, simply because these are much easier to compile. Indirect techniques for estimating specific demographic, cultural or economic variables for subnational areas, are much less developed than models of population dynamics over time.

Technical - Modeling

- The overview of spatial population modeling approaches has made it clear that there is no single optimal strategy for modeling the distribution of human population based on limited information. It very much depends on the resolution, quality, and format of demographic information, and on the availability of auxiliary information. Where only population totals for relatively homogeneous administrative regions are available, the pycnophylactic interpolation approach represents a reasonable means of estimating spatial population patterns. Where, on the other hand, data are available for enumeration areas or even census blocks, there is little need for modeling. Most cross-national applications will be located somewhere between these two extremes. In these cases, an ideal strategy would combine approaches that have so far been employed in practice. For example, the meticulous, country-by-country approach employed by the CIR could be modified by a more formal inclusion of auxiliary information. This would relax some of the restrictive assumptions inherent in the CIR data (e.g., circular city shapes) and would make the output data sets more replicable. Of course, the resources required for such an approach would be considerably larger than those for the more automated techniques used for the development of continental databases.
- The use of satellite data in population distribution modeling, as often advocated, has so far not been of great help. Sensors with a sufficient spatial and spectral resolution are typically not practical for large areas. Sensors that provide large area coverage at reasonable cost (e.g., AVHRR) have so far shown little promise in improving the accuracy of population estimates significantly. DMSP data may be useful at least for delimiting urban areas. A "*census from heaven*" (Stern 1986), however, appears to be an unlikely option at least for the next few census rounds.
- Irrespective of the modeling approach, the assumptions that went into the models need to be made explicit. This ensures that the results can be replicated and that future modeling approaches will result in output that is comparable. This is particularly important if future studies incorporate these data for change analysis. Sufficient documentation will also avoid circularity in subsequent analysis where one of the population model inputs is used as an explanatory variable.
- Error analysis has been largely neglected in the development of currently available spatial population databases (admittedly including those that the author has worked on). Cross-validation is a viable option for assessing the errors in the interpolation of point data representing a continuously varying indicator such as temperature: each point observation is deleted in turn, the value for that location is estimated using the neighboring data points, and an aggregate error estimate or an error surface is generated. Similarly, kriging yields a spatially disaggregate estimate of the prediction error that can be used in evaluating the fit for particular areas. Unfortunately, these approaches are not appropriate for areal data where the

interpolation relies on an estimate of variation within each polygon.

A feasible approach is to model population distribution at a higher aggregation level than is actually available, and to compare the resulting totals at the lower level with recorded information. The resulting error measures will not be as reliable an estimate of the true error at the higher resolution level as, for example, cross-validated residuals for point data. But sensitivity analysis based on this approach will help identify particular regions that are not well modeled and for which additional data should be collected.

Future work

- A "first-cut" global population database is now available (Tobler *et al*. 1995), and the next goal should be to improve the data input and modeling strategy to produce "second generation" databases for all parts of the world. For several regions such work has already been completed or is ongoing. Global coverage at 2.5 or 5 minute resolution appears reasonable at the moment. One kilometer data that would match, for example, the global DEM currently produced by EDC and UNEP/GRID Sioux Falls would be meaningful only for those regions of the world where high resolution data at the sub-county or sub-district level are available. Concurrently, the development of national level databases that include a larger set of variables should be supported. Examples are the Rwanda, Mexico, and Nepal studies and the work by the UNSD Software Development Project which have all been mentioned previously.
- More emphasis should be put on the development of data on settlements. With increasing urbanization, a larger and larger share of the population resides in cities. These have the advantage of being spatially relatively well defined. Reliable information on the location, geographic size and population of towns and cities will therefore greatly facilitate population modeling for large areas.
- Data sharing and copyright issues remain one of the most formidable barriers to population data dissemination. The increasing commercial value of geodemographic data is simultaneously a blessing and a curse. On the one hand, the market value of population data can lead to the efficient development and distribution of administrative boundary and population data. On the other hand, many potential users will not be able to afford the often substantive costs, and - even more seriously - the copyrights that are put on the data and its derivative products interfere with the sharing of information upon which research and policy analysis need to rely.
- Most of the work on modeling large area population distribution has been driven by the global change community. With few exceptions, demographers have not been involved in these efforts to the extent that would be desirable. A closer, interdisciplinary integration of the relevant fields of demography, geography, economics and the physical sciences would be desirable.

7. References

Adams, J. (1968), A population map of West Africa, Graduate School of Geography discussion paper No. 26, London School of Economics, London.

Africa Institute (1965), *Africa - maps and statistics*, Africa Institute/Afrika Instituut, Johannesburg.

Bähr, J. (1992), Bevölkerungsgeographie, 2nd edition, Ulmer, Stuttgart.

Batini, C., S. Ceri, and S.B. Navathe (1992), *Conceptual database design. An entity-relationship approach*, Benjamin/Cummings, Redwood City, CA.

Batty, M. (1995), New ways of looking at cities, *Nature*, 377, 19 October, 574.

Batty, M. and P. Longley (1994), *Fractal cities: A geometry of form and function*, Academic Press, San Diego.

Benzine, D.E. and P. Gerland (1995), Accessing and using the Internet, Resource paper prepared for the TSS/CST Workshop on Data Collection, Processing, Dissemination and Utilization, New York, May 15-19, Department for Economic and Social Information and Policy Analysis, Statistical Division, New York (http://www.undp.org/popin/softproj/us/papers.htm).

Bogue, D., E.A. Arriaga, and D.L. Anderton, eds. (1993), *Basic readings in population research methodology*, 8 vols., United Nations Population Fund, New York.

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

Bröcker, J. (1989), How to eliminate certain defects of the potential formula, *Environment and Planning A*, 21, 817-830.

Brunner, J., K. Dalsted, and A. Arimi (1995), The use of aerial video for land use/land cover characterization and natural resource management project impact assessment in Niger, Report to the U.S. Agency for International Development/Niger.

Brunner, J., N. Henninger, U. Deichmann, and B. Ninnin (1995), *West Africa Long Term Perspective Study - Database and user's guide*, World Resources Institute, Washington, D.C. and Club du Sahel/OECD, Paris.

Carroll, G.R. (1982), National city-size distribution: what do we know after 67 years of research? *Progress in Human Geography*, 6, 1-43.

Carter, S., M. Collinson, K. Dvorak, J. Lynam and S. Romanoff (1992), Development of a socio-economic database for African agriculture: Summary of issues, paper prepared for the CGIAR/NORAGRIC/UNEP Meeting on Digital Data Requirements for GIS Activities in the CGIAR, Arendal, Norway.

Cho, L-J. and R.L. Hearn (1984), *Censuses of Asia and the Pacific: 1980 round*, East-West Population Institute, East-West Center, Honolulu.

Clark, C. (1951), Urban population densities, *Journal of the Royal Statistical Association A*, 114, 490-496.

Clarke, J.I. and D.W. Rhind (1992), *Population data and global environmental change*, Paris, IISC/UNESCO.

Clayton, C. and J. Estes (1980), Image analysis as a check on census enumeration accuracy, *Photogrammetric Engineering and Remote Sensing*, 46, 757-764.

Cohen, J. (1995), *How many people can the Earth support?*, Norton, New York.

Crissman, L.W. (1993), The spatial information infrastructure for Asian studies in Australia (SIIASA), Asian and International Studies Department, Griffith University, Nathan, Queensland.

Davis, B.A., J.R. George and R.W. Marx (1992), TIGER/SDTS - Standardizing an innovation, *Cartography and Geographic Information Systems*, 19, 5:321-327.

Deichmann, U. (1994a), *A medium resolution population database for Africa*, Technical paper and digital database, National Center for Geographic Information and Analysis, Santa Barbara.

Deichmann, U. (1994b), Applications of GIS for demographic and related statistics: A demonstration for Nepal, United Nations Statistics Division and United Nations Population Fund, New York.

Deichmann, U. and L. Eklundh (1991), Global digital datasets for land degradation studies: A GIS approach, United Nations Environment Programme, Global Resource Information Database, Case Study No. 4, Nairobi, Kenya.

Diakonis, C.A. (1968), *Ekistics - an introduction to the science of human settlements*, New York, Oxford University Press.

Dijkstra, E.W. (1959), A note on two problems in connexion with graphs, *Numerische Mathematik*, 1, 269-271.

Domschke, E. and D.S. Goyer (1986), *The handbook of national population censuses. Africa and Asia*, Greenwood Press, Westport, Conn.

Dvorak, K.A., ed. (1993), *Social science research for agricultural technology development. Spatial and temporal dimensions*, CAB International, Wallingford.

Eastman, R. *et al.* (1993), *GIS and decision making*, Explorations in Geographic Information Systems Technology, United Nations Institute for Training and Research, Geneva.

Ehrlich, P.R., A.H. Ehrlich, and G.C. Daily (1993), Food security, population and environment, *Population and Development Review*, 19, 1:1-32.

English, J., M. Tiffen, and M. Mortimore (1994), *Land resources management in Machakos District, Kenya 1930-1990*, Environment Paper No. 5, Washington, D.C., World Bank.

ESCAP (1988), *Censuses of population and housing in Asia and the Pacific: Towards the 1990 round*, Economic and Social Commission for Asia and the Pacific, Bangkok.

Feeney, G. (1994), Fertility decline in East Asia, *Science*, 266, 5190:1518-1523.

Fetter, B., ed. (1990), *Demography from scanty evidence, Central Africa in the colonial era*, Lynne Rienner Publishers, Boulder and London.

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, 211-224.

Flowerdew, R., M. Green, and E. Kehris (1991), Using areal interpolation methods in geographic information systems, *Papers in Regional Science*, 70, 303-315.

Foster, J.L. (1983), Observations of the Earth using nighttime visible imagery, *International Journal of Remote Sensing*, 4, 785-91.

Geertman, S.C.M. and J.R. Ritsema van Eck (1995), GIS and models of accessibility potential: an application in planning, *International Journal of Geographic Information Systems*, 9, 1:67-80.

Ghosh, M. and J.N.K. Rao (1994), Small area estimation: an appraisal, *Statistical Science*, 9, 1:55-93.

GIS World (1995), *GIS World sourcebook 1995*, GIS World, Inc., Fort Collins, CO.

Goodchild, M.F., L. Anselin, and U. Deichmann (1993), A framework for the areal interpolation of socioeconomic data, *Environment and Planning A*, 25, 383-397.

Goyer, D.S. and E. Domschke (1983), *The handbook of national population censuses. Latin America and the Caribbean, North America, and Oceania*, Greenwood Press, Westport, Conn.

Hartwell, R.M. and M.C. Hartwell (1993), Long-term time: historical datasets - Progress report and proposals for future development, Technical Paper, Center for Studies in Demography and Ecology, University of Washington, Seattle.

Heilig, G. (1994), *Demographics ’94 - User’s guide*, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Honeycutt, D. and J. Wojcik (1990), Development of a population density surface for the conterminous United States, *Proceedings GIS/LIS*, Anaheim, Vol. 1, 484-496.

IDP *et al.* (1988), *Population size in African countries*, originally in French: *Evaluation des effectifs de la population des pays africains,* Groupe de Démographie Africaine, IDP-INED-INSEE-MICOOP-ORSTROM, Paris.

IMF (1994), Proposal for a new set of country codes, Paper presented at the Fifth Meeting of the IMF Committtee on Balance of Payments Statistics, Basel, Switzerland, October 20-21, Statistics Department, International Monetary Fund, Washington, D.C.

Jones, H. (1990), *Population geography*, 2nd edition, Chapman, London.

Keersmaecker, M.L. de (1990), Testing urban density gradient models using satellite data, *Sistemi Urbani*, 2, 231-240.

Kirtland, D., L. Gaydos, K. Clarke, L. De Cola, W. Acevedo, and C. Bell (1994), An analysis of human-induced land transformations in the San Francisco Bay/Sacramento area, *World Resource Review*, 6, 2:206-217.

Klein, M. (1995), The data business goes global: new markets, new software, *GIS Europe*, March, 34-35.

Krugman, P. (1991), *Cities in space: three simple models*, National Bureau of Economic Research Working Paper No. 3607, Cambridge, MA.

Lang, L. (1995), Geographic information systems - Mapping for the masses, *Hemispheres*, October, 57-62.

Langford, M., D.J. Maguire, and D.J. Unwin (1991), The areal interpolation problem: estimating population using remote sensing in a GIS framework, in E. Masser and M. Blakemore, *Handling geographical information: Methodology and potential applications*, Harlow, Longman, 55-77.

Langford, M. and D.J. Unwin (1994), Generating and mapping population density surfaces within a geographical information system, *The Cartographic Journal*, 31, June, 21-25.

Langran, G. (1992), *Time in geographic information systems*, Taylor and Francis, London.

Lavely, W. (1994), China in time and space: An outline proposal, Center for Studies in Demography and Ecology, University of Washington, Seattle.

Leddy, (1994), Small area populations for the United States, Paper presented at the Association of American Geographers Annual Meeting in San Francisco, Geographic Studies Branch, International Programs Center, US Bureau of the Census, Washington, D.C.

Lo, C.P. (1986), *Applied remote sensing*, Longman, London.

MacIntire, J., D. Bourzat, and P. Pingali (1992), *Crop-livestock interactions in sub-Saharan Africa*, World Bank, Washington, D.C.

MacKellar, F.L. and D.R. Vining, Jr. (1995), Population concentration in developing countries: new evidence, *Papers in Regional Science*, 74, 3:259-293.

Makse, H.A., S. Havlin, and H.E. Stanley (1995), Modelling urban growth patterns, *Nature*, 377, 608-612.

Martin (1991a), Representing the socioeconomic world, *Papers in Regional Science*, 70. 3:317-327.

Martin (1991b), *Geographic information systems and their socioeconomic applications*, Routledge, London.

Martin, D. and I. Bracken (1991), Techniques for modelling population-related raster databases, *Environment and Planning A*, 23, 1069-1075.

Marx, R.W., ed. (1990), Special issue: The census bureau's TIGER system, *Cartography and Geographic Information Systems*, 17, 1:9-113.

McDonald, J.F. (1989), Econometric studies of urban population density - a survey, *Journal of Urban Economics*, 26, 3:361-385.

Meyer, W.B. and B.L. Turner II, eds. (1994), *Changes in land use and land cover: a global perspective,* Cambridge University Press, Cambridge.

Moxey, A. and P. Allanson (1994), Areal interpolation of spatially extensive variables - A comparison of alternative techniques, *International Journal of Geographic Information Systems*, 8, 5:479-487.

Munro, D.A., ed. (1990), *Cambridge world gazetteer: a geographical dictionary*, Cambridge University Press, Cambridge.

National Research Council (1993), *Demographic change in sub-Saharan Africa*, National Academy Press, Washington, D.C.

National Research Council (1994), *Memoranda from a workshop on research needs and modes of support for the human dimensions of global change,* Committee on the Human Dimensions of Global Change, Washington, D.C.

Ninnin, B. (1994), *The influence of markets on the distribution of rural population in West Africa*, Working Paper No. 4, Club du Sahel, Organization for Economic Cooperation and Development, Paris.

Noin, D. (1979), *Geographie de la population*, Masson, Paris.

Noin, D. (1991), *Atlas de la population mondiale*, RECLUS, Montpellier.

Nordbeck, S. (1971), Urban allometric growth, *Geografiska Annaler*, 53B, 54-67.

Nuttall, I. (1995), Epidemiological data - Schistosomiasis (reports by country), Division of Control of Tropical Diseases, World Health Organization.

Olson, J.M. (1994), *Demographic responses to resource constraints in Rwanda*, Rwanda Society-Environment Project, Working Paper No. 7, Michigan State University, East Lansing.

Paulsson, B. (1992), *Urban applications of satellite remote sensing and GIS analysis*, Urban Management Discussion Paper No. 9, World Bank, Washington, D.C.

Pingali, P., Bigot Y., and H.P. Binswanger (1987), *Agricultural mechanization and the evolution of farming systems in Africa*, Johns Hopkins University Press, Baltimore.

Plane and Rogerson (1994), *The geographical analysis of population*, Wiley, New York.

Platek, R., J.N.K. Rao, C.E. Särndal and M.P. Singh, eds (1987), *Small area statistics*, Wiley, New York.

Population Census Office (1987), *The population atlas of China*, Population Census Office and Institute of Geography, Chinese Academy of Sciences, Oxford University Press, Hongkong.

Pressat, R. and C. Wilson, eds. (1985), *The dictionary of demography*, Basil Blackwell, Oxford.

Prévost, Yves (1995), Data standards and harmonization: a view from the World Bank, paper presented at the International Union for Surveys and Mapping Workshop on Current Status and Challenges of Geoinformation Systems, Hannover, 25-28 September.

Republique du Zaire (1988), *Zaire - Recensement scientifique de la population - Juillet 1984 - Totaux Definitifs*, Institut National de la Statistique, Kinshasa.

Rhind, D. (1991), Counting the people: The role of GIS, in D.J. Maguire, M.F. Goodchild and D.W. Rhind, *Geographical information systems - principles and applications*, Longman, London, Vol. 1, 127-137.

Rondinelli, D.A. (1985), *Applied methods of regional analysis: the spatial dimensions of development policy*, Westview Press, Boulder.

Roy, B.K. (1988), *Census atlas - National volume - Census of India,* Office of the Registrar General, New Delhi.

Rossi, R.J. and K.J. Gilmartin (1980), *The handbook of social indicators*, Garland STPM Press, New York & London.

Segal, A. (1993), *An atlas of international migration*, Hans Zell Publishers, London, N.J.

Silverman, B. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, New York.

Snrech, S. (1995), *West Africa Long Term Perspective Study - synthesis report*, Club du Sahel, Organization for Economic Cooperation and Development, Paris.

Snyder, J.P. (1982), *Map projections used by the U.S. Geological Survey*, Government Printing Office, Washington, D.C.

Statistics Sweden (1993), *National atlas of Sweden*, SNA Publishing, Stockholm.

Stern, M. (1986), *Census from heaven?* Meddelanden fran Lunds Universitets Geografiska Institution, Avhandlingar XCIX, University of Lund, Sweden.

Stetzer, F.C. (1990), Applications of spatial autocorrelation analysis to colonial African census data, in B. Fetter, ed., *Demography from scanty evidence, Central Africa in the colonial era*, Lynne Rienner Publishers, Boulder and London.

Stewart, J.Q. and W. Warntz (1986), The physics of population distributions, in B.J.L. Berry and D.F. Marble, eds, *Spatial analysis: a reader in statistical geography*, Prenctice Hall, Englewood Cliffs, 130-146.

Stycos, J.M. (1994), Population, projections, and policy: A cautionary perspective, Environmental and Natural Resources Policy and Training Project Working Paper, Midwest University Consortium for International Activities, Madison, Wisconsin.

Sweitzer, J. and S. Langaas (1994), Modelling population density in the Baltic drainage basin using the Digital Chart of the World and other small scale data sets, in V. Gudelis, R. Povilanskas, and R. Roepstorff (eds.), *Coastal conservation and management in the Baltic region*, Proceedings of the EUCC-WWF Conference, 2-8 May, 1994, Riga-Klaipeda-Kaliningrad, 257-267.

Technical Paper, Stockholm, Beijer International Institute of Ecological Economics.

Tang, Q. (1994), Construction of the county boundary database of China, Technical Paper, Center for Studies in Demography and Ecology, University of Washington, Seattle.

Tobler, W.R. (1969), Satellite confirmation of settlement size coefficients, *Area*, 1, 30-34.

Tobler, W.R. (1970), A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 234-240.

Tobler, W.R. (1979), Smooth pycnophylactic interpolation of geographical regions, *Journal of the American Statistical Association*, 74, 367:519-530.

Tobler, W., U. Deichmann, J. Gottsegen and K. Maloy (1995), *The global demography project*, Technical Report TR-95-6, National Center for Geographic Information and Analysis, Santa Barbara.

Tufte, E.R. (1990), *Envisioning Information*, Graphics Press, Cheshire, Conn.

Tryfona, N. and J. Sharma (1995), *On information modeling to support interoperable spatial databases*, Technical Report TR-95-12, National Center for Geographic Information and Analysis, Orono, Maine.

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 5th Scandinavian Research Conference on Geographical Information Systems*, 12-14 June 1995, Trondheim, Norway, 145-154 (see also: http://ilm425.nlh.no/gis/dcw/dcw.html).

UNEP (1992), Global atlas of desertification, Edward Arnolds, London.

United Nations (1988), *The geography of fertility in the ESCAP region*, Economic and Social Commission for Asia and the Pacific, Asian Population Studies Series, No. 62-K, Bangkok.

United Nations (1989), *Handbook on social indicators*, Studies in Methods Series F, No. 49, Department of International Economic and Social Affairs, Statistical Office, New York.

United Nations (1992), Handbook of population and housing censuses, Part I, Studies and Methods, Series F, No. 54, Department of Economic and Social Development, New York.

United Nations (1993a), *World Urbanization Prospects*, Department for Economic and Social Information and Policy Analysis, Population Division, New York.

United Nations (1993b), Report on the programme to monitor the achievement of social goals in the 1990s and related methodological work, Report by the Secretary General, Statistical Commission, Twenty-seventh session, United Nations Statistical Commission.

United Nations (1994a), *Popmap - User's guide and reference manual*, Department for Economic and Social Information and Policy Analysis, Statistics Division, New York.

United Nations (1994b), *Population and the environment in developing countries: Literature survey and research bibliography*, Department for Economic and Social Information and Policy Analysis, Population Division, New York.

United Nations (1994c), Standard country and area codes for statistical use - Interim list, Department for Economic and Social Information and Policy Analysis, Population Division, New York.

United Nations (1995a), *World population prospects*, Department for Economic and Social Information and Policy Analysis, Population Division, New York.

United Nations (1995b), *Compendium of human settlements statistics 1995*, Department for Economic and Social Information and Policy Analysis, Statistics Division, New York and Centre for Human Settlements (Habitat), Nairobi.

Uvin, P. (1994), Violence and UN population data, *Nature*, 372, 6506:495-496.

Veldhuizen, J. van, R. van de Velde and J. van Woerden (1995), Population mapping to support environmental monitoring: some experiences at the European scale, *Proceedings, Sixth European Conference on Geographical Information Systems*, EGIS Foundation, Utrecht.

Vu, T.L. and C. Taillard (1994), *Atlas du Viet-Nam*, Reclus, La Documentation Francaise, Montpellier.

WHO/UNICEF (1995), *Joint programme on data management and mapping for public health - HealthMap*, World Health Organization, Division of Control of Tropical Diseases, Geneva, and United Nations Children's Fund, New York.

Willmott, C.J. and K. Matsuura (1995), Smart interpolation of annually averaged air temperature in the United States, *Journal of Applied Meteorology*, 34, 12:2577-2586.

Wint, W. and D. Bourn (1994), Anthropogenic and environmental correlates of livestock distribution in sub-Saharan Africa, Report prepared for the Overseas Development Association, Oxford, UK.

Woods, R. (1982), *Theoretical population geography*, Longman, London.

WRI (1995), Africa data sampler user's guide - CD-ROM version, World Resources Institute in collaboration with World Conservation Monitoring Centre and PADCO, Inc., Washington, D.C.

Appendix A

List of Priority Social Indicators proposed by the interagency working group of UNICEF, UNFPA, UNDP and UNSD (United Nations 1993b)

1. Infant mortality rate.
2. Under five mortality rate.
3. Maternal mortality rate.
4. Number of deaths from neonatal tetanus per 1000 live births.
5. Number of deaths from neonatal diarrhea per 1000 live births.
6. Number of deaths from neonatal pneumonia per 1000 live births.
7. Proportion of under five underweight.
8. Proportion of infants breastfed exclusively for the first four to six months of age.
9. Proportion of population with access to safe drinking water.
10. Proportion of population with access to sanitary means of excreta disposal.
11. Proportion of population with access to adequate shelter.
12. Proportion of children who suffer physical or mental abuse.
13. Primary school enrollment (gross/net).
14. Secondary school enrollment.
15. Proportion of first graders completing grade four.
16. Mean years of schooling per person 25+.
17. Adult literacy rate.
18. Tertiary science graduates ratio.
19. Scientists and technicians per 1000 population.
20. Life expectancy.
21. Distribution of age of mother at first birth.
22. Mean number of children ever born.
23. Median number of months since previous birth.
24. Proportion of births to females aged 20 to 34 years.
25. Proportion of households with female heads.
26. Contraceptive prevalence rate.
27. Total expenditure in social sectors as percentage of GNP.
28. Public expenditure in the social sectors as a percentage of total public expenditure.
29. Females per 100 males for population under age 10 and 60+.
30. Women per 100 men in wage and salaried employment (urban and rural).
31. Women per 100 men unpaid family workers (agricultural and non-agricultural).
32. Females per 100 males in rural to urban migration.
33. Percentage of women participating in grass-roots and community organizations.
34. Women per 100 agricultural holders.