Geographic Information Systems in Health

Basic Concepts

Special Program for Health Analysis
Pan American Health Organization
GEOGRAPHIC INFORMATION SYSTEMS IN HEALTH

BASIC CONCEPTS

Special Program for Health Analysis

PAN AMERICAN HEALTH ORGANIZATION
WORLD HEALTH ORGANIZATION
Geographic Information Systems in Health, Basic Concepts

Credits

The content of this manual was elaborated by:

Dr. Carlos Castillo-Salgado, Chief
Special Program for Health Analysis
Pan American Health Organization

Dr. Enrique Loyola
Special Program for Health Analysis
Pan American Health Organization

Mr. Manuel Vidaurre
Special Program for Health Analysis
Pan American Health Organization

Eng. Ramón Martínez
Instituto de Medicina Tropical “Pedro Kouri”
Cuba

Dr. Patricia Gassibe
Universidad Santiago de Chile
Chile

Geog. José M. Klarián
Universidad Tecnológica Metropolitana
Chile

Geog. Roberto González
Unidad de Análisis y Tendencias en Salud
Ministerio de Salud, Cuba

Geog. Miguel González
Universidad de Guadalajara
Mexico

Dr. Antonio Beas
Universidad de Guadalajara
Mexico

Other participants in the preparation of the manual:

Dr. Antonio Escamilla
Johns Hopkins University
United States

Ms. Anne Roca
Special Program for Health Analysis
Pan American Health Organization

Dr. Margaret Hazlewood
Special Program for Health Analysis
Pan American Health Organization

Special acknowledgements for their review and comments:

Dr. Oscar Mujica
Special Program for Health Analysis
Pan American Health Organization

Dr. Rocío Cuevas
Special Program for Health Analysis
Pan American Health Organization

Dr. Gabriela Fernández
Special Program for Health Analysis
Pan American Health Organization
PAHO Library Cataloguing in Publication Data

Pan American Health Organization

ISBN 92 75 12342 X

I Title
1. EPIDEMIOLOGY - Geographic Information Systems. 2. GEOGRAPHY - Health, Medical. 3. INFORMATION SYSTEMS - Health Geography. 3. CARTOGRAPHY - Health, Medical.

© Pan American Health Organization, 2000

Publications of the Pan American Health Organization enjoy copyright protection in accordance with the provisions of Protocol 2 of the Universal Copyright Convention. All rights are reserved.

The designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the Pan American Health Organization concerning the status of any country, territory, city, or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.
# Table of Contents

**Introduction**  
i

1. Geographic Information Systems Applied to Epidemiology  
   - Presentation  
   - Objectives  
   - Expected results  
   - Summary  
   - The role of epidemiology in public health  
   - Public Health Surveillance (PHS) systems for decision-making  
     - Identification of information sources, data collection, and data processing  
     - Data analysis and interpretation  
     - Dissemination of information  
   - The process of epidemiological analysis  
     - Measures of the frequency and association of health events  
   - Methods and techniques for epidemiological analysis  
   - Geographic Information Systems (GIS) in Public Health  
     - GIS applications in public health and epidemiology  
   - Glossary of selected epidemiological concepts for GIS in health  
   - References  

2. Cartography, Geographic Information Systems, and Spatial Analysis  
   - Presentation  
   - Objectives  
   - Expected results  
   - Summary  
   - Geography and cartography  
     - Scale  
     - Distance  
     - Elevation  
     - Orientation  
     - Location  
     - Projection  
   - Elaboration of maps  
     - General or base maps  
     - Thematic maps  
   - Geographic information systems  
     - Input  
     - Storage  


Processing 50
Output 55
GIS software programs used in public health and sources of data 57
Glossary of selected geographic and cartographic terms 60
References 66

3. Relational Database Systems in Geographic Information Systems 68
Presentation 68
Objectives 68
Expected results 68
Summary 68
Introduction to relational database systems 69
Informal description of the model 69
Relational Model (RM) 72
Keys 73
Relational assignment 75
Processing databases 79
Query languages 79
Structured Query Language (SQL) 79
System catalog 86
Additional examples of applications of the SQL in epidemiology 87
Glossary 88
Appendix: MapInfo Functions 90
Appendix: Relational Algebra (RA) 92
References 94
INTRODUCTION

In the past few years, the political, administrative and territorial decentralization initiatives have been a subject of growing concern all over the world. Reform of health systems is a component of the decentralization process. In this regard, the Pan American Health Organization (PAHO) is committed to support national policies and strategies directed at reorganizing the health sector. Our goal is to strengthen the decentralization processes and develop local health systems, within the Primary Health Care framework and, ultimately, towards the accomplishment of Health for All.

Decentralization of health services is seen as the transfer of technical and administrative control from the central to the municipal level. This strategy allows more specific identification of health problems, their characteristics in the population, as well as their time of occurrence. These health problems can be grouped by geographical areas, at either national, regional or local levels.

In 1995, the Special Program for Health Analysis (SHA) of PAHO launched a technical cooperation project (PAHO, 1996), to strengthen the application of epidemiology to health services. The objectives of the project are to simplify and disseminate the development and utilization of Geographic Information Systems (GIS) and to provide technical support for their applications in epidemiology and public health (SIG-Epi, for its Spanish name). The development of simplified applications is targeted to direct users, such as local health services, units of health situation analysis, educational institutions, and planning departments, among others. Given the present constraints to obtain basic digitized maps, an important objective of SHA is to serve as a clearinghouse of data and information from libraries on country geographic borders, that would be shared by users of PAHO member countries. The project’s activities include support for training of human resources, development of digitized cartographic databases, epidemiologic analyses, and elaboration of thematic maps of priority areas.

In order to successfully accomplish the objectives, SHA has promoted a network of Collaborating Groups to facilitate training and development of SIG-Epi in the region. Presently, three groups are actively involved: one in Chile (Universidad Santiago de Chile and Universidad Tecnológica Metropolitana), one in Cuba (Departamento de Bioestadística y Computación, Instituto Pedro Kouri, and Unidad de Análisis y Tendencias de la Salud, Ministerio de Salud), and the other in México (Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara). These groups have helped to expand the coverage of human resource development, developed different GIS materials, generated more diverse and simplified applications, and enhanced the capability for direct technical support to GIS users. Hopefully, the network of GIS groups will continue to grow in other countries.

SHA has assisted in the organization of several SIG-Epi workshops, for more than 600 participants in 15 countries, since 1993. These workshops have been designed with levels of complexity ranging from basic to intermediate levels. Advanced level training is provided by other collaborative institutions. Basic level training is directed to local health services providers; the intermediate level training to professionals from regional and national health service levels, as well as to academic and research institutions.

The first series of workshops highlighted the need for educational materials in epidemiology and public health, that were tailored to the needs of users. These materials are now being offered in this manual by SHA and the Collaborating Groups. Three disciplines are represented in GIS in Health: Epidemiology, Geography, and Informatics. It is expected that this manual will become a useful tool to support the assessment of health conditions as well as the formulation of health policies and decision making.

Chapter One (Geographic Information Systems Applied to Epidemiology) presents the methods and uses of epidemiology as they relate to the development and application of GIS in public health.
Chapter Two (Cartography, Geographic Information Systems, and Spatial Analysis) discusses basic geographic concepts, cartography, and aerial photography, in relation to GIS concepts and health situation analysis.

Chapter Three (Relational Database Systems in Geographic Information Systems) introduce basic concepts of relational database systems and the structured query language, including some of their applications to epidemiology.

Two additional documents have been developed and published separately: the SIG-Epi User’s Manual and the basic EPIMAP Handbook. SIG-Epi is also a PAHO software for spatial epidemiologic analysis that works in Windows environments and that includes other features allowing more integrated management of different georeferenced databases. EPIMAP v 2.10 is a user-friendly public domain software developed by CDC. Even though it is not a comprehensive GIS program, it is a very useful program for cartographic analysis of health events. At the time of publishing, the new version included in EpilInfo 2000, a Windows-based tool, has been released with more GIS capabilities.

SHA and the Collaborating Groups have developed several GIS applications on mortality atlas, health situation analysis, environmental health, accidents and injuries due to earthquakes in Chile, malaria morbidity in the North of Guatemala, morbidity in health jurisdictions of Jalisco, Mexico, assessment of epidemic diseases in Cuba, and basic health indicators in the Americas, malaria surveillance in Brazil, natural disasters in Central America, among others, as examples on the use and methods of GIS in public health.

---

2. CDC/EpiMap is available at ftp:ftp.cdc.gov/pub/Software/epimap
CHAPTER 1:

Geographic Information Systems Applied to Epidemiology
1. **Geographic Information Systems Applied to Epidemiology**

**Presentation**

Negative or positive health events seldom occur by chance. It is commonly accepted that, in general, these events are closely related to environmental and social conditions. For example, some health events tend to have a seasonal occurrence related to changing temperatures, some others are associated with living and working conditions of given population groups and with haphazard migrations, and still others are the consequence of catastrophic events such as earthquakes and floods. Their common characteristic is that they are spatially located, that is, they occur within a geographical boundary.

Epidemiology, a leading discipline of Public Health, has developed a variety of methods to assess the frequency, distribution, and determinants of health events occurring in human populations, in time and space. The ultimate purpose of epidemiologic knowledge is to improve the health conditions of populations. It guides decision-makers on the definition of health policies, plans, programs and services and the evaluation of their impact. The epidemiologist uses geographical referent to determine elements of cause and effect. Recently, the analytical potential of Epidemiology has been greatly expanded by technological developments in disciplines such as Informatics, Geography, and Biostatistics.

This chapter presents epidemiological principles and methods that are useful to determine the health needs of populations. It emphasizes the potential of Geographic Information Systems (GIS) as epidemiological analysis tools for describing the magnitude of health problems, identifying specific health determinants, and supporting health decision-making. This chapter also includes a discussion on Public Health Surveillance as a basic element for GIS epidemiological analysis, an introduction to common epidemiologic methods and techniques applied to public health, and an appendix of selected Epidemiology and Public Health concepts and definitions.

**Objectives**

- To introduce the basic concepts of Epidemiology and Geographic Information Systems.
- To present the contributions of GIS to epidemiological analysis and to Public Health practice.
- To discuss methods and techniques of Epidemiology related to the analysis of geographic information.

**Expected Results**

After reviewing this chapter, the user should be able to apply the basic concepts of Epidemiology and GIS, and apply them to public health practice.

**Summary**

Epidemiology studies the frequency, distribution and determinants of health events in human populations, as well as the application of this knowledge to design, implement, and evaluate interventions for preventing and controlling health problems. The characterization of health events in time and space helps to identify associated factors and to design actions for the prevention or elimination of negative health occurrences. This chapter reviews basic epidemiological methods and their application to GIS. The integration of Epidemiology and GIS methods and techniques simplifies the epidemiologic analysis of health events and their determinants, in support of public health interventions and decision-making. The potential benefits resulting from this integration are illustrated using real-life examples such as: the analysis of risk areas for the transmission of cholera in Santiago, Chile; the 1997 Health in the Americas situation analysis; the analysis of risk predictors of malaria in the Mexico-Guatemala border; and the analysis of the meningococcal meningitis epidemic in Cuba, among others.
The role of Epidemiology in Public Health

Epidemiology studies the frequency, distribution, and determinants of health events in human populations, as well as the application of this knowledge to design, implement, and evaluate interventions targeted to prevent and control health problems.

One of the main applications of Epidemiology is to facilitate the identification of geographical areas and population groups at greater risk of premature disease or death and therefore having a greater need for health promotion, preventative and/or curative interventions. Epidemiologic findings have shown that the frequency, distribution, and importance of risk factors that determine health are not always equally distributed among population groups. Some groups share similar health risk factors. The characteristics of these groups may be used to devise specific social and health interventions to decrease or eliminate specific health risks. This perspective of Epidemiology implies that health services must respond not only to health care demands, but also to unmet health needs.

Epidemiological interventions must be followed by the evaluation of their impact on the health of populations, to monitor and modify health interventions and programs accordingly. This dynamic process of diagnosis-action-evaluation-correction is part of the methodological approach known as Epidemiological Stratification (Castillo-Salgado, 1993).

In many countries, health resource constraints and decentralization of health services compel social and health program managers to be more effective and efficient in their decision-making. To meet these challenges, decision-makers need reliable information systems that are able to identify areas and populations with greater unmet health needs. A comprehensive epidemiological theory, the availability of greater computer power, and the advancement of information sciences and telecommunications greatly enhance the efficiency and effectiveness of current health information systems.

Considering that nearly 80% of the information needs of local health system decision- and policy-makers involve geographical positioning (Williams, 1987), the use of maps, especially in computerized format, contributes to more effective decision-making. Maps are one of the most efficient types of graphical display of information, since they summarize -in a limited space (the size of the map)- a great amount of information on the magnitude and distribution of events. The same amount of information would require much more space using other methods, for example, in a table or matrix (Tufte, 1983). In addition, maps display spatial patterns of health events, something that cannot be done easily using a table.

Figure 1.1 shows two ways of displaying data on malaria in the Americas: tabulated data and a thematic map. By looking at the table, it is quite difficult to identify the countries that have an Annual Parasite Index (API) greater than 10 and a downward trend in the number of cases between 1990 and 1993; the map effectively displays these features.

Geographic Information Systems (GIS) are computerized tools to manage, process, and analyze data (including multiple variables simultaneously); they allow the integration of large amounts of data from different sources in maps, graphs, and tables. This means that GIS allow the multiple processing of data that would normally require 2 or 3 software packages. Thus, GIS are among the most effective existing technologies that facilitate the use of information and decision-making processes in public health.

GIS were developed with military strategic purposes over 30 years ago. After the end of the Cold War a decade ago, they were made available for non-military purposes, such as environmental protection, exploring mining and forestry resources, marketing, improvement of services, real estate registries, etc. As has been the case in other areas of knowledge, the application of GIS to epidemiology and public health is quite recent. Reasons for the late advent of GIS are the limited expertise with methods, techniques and processes, and the unavailability of simplified computerized tools to conduct epidemiological analyses. To close these gaps, the Special Program for Health Analysis (SHA) of the Pan American Health Organization launched the Geo-
An important source of information for epidemiologic analysis and public health activities is provided by public health surveillance systems. Processing such information through a GIS system with epidemiologic analytical tools is a very effective and efficient way to conduct a health situation analysis for health services planning and organization, as well as for evaluating health strategies.
Public Health Surveillance (PHS) Systems for decision-making

Public Health Surveillance (PHS) is defined as the ongoing and systematic process of observation, collection, analysis, interpretation, and dissemination of information related to health events and determinants. It also includes the evaluation of sanitary and social interventions for user's feedback and adjustment of interventions.

The essential functions of PHS are:

- Identification of information sources, data collection, and data processing.
- Data analysis and interpretation.
- Dissemination of information.

Identification of Information Sources, Data Collection, and Data Processing

PHS systems require regular collection of data, specially those data related to infectious diseases or priority health problems, i.e., those for which reporting is mandatory by law or regulations. Historically, the emphasis of PHS on infectious diseases is explained by two reasons: (a) the significance of emerging or re-emerging diseases on population morbidity and mortality, and (b) the failure to control these diseases may result in epidemics that have important economic, social, and political consequences.

The development and operation of surveillance systems have been devoted primarily to the prevention and control of epidemics and outbreaks. However, the growing importance of noncommunicable diseases in the epidemiological profile of several countries, –often associated to the demographic transition and the changing economic conditions of societies– call for the development of new methods and approaches to health events and determinants. To this end, the Pan American Health Organization has advanced the concept of Risk Factor Surveillance as a component of Public Health Surveillance, which involves gathering and developing new information sources. These activities are simplified by the increasing availability of health information systems, the challenge being their accessibility, communication and integration.

Health information sources for PHS systems include:

- Registries for mandatory notifiable diseases.
- Hospital registries: hospital discharge cause-specific morbidity reports, laboratory registries, special reports (for example, nosocomial infections), and pathology reports. These information sources are the basis of nosocomial surveillance systems.
- Health care registries: patient visits, and health preventive and curative interventions (for example, vaccination) at different levels of care.
- Registries of environmental and occupational quality monitoring: monitoring of air and water environmental quality, earthquakes, etc.
- Population census, housing and economic registries.
- Specific surveys of different health programs: drug abuse, health, nutrition, availability and access to health services, quality of care, etc.
- Registries of cancer, birth defects, exposure to radiation, chemical, and toxic biological substances, etc.
- Sentinel center registries, including data reported by physicians, laboratories, and community groups.
- Remote satellite information registries for monitoring climates, environmental and oceanic temperatures, meteorological phenomena, rainfall, etc.
- Other surveys, including those on standards of living conditions.

These information sources usually work independently or side by side without communication and integration schemes. GIS support the integration of information sources due to their flexibility for database management and capability for simultaneous utilization of multiple variables.

PHS data required for monitoring the health situation and trends and for most epidemiological analy-
ses are distributed in various, dispersed information systems. A major regional effort to overcome such limitation (by consolidating an information system of basic, valid and consistent health-related data), is the “Core Health Data Initiative” of PAHO. This initiative is the result of the joint effort of information exchange between countries in the region of the Americas and PAHO. PAHO has been working on the development of a system of basic data and indicators that integrates a series of more than 120 health and health-related data, including health determinants, some of them disaggregated by age and sex since 1990. The Core Health Data System is a Web-based distributed engine application that includes a set of reference documents, a set of country health profiles, a tabulator and a geographic information system. The last two are dynamically linked to the database, where data are stored and processed.

PHS data collection from information sources is conducted at different time intervals, depending on the intended uses of information. Data collection requires proper case definitions and standardized and validated risk indicators for uniformity and comparability. Also, proper protocols are required for standardization and validation of data from diagnostic support resources, such as laboratories.

The data collection instruments should be designed according to the goals and objectives pursued by the information system. Basically, they should collect all the variables included in a pre-designed plan of analysis. Ideally, these data collection instruments should be comparable with national and international standardized instruments. The Tenth International Classification of Diseases (ICD-10) provides the latest description of conditions used to code mortality and morbidity causes. List servers are available at PAHO for review and technical discussions on disease classification. Some critical points should be considered for data comparison across different geographical units, for instance: the quality of diagnostic data and the consistency and coverage of registries. Also, the data collection instrument should contain information on the main risk factors in a way that allows stratification of those factors for priority setting.

Computers are indispensable for efficient data processing. Economic scarcity in many countries limits the availability of computer resources to the regional and central levels. However, computer data processing, which is increasingly available at the local level, should increase analysis and management capabilities. The unavailability of this type of technology restricts data processing to manual methods and calculators. Data processing and analysis procedures, (manual or semiautomatic) as well as the production of graphs and maps, should be planned for in advance to establish comparisons and tabulations of variables, and to allow the proper display and presentation of information. Proper planning will enhance the quality of information and the effectiveness of PHS.

Data Analysis and Interpretation

The purpose of data analysis and interpretation from PHS is to identify the frequency, distribution, and trends of health events and risk factors, to aid in the design of health interventions, and to evaluate their impact on the health of populations.

Biostatistics is a fundamental tool for analysis of health events and risks. Due to the widespread availability of powerful computers and statistical analysis software, the typical time and place data analysis is rapidly changing to more sophisticated statistical analyses and mathematical modeling; however in spite of their appealing complexity and power, good scientific thinking is still the basis for proper analysis.

The section on the Process of Epidemiological Analysis below presents a brief review of methods and techniques of analysis of health events, investigation of outbreaks and epidemics, and evaluation of health programs.

Dissemination of Information

This important component of a PHS system serves two objectives: to provide health information to health authorities, public and private entities, community organizations, and the population at large; and to provide timely feedback to the local health level. Information systems, and particularly surveillance

information systems, are not ends in and of them-

selves. All their activities are directed at becoming a

useful instrument for priority setting, decision-mak-
ing and evaluation, to devise optimal prevention and
control interventions. These basic and seemingly

obvious goals are frequently overlooked, rendering
surveillance systems ineffective. Thus, PHS system

managers at all levels should participate not only in
information systems activities such as designing re-
port outputs, but also in advancing proposals for health
interventions and organizational design at different
decision-making levels.

The process of Epidemiological Analysis

Once data from PHS systems, epidemiologic
studies, and other information sources are ready for
analysis, proper methods and techniques are avail-
able to analyze the frequency, distribution, and trends
of health events and risk factors, as well as their
possible interrelationships. Different epidemiological
methods and techniques are utilized for various pur-
poses such as health situation description and analy-
sis, outbreak/epidemic investigation, etiological hy-
potheses studies, health program evaluation, and
health services research. Their description will be
linked to georeferenced data managed by GIS.

Measures of the Frequency and Association
of Health Events

One of the basic activities of epidemiological
analysis is the description of the frequency of health
events and risk factors using the following measures:

- Number of events
- Proportions
- Ratios
- Rates

The number of events (e.g., number of menin-
gococcal disease cases, number of deaths, popula-
tion, number of earthquakes or volcano eruptions,
etc.) indicates the absolute magnitude of an event
and it is commonly used to present the frequency of
a rare event (less than 20 cases is conventionally

considered a rare event) (Cuzick & Elliot, 1992). The
number of events is used frequently in outbreak stud-
ies, cluster analysis, and in health services planning
and programming, particularly when there is a need
to estimate resources (human, material, etc.). The
advantage of this measure is its simplicity, since it is
merely a sum; its disadvantage is a limited ability to
compare geographical units in the absence of a de-
nominator (usually the population exposed to the
event).

Proportions (e.g., proportional mortality by
cause, percentage of urban population) are measures
that relate the number of events occurring in a given
place and at a given time with a reference popula-
tion. They may also be expressed as percentages.
Among the most commonly used measures of mor-
bidity, disability, and mortality are the prevalence
(number of events occurring at a given time) and the
cumulative incidence or risk of occurrence of a new
health event in a given period. Unlike absolute num-
bers, these are measures of relative frequency that
adjust the magnitude of a health event by the size of
the reference population; thus, the denominator in-
cludes the numerator. For example, an equal number
of cases of disease represent different risks in popu-
lations of different size. These measures are used in
descriptive, ecologic, and analytic epidemiological
studies.
Ratios (e.g., male:female ratio, 20% highest/20% lowest income ratio, population density per km², household overcrowding index, etc.) are measures that relate two events occurring in different populations and areas. Their main characteristic is that the numerator is not included in the denominator. Among the most commonly used ratios in public health are the standardized mortality ratio, the relative risk or risk ratio, and the odds ratio.

Rates (e.g., cholera incidence rate, malaria mortality rate) represent a change in health status per time unit. In epidemiology, rates correspond to incidence density measures, where the numerator is the number of new events and the denominator is person-time. This concept means that 100 person-years accrue from observing 100 persons for one year or ten persons for ten years. Expressing the prevalence, cumulative incidence or risk as a rate is incorrect and should be avoided.

Frequency and association measures are commonly used in epidemiology and in the analysis of geographically referenced information.

Rates, ratios, and proportions are very useful measures for mapping patterns of health events in populations of different size, particularly when both the number of events and populations vary significantly from place to place. However, special care should be taken when the number of cases is very small since this may greatly affect a rate. If this is the case, the following alternatives may be considered:

- In addition to rates, also show the absolute number of health events to convey their magnitude and the basis for calculation.
- Extend the observation time of the event to include a reasonable number of events.
- Exclude from the analysis small geographical units or outlier rates (for example, those falling beyond twice the standard error of the mean).
- Group small geographical units that have similar epidemiological characteristics into larger geographical units with a number of events suitable for statistical analysis. Grouping is made possible using GIS spatial redistribution functions (redistricting).
- Use spatial statistic methods, such as smoothing and testing randomness of rates.

These alternatives are not perfect and have both advantages and disadvantages, but they may be feasible under specific conditions.

Epidemiologic measures may be further refined to control the distortion arising from the comparison of populations with different distributions of age, gender, or other variables related to exposure to the factor or event of interest. The comparability of rates and ratios is accomplished through adjustment or standardization. The Pan American Health Organization and the Xunta de Salud de Galicia developed the computer software EPIDAT⁴ that includes techniques to carry out rate adjustments. SIG-Epi also includes tools for standardization and smoothing of rates that is presented in the specific Manual.

Methods and Techniques for Epidemiological Analysis

Epidemiology allows the description of spatial patterns of health events, their association with demographic characteristics, and their variation in time. The information resulting from this process is used to generate hypotheses on the etiology or causality of health events. These hypotheses generated from descriptive studies are subsequently tested in analytical studies. After obtaining frequency measures of health events and risks, special epidemiologic methods are used to accomplish the following specific purposes:

- Health situation analysis
- Outbreak and epidemic investigation
- Ecologic studies
- Migrant studies
- Evaluation of health interventions

⁴ http://www.paho.org/
Health situation analysis (HSA) is a set of methods and techniques to analyze the health conditions and determinants among populations in delimited geographical areas. From a time perspective, HSA includes strategic analysis and trend analysis. HSA are useful to identify differences in the frequency of health problems, patterns of morbidity and premature mortality risks, their social, economic, political and health resources factors, and their distribution across places and populations. These results may be presented on thematic maps to visually display information in a summarized and effective way. For example, the Atlas of Health Inequalities in the Americas (PAHO/SHA) contains a compilation of indicators of health inequity, geographically distributed by countries.

It should be noted that situation analysis results at the regional level do not necessarily represent the health situation at the national level, nor those at the national level represent the situation at the local levels. This is due to the presence of heterogeneity within geographical areas, to sociodemographic characteristics of populations, and to the aggregation levels of information. For example, the analysis of infant mortality in the Region of the Americas in 1995 revealed a regional average value of 19 deaths per 1,000 live births (PAHO, 1997). This figure is not useful to identify differences among countries of the region (a minimum value of 7 and a maximum of 74). Nevertheless, an analysis of smaller geographical areas or units at the sub-national level can show important differences within countries, such as differences in the death rate between 9 and 160 as shown in figure 1.2, consistent with the differential pattern seen at the regional level (PAHO, 2000). In these situations, GIS are useful instruments to visualize differences in health conditions and to locate critical areas and population groups for targeting health interventions.

Outbreak and epidemic investigations have special public health importance. An epidemic is defined as the occurrence of a number of disease cases greater than expected in a specific area or specific population group, in a given time period. An outbreak is the occurrence of a number of related disease cases greater than expected, in a more delimited area and time. A classical example is the cholera epidemic investigation by John Snow in 1854. After estimating cholera mortality rates in different areas of London along the Thames river, he compared them among areas receiving water from different sources, and found higher rates in areas supplied by one of the water supply companies. Snow tracked clusters of deaths down to a water pump on Broad Street. His findings led him to generate a hypothesis on the mode of transmission of cholera and devise interventions for epidemic control, even without identification of the etiological agent. The outbreaks of dengue fever, meningitis, measles, and cholera in the Americas are contemporary problems that have been properly studied using this approach.

Statistical methods such as time-series analysis are useful to detect unusual numbers of health events in different geographical areas or time periods. Small area and outbreak studies may be used to compare the frequency of a health event among different areas where an epidemic is suspected (English, 1992). One approach consists of the comparison of the number of events observed in a given geographical re-

---

igion, with the number of events expected if the incidence were similar to that of larger areas of reference. An assumption is that the observed number of events follows a Poisson distribution and that any difference from the expected number is due to random variation (and not to confounders and potential biases). Any statistically significant departure from this hypothesis (assessed by the p value) would be suggestive of a cluster. An alternative methodological approach assumes that each event represents the center of a cluster (see definition below). This method compares the number of observed events in other clusters (circles) with the number of expected events. This is the basis of analysis of clusters. The events that contribute more to statistical significance are considered as potential etiological factors.

If health events are not occurring uniformly, that is, if the patterns of occurrence differ during a specific time period among population groups, across geographical regions, or both, then determinants should be identified to explain such irregular pattern.

To evaluate the presence of a potentially irregular pattern, the following approach is recommended:

- Observe possible statistical associations suggesting a non-random pattern.
- Formulate a hypothesis for the spatial distribution of the health event.
- Test the hypothesis for possible determinants of the distribution.

These analyses are conducted as part of cluster studies (Aldrich et al., 1993). A cluster is defined as a set of events grouped within a defined geographical area, occurring in a number and concentration unlikely due to chance.

Figure 1.3 shows a flowchart for a systematic approach to cluster studies (CDC, 1990, modified by PAHO). The SIG-Epi Manual presents a more detailed discussion of the quantitative approach to cluster studies using GIS.

Care should be taken to identify false epidemics produced by artifacts of the surveillance system, such

---

**Figure 1.3: Stages of a cluster study**
as those due to higher or lower levels of case reporting. A survey on attitudes toward the surveillance system and the presence of a number of severe or complicated cases of disease may increase the awareness of disease occurrence and therefore raise the number of reported cases.

Ecologic studies (or geographical correlation studies) are used to describe quantitatively, the relationship of the frequency and geographical variation of a health-related event or a potential risk factor, among different population groups. These studies usually include environmental and social variables at aggregated levels; the place and time period of residence or work in a given area are utilized as surrogate exposure measures, assuming that a population group shares the same average value of exposure. Results may be presented on maps and correlation graphs. Statistical analysis includes correlation and regression models. These methods are used to evaluate association, assess the degree of association, and predict values of a variable as a function of the values of another variable.

Epidemiologic data are obtained by direct individual measurements (e.g., age, blood glucose levels, smoking) or from observations on groups, organizations, or regions (e.g., age-specific population groups, mean blood glucose levels among females aged over 60, the prevalence of teenage smoking, the levels of air pollutants in a district). In the first case, the variables are properties of the individual, and in the second case they are ecologic variables that are properties of groups, organizations, or regions. The term ecologic derives from the common utilization of geographical areas to define clusters or population groups, rather than individuals, as the units of observation and analysis. In general, group data from small areas are more likely to be applicable to individuals. Exposure measurement may be continuous (e.g., mean blood cholesterol in a region) or discrete (e.g., presence of tertiary level health care units in a region). A study of the association between blood cholesterol levels and diabetes mellitus with ischemic heart disease mortality in several geographical regions of Mexico is an example of an ecologic study. (Escobedo-De La Peña et al., 1994).

Ecologic studies are frequently used for epidemiologic research due to the following reasons (Morgenstern, 1998):

- **Low cost and convenience.** They are relatively inexpensive and quick, since data are usually available and easy to categorize.

- **Limited ability to measure health-related variables at the individual level.** In Environmental Epidemiology and other areas of research, it is frequently difficult or expensive to get precise measurements of a large number of individuals. In this case, ecologic studies are the only alternative for measuring exposures, particularly when studying clusters of diseases in small areas.

- **Design limitations of studies where individuals are the unit of analysis.** It may not be practical to study exposure effects when exposure levels show little variation. Ecologic studies that cover larger areas may detect variation in the mean exposure of population groups.

- **Interest in ecologic effects.** Ecologic effects may be particularly relevant to evaluate health policies and programs. An assumption is made that the sum of individual effects is not equal to the group effect.

A useful feature of ecologic studies is the estimation of the ecologic risk ratio, a measure of the presence of a characteristic or risk factor (a concept different from the relative risk in analytical epidemiological studies of individuals), at a relatively low cost. The value of the slope of a regression line relates the frequency of a health problem in the population (dependent variable) to the frequency of the exposure factor (independent variable). It provides an estimate of the risk ratio and represents the risk difference using reference geographical area (the value of the intercept) (Morgenstern, 1995).

Many ecologic studies are analyzed using only the correlation coefficient. This coefficient may not be used to make inferences, i.e., it is not a measure of the nature of the relationship; it only quantifies the presence and strength of a linear relationship.

---

6. This term indicates the measurement of exposure at the group level, to distinguish it from the risk ratio or relative risk at the individual level.
Care should be taken in the interpretation of results from ecologic studies. They may be less affected by random error of exposure measurements, but they are vulnerable to potential confounders (for an explanation, see confounding in the Appendix). The validity of these studies depends on whether the exposure indicator (area of residence) adequately estimates the usual exposure of an individual. For example, in studies that estimate the health effects of exposure to air contaminants, commonly used exposure variables are the mean (daily) levels of contaminants measured by area aerial monitors. Similar levels of exposure are assessed in individuals who live near a given monitor. However, these levels may vary greatly among individuals due to traveling to areas that have different concentrations of contaminants throughout the day (e.g., commuting to work). To the extent that the area monitor estimates the individual's actual levels of exposure, the ecologic estimate will be less biased. Another caveat in the interpretation of results from these studies is that the relationship between the exposure and the disease at the group level may not be the same as that at the individual level; this type of bias is known as “the ecologic fallacy” (Greenland and Morgenstern, 1989).

Another type of epidemiologic studies that uses geographical mapping is migrant studies that compare the risk of a health outcome among people who migrate from high-risk to low-risk areas or vice versa. They intend to discern the effects of environmental exposure from the effects intrinsic to the individual (for example, genetic factors). The results are typically presented as tables and graphs. An example of this type of design is a study of Mexican American children in border areas of the United States and Mexico (Castillo-Salgado, 1987).

Other analytical studies such as case-control or cohort studies, including measurement of individual health status and group exposures, may also benefit from the use of GIS. In this case, single or multi-level exposures (environmental or social variables) may be assessed, thus allowing estimation of the relative risk or the odds ratio. The overlaying capacity of GIS will allow the combination or categorization of several variables in different data layers simultaneously. For example, combining levels of accessibility to potable water systems with crowding levels will produce areas where both variables are high, areas where both are low and areas where one is low and the other is high.

PHS information may be used also in the evaluation of health interventions to assess the impact of health promotion and health risk prevention and control programs. Statistical analyses of these studies also use the ecologic risk ratio as impact measure. Impact evaluation is accomplished by comparing a given situation with historical controls (the same area before and after an intervention), or with concurrent controls (areas with different interventions in the same period), or a combination of both.

An additional type of design is quasi-experimental studies which allow evaluation of the impact of health information. Their advantage over ecologic studies is a greater control of study conditions and observations, by limiting the effects of some factors (e.g. changing the case definitions, the extent of prevention and control interventions, the diagnostic coverage, and the population location) that may contribute to health outcomes, therefore enhancing the internal and external validity of the study. An example of this type of design is the evaluation of the impact of measles vaccination on mortality rates from measles in the Americas.
Geographic Information Systems (GIS) in Public Health

A Geographic Information System (GIS) is a powerful computerized technology for analyzing large databases in a geographical context (Vine et al., 1997). Chou (1997) defines a GIS as an organized set of hardware, software, and geographical and personal data, designed to enter, store, update, manage, analyze, and present different forms of geographically referenced information, in a comprehensive fashion. Antenucci (Antenucci et al., 1991) defines GIS as a computerized system to store and associate geographically referenced data that has mapping and graphing abilities to perform a variety of operations such as production of maps, data processing, analysis, and modeling. A more operational definition considers GIS as a support system for decision making that involves the integration of spatially referenced data to perform problem-solving operations (Cowen, 1988).

The relationship between PHS and GIS is a dynamic, permanent, and continuing process of information exchange. PHS is the source of basic information regarding health events, risk factors, determinants, and interventions. Traditionally, maps have been used to describe the spatial distribution of health-related events. More recently, maps have been used to guide hypotheses on disease etiology, considering the spatial variation of social and environmental factors. The ability of GIS to process and manage large amounts of data and generate maps greatly enhances the analysis and synthesis of public health information. The results from analyses of geographically referenced data using GIS (new hypotheses and conclusions) feedback PHS systems, thus enhancing their quality in timeliness of information for decision making. These abilities of GIS greatly simplify the following activities:

- The location of health events in time and space
- The identification and monitoring of health events and risk factor behaviors in a given time period (weeks, months, years)
- The identification of time and space distribution patterns of risk factors and health outcomes
- The identification of geographical areas and population groups with greater health needs, and their possible solutions, by analyzing multiple variables (multivariate models)
- The evaluation of the impact of health interventions

The concept of Geographical Information Systems applied to Health encompasses the design, development, and utilization of GIS tools for the description of health situations; epidemiologic analysis; and public health management. The abilities of GIS to integrate and process data contribute to its great potential for application in different public health areas, and offer new opportunities to describe and analyze the relationships among environmental attributes and health events in geographical settings. Some of the main applications of SIG-Epi, are presented below (Pan American Health Organization, 1996):

- The spatial description of a health event
- The identification of environmental and occupational risks
- The health situation analysis in a geographical area
- The analysis of health situation patterns and differences at various aggregation levels
- The identification of high-risk groups and critical areas
- Public health surveillance and monitoring
- The generation of operations research hypotheses and new study areas
- The planning and programming of health activities
- The evaluation of sanitary interventions

Based on the previous definitions, the GIS systems operational framework consists of input, storage, processing, and output operations, (Figure 1.4). Input involves entering geographical or spatial data attributes, characteristics of a population or a geographical area and its environment. Data entering may be carried out through different methods: at the study site, directly or from other databases, and remotely by satellite. Data storage consists of construct-
ing digital databases for future use. At the time of process, data are retrieved, managed and transformed, and analyzed. Finally, data are further transformed into information and presented in maps, graphs, and, tables (output). Chapter Two on Geography, Cartography, and Geographic Information Systems presents these components in greater detail.

**GIS Applications in Public Health and Epidemiology**

A first step in data analysis using GIS is the spatial description of the situation of a health event in a geographical area. As an example, GIS were used in Cuba for meningococcal disease surveillance, to determine high-risk mortality areas, to estimate the absolute magnitude of the problem, and to identify the most frequent type of disease (SHA, 2000a). Databases containing individual, population and health care data were used to construct subject maps of mortality by municipality, overlapped with a cartogram of the number of cases and the proportions of disease types by province (Figure 1.5). The analysis of the causes of mortality in the municipalities of the US-Mexico border is another example. Differences were observed by cause and age/sex distributions. The age-adjusted data shows consistently higher mortality due to communicable diseases, accidents and diabetes in Mexican communities, and higher tumor-related mortality rates in the US.

The analysis of the causes of mortality in the Sister Communities (municipalities) of the US-Mexico border is another example (PAHO, 2000). Mortality data from the 1995-1997 period were provided by national authorities and then processed to identify the leading causes of death and to classify all causes (based on the ICD-9) according to six broad groups. Estimated populations for the same years based on projections from the 1990 Census in both countries were also obtained to calculate cause-specific mortality rates by geographic place, sex and age group. The proportionate mortality for the five leading causes of death indicated that about 19% of them in Mexican Communities and 30% in the US were due to diseases of the heart, in particular ischemic heart disease (around 65%). Malignant tumors were the second leading cause of death in Mexican (12%) and US (23%) Communities. However, the third most common cause of death in Mexico was accidents and adverse effects in Mexico, while it was the fifth (5%) in the US. There were differences by sex and age groups. The age-standardized death rates provided additional information on the spatial distribu-
Figure 1.5: Spatial description of a health event. Mortality from meningococcal infection by municipality and clinical types by province, Cuba, 1989-1993.

As indicated earlier in this chapter, caution should be taken when analyzing spatial data. Geographic boundaries may tend to distort the perception of high risk areas by paying too much attention to the size of the polygon represented. To limit this visual effect, data may be transformed to a continuous format (grid) and then interpolated to identify the high-risk strata made of several rather than individual geographic areas.

GIS are useful for the identification of areas with greater environmental risk and areas where industrial and technological development have generated new occupational risks for workers and the population at large. As an example, in Chile, a country where natural disasters such as earthquakes, droughts, floods, and volcanic eruptions are frequent, maps were created to locate these events (SHA, 2000b) (Figure 1.7). This was accomplished by compiling the registries of natural disasters occurring between 1960 and 1997, and locating them geographically across regions of the country. These maps show that the geographical diversity of the country is related to the location of disasters. Earthquakes are more frequent in the central and surrounding areas. Floods tend to occur with greater frequency and severity in the central area and in the southern regions where there is a greater number of superficial water sources. Droughts occur in the central-northern area, where rainfall supports agricultural and livestock production. Volcanic eruptions occur in the South, where active volcanoes are located. In contrast to the previous example, a thematic map in a single view, including various layers of locations, allows for a summarized perspective (Figure 1.8). Identification of
Figure I.6: Leading causes of death in the Sister Communities of the United States-Mexico border, 1995-1997 (age-adjusted rates per 100,000 population)

Figure I.7: Identification of areas of higher occupational and environmental risk. *Earthquakes, Floods, Droughts, and Volcanic Eruptions in Chile during the Period 1960 - 1997.*
areas where different natural disasters occur in Colombia may help to orient decisions on how to organize emergency preparedness and response according to local situations. When all the data are combined into GIS categories, it is possible to identify the overall risk areas. In addition, to assess the potentially affected populations, a layer containing the main cities (municipality capitals) was included. It may be observed that the largest areas in the southeast and northeast of the country are at lowest risk. However, most of the main cities, where population density is also higher, lie in average to high risk areas.

This information is critical to target surveillance efforts and to devise strategies for effective responses in the event of a natural disaster.

GIS may be used to analyze health situation patterns/differences at various levels of aggregation (from regional to local levels). As an example, Chapter I of the Annual Report from the Director of PAHO 1996 (Pan American Health Organization, 1997) presents the analysis of a number of indicators used to identify health conditions of populations of the Americas. First, countries were classified into five groups, from highest to lowest Gross Domestic Product (GDP) to make them comparable by controlling the confounding effect of different GDPs. Subsequently, four selected health indicators were mapped at the country level: life expectancy at birth, literacy, availability of drinking water, and basic sanitation. The analysis of georeferenced data evidenced that, in general, people in countries with higher GDPs had greater life expectancy. Other indicators led to similar findings, i.e., countries with greater resources had more favorable health conditions (Figure 1.9). Subsequent analysis of smaller geographical units (states) showed significant differences within these countries, that were not evident using national average values of indicators.

GIS may be used to identify high-risk groups and critical areas. This is simplified by their ability to process and analyze multiple variables simultaneously. As an example, GIS were used to analyze data from

**Figure 1.8:** Identification of areas of higher environmental risks. Natural disaster areas by type and their proximity to main cities in Colombia.
the Metropolitan Area of Guadalajara, Mexico, in 1990 (PAHO, 1997); the distribution of three indicators was analyzed: total population, overcrowding, and availability of indoor drinking water by basic geostatistical area (AGEB) or census unit. This information was obtained for more than 1600 geographical units of the 1990 National Census of Population and Housing database. The analysis consisted of grouping the indicators by quintiles and mapping the results individually (Figure 1.10). The large number of observation units made it difficult to identify areas with greater health needs (those with the most unfavorable conditions). Using GIS data management tools, data were reprocessed and summarized to identify the geographical units with greater health needs (combined effect). GIS proved to be a useful tool for the identification of unmet health needs.

Another example relates needs of certain risk groups with the availability of health care facilities in Washington, D.C. (Figure 1.11). This analysis is based on different information layers at a lower aggregation level (smaller geographic units). Block group level information from the 1990 U.S. Population Census was analyzed to identify critical areas. Critical areas
Figure 1.10: Identifying high-risk groups and critical areas. *Availability of drinking water, overcrowding, and total population, Metropolitan area of Guadalajara, Mexico, 1990.*

Figure 1.11: Identification of high-risk groups and critical areas. *Geographic coverage of health care facilities to high risk populations in Washington, D.C. 1990*
were defined as those with the highest quintile of Hispanic population proportion as well as those where the highest quintile of households with children having women as heads occurred. Blocks that met the criteria were selected with a query, and then the selection was transformed into another layer (shown in darker yellow). A set of georeferenced public and private institutions, including hospitals, in the District of Columbia, Virginia and Maryland was used; after their selection, they were also displayed in the map (blue crosses). An area of 1 mile around the hospital was delineated (red lines) to indicate coverage. From the composed thematic map it is possible to determine a lack of coverage for a large portion of the critical areas in the center. To show the capability of multi-layer handling of GIS when data are georeferenced, an aerial photograph of the areas (the Washington Mall) in the lower part of the map provides additional information on the appearance or on potential barriers.

GIS are very useful to support public health surveillance and monitoring. As an example, a study conducted in Brazil in 1992 using GIS showed that nearly 50% of the country’s one million cases of malaria occurred in a single state (Rondonia), which represented only a small fraction (less than 1%) of the total population of the country (Castillo-Salgado, 1993) (Figure 1.12). It was also found that only some of this state’s municipalities were the most affected and that the occurrence of malaria was limited within a specific area. This information was very valuable to estimate and target health resources to control a substantial proportion of malaria in the country.

To illustrate this situation graphically, two maps were prepared with the 1996 Annual Parasite Index (or API, which represents the number of malaria positive slides per 1,000 population in a given area): one at the state level, and the other at the municipal level (Figure 1.13). As shown in the state map, the API was almost 10 times higher than the national average in 6 out of the 9 Amazon states in the north and northwest. However, as may be seen in the municipality map, the distribution of malaria in the Amazon region of Brazil is heterogeneous. More cases tended to occur in certain areas, in spite of their relative adjustment by considering the population at risk. The API in the municipalities with darker colors (mainly in the corners of the Amazon region) indicated that, on average, almost one third of their inhabitants had one confirmed malaria episode in 1996, that is, almost 100 times more than the national average. Any effort to control the malaria problem should consider these large differences with critical areas and define the factors associated with them.

An additional example is that of surveillance and monitoring for meningococcal meningitis in Chile (SHA, 2000c), a disease requiring mandatory reporting of all suspect cases. From 1987 to 1993, there was an increase in the number of cases in Iquique (northern part of the country). In the early 1990s, an increase in the number of cases was also seen in Antofagasta (north of the country) and in the Metropolitan Region (Santiago), particularly affecting the group of children under five years of age. The occurrence of this disease reached epidemic proportions in August 1993, when the number of cases exceeded more than twice the number of cases registered in the same month of the previous year. Using GIS, cases were mapped by residence, which allowed the identification of associated variables such as the percentage of the population living in poverty and the population density (Figure 1.14). It was also possible to analyze the variation in the distribution of cases by geographical area (municipality) in the metropolitan region (Santiago). Before 1993, cases were homogeneously distributed; in 1993, cases were concentrated in municipalities located on the northern, central, and southern parts of the metropolitan region (Santiago), areas with the highest population densities.

GIS have also been used to generate and evaluate research hypotheses and to open new study areas. As an example, in a study conducted in the border areas of Chiapas in Southern Mexico and Petén, Guatemala (SHA, 2000d), different ecologic variables were analyzed as predictors of malaria. The approach consisted of mapping the municipalities reporting the distribution and frequency of vector species, including Anopheles albimanus, An. darlingi, An. pseudopunctipennis, and An. vestitipennis (Figure 1.15). It is known that ecologic conditions such as temperature, rainfall, and type of vegetation, determine the distribution of anopheles species.
Figure 1.12: Public Health surveillance and monitoring. *Malaria in the Amazon region and the State of Rondonia, Brazil, 1993.*

Figure 1.13: *Malaria Annual Parasite Index (API) in Brazil, 1996. State and municipal level aggregation*
Since altitude isolines are closely related to these ecologic variables, they were mapped to compare their distribution with that of insect vectors. It was concluded that vectors were not homogeneously distributed, but were spatially related to different altitude levels.

Because of the heterogeneity of socioeconomic conditions, agricultural development, and transformation of natural habitats, analysis was limited to a smaller (and presumably more homogeneous) area of Petén, to control for potential biases. Information available for analysis included incidence of malaria at the local level, (measured through the Annual Parasite Index (API)), altitude levels, and land use, (which is a proxy measure of human activities and development). These three variables were simultaneously mapped and it was seen that the risk of malaria in localities was not associated with altitude levels (since all the villages analyzed were below 200m over the sea level and therefore neither with vector-related conditions) (Figure 1.16). However, it was observed that localities with a higher risk of malaria were more frequently those bordering agricultural limits, that is, where deforestation was still incipient. This would create a favorable habitat for more efficient vectors such as *Anopheles vestitipennis* and *An. darlingi*, which thrive better in wilderness conditions. Findings from this study suggest that ecological factors such as altitude and deforestation, as well as socioeconomic factors, are major determinants of malaria risk; these findings are yet to be validated by analytical studies. GIS capabilities to manage different layers of data makes possible to overlap several of them simultaneously and to generate new spatial information elements derived from areas of intersection of spatial variables.

GIS are also useful to support health services planning. GIS were used to evaluate the adequacy of health services to health care needs related to severe malaria in the Department of Petén (north of Guatemala) (Loyola et al., 1998). First, localities with greater risk of malaria caused by *Plasmodium falciparum* were identified (Figure 1.17). Second, health services were mapped to determine whether they were geographically accessible and distributed according to health needs. Access roads, rivers and
Figure 1.15: Generation and evaluation of new research hypotheses and new study areas. 
*Altitude levels of the region of Chiapas, México, and Peten, Guatemala*

![Map showing altitude levels in Chiapas, México, and Peten, Guatemala](image)

- An. pseudopunctipennis
- An. vestitiperennis
- An. albimanus
- An. darlingi

Figure 1.16: Generation and evaluation of research hypotheses and new study areas. 
*Incidence of malaria, altitude, and land use in Peten, Guatemala, 1994*

![Map showing incidence of malaria, altitude, and land use in Peten, Guatemala](image)

- Agricultural frontier
- Incidence of malaria
coverage or catchment areas (within a 12 km. radius around health posts) were mapped. Catchment or buffer areas were generated using GIS buffering functions. It was found that some high-risk localities have little or no road access to health services, while others have an excessive supply of health services, as evidenced by the overlapping of catchment areas. These findings suggested the need to relocate health posts or deliver health services by mobile units to cover underserved areas along rivers. In this study, GIS aggregation functions made it possible to estimate the number of covered and uncovered populations and to calculate the resources needed to provide additional health services. Finally, GIS capabilities to calculate distances were used to trace routes and to estimate transportation resources to access health services.

In Chile, GIS were used to study the distribution, accessibility, and response capacity of the health services in the IV Region (Coquimbo) (SHA, 2000e).

A survey was conducted in primary care centers and hospitals to find the reasons and origin of doctor visits; data were collected on socioeconomic variables (pensions), basic sanitation, and health (permanent disability). Collected data were geographically referenced to determine the distribution of health care centers at different levels, and to establish the origin of doctor visits, in order to outline catchment areas. Findings showed that health services were properly distributed in areas of greater need, that is, where there were greater proportions of unemployed, retirees, and people with permanent disability (Figure 1.18). The GIS isoline interpolating modeling utility was used to establish access times (isochrones) of patients to health centers, taking into account the type of access route (asphalt, dirt road, etc.). It was determined that primary care centers were visited by patients due to geographical proximity (Figure 1.19), while attendance to hospitals depended on the availability of medical specialties, thus indicating that the level of response capacity was appropriate.

**Figure 1.17:** Health services planning and programming. Health services coverage, road access, and risk of malaria in Petén, Guatemala, 1991.
Figure 1.18: Health care planning and programming. *Access time to health care services by locality, Coquimbo, Chile.*

Figure 1.19: Health care planning and programming. *Distribution of health care services according to number of unemployed, retirees, or permanently disabled people, Coquimbo, Chile*
GIS have been applied to the evaluation of sanitary interventions. For example, GIS were used in Cuba to monitor meningococcal meningitis in order to identify municipalities with greater risk. It was possible to estimate the disease risk, and the effect of health interventions, such as the introduction of a vaccine in the population (SHA, 2000f). These objectives were accomplished by analyzing different databases containing individual data on cases and consolidated population data. Thematic maps of morbidity at the municipality level were created (Figure 1.20). To monitor the effect of the intervention, morbidity maps were compared over several periods, before and after vaccination, and to measure subsequent changes.

A final integrated example of the analytical capabilities of GIS in complex health processes is provided by the analysis of the effects of Hurricane Mitch in Central America (USGS, NASA, ESRI, CIAT and PAHO, 1999). Both vector and raster cartographic and attribute databases of Central America were used, having either on-site or remote data sources. Data processing included joining of multiple files, data transformation, data classification, establishment of buffer zones, distance measurements, image processing, and modeling, among other procedures. This health application shows that GIS may be used to identify high risk areas prior to the occurrence and not only following a natural disaster. It also shows the capacity of GIS on the use of risk factors and outcomes indicators for purposes of planning and organizing the preparedness and relief response before and after a disaster. In October-November, 1998, Hurricane Mitch hit the Atlantic Coast of Honduras with devastating effects (Figure 1.21. Through the use of different maps and types of information layers, it was possible to get a more comprehensive idea of the whole process that took place in that country. Display of the river network configuration and definition of the flood potential in some areas (based on an analysis of the slopes from digital elevation modeling) allowed to recognize a priori specific areas of major drainage coinciding with lowlands in the northern, southern and eastern coastal regions. While affecting the Honduran coast, Hurricane Mitch reached a width of nearly 500km, with high-speed

**Figure 1.20:** Evaluation of sanitary interventions. *Meningococcal infection in Cuba: Monitoring changes of meningococcal disease risk by year before and after introduction of a vaccine (1989-1994)*

<table>
<thead>
<tr>
<th>Year</th>
<th>Incidence rate per 100,000 pop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>6.7 - 6.92</td>
</tr>
<tr>
<td>1990</td>
<td>4.0 - 6.16</td>
</tr>
<tr>
<td>1991</td>
<td>2.0 - 4.82</td>
</tr>
<tr>
<td>1992</td>
<td>1.34 - 2.18</td>
</tr>
<tr>
<td>1993</td>
<td>0 to 1.34</td>
</tr>
<tr>
<td>1994</td>
<td>0.7 - 0.92</td>
</tr>
</tbody>
</table>
Figure 1.21: Use of GIS for integrated health applications. Effects of Hurricane Mitch in Honduras, 1998
winds (over 175 mi/h) and torrential rains (more than 30 in/d). After five days of transit over Honduras, devastation of the vegetation was evident on the ground, but that was also captured from satellite images, as shown by a normalized vegetation index (NDVI) from filtered and transformed remote data. The areas identified above were some of the most affected by the disaster. Aerial photographs from the city of Choluteca, in the south of Honduras, allowed verifying on land the damage suggested by climatic and satellite models. An indication of the magnitude of the floods is the river width, which went from 0.2 to almost 1 mi in the most affected area. As a result, an inhabited area of nearly 0.3 sqmi of the city was washed out. Surveillance reports then provided an overall account of the magnitude of people affected by the hurricane according to adminis- trative level. By linking attribute data provided by different institutions, it was possible to identify a north-south axis of areas where more people were affected.

In summary, epidemiology provides the essential tools for analysis and identification of unmet health needs. Health service managers may benefit from GIS capabilities by applying them to efficient resource allocation. GIS technology may be widely used as a support tool for public health activities. The ability of GIS to integrate information may simplify, expedite, and automatize epidemiologic analysis of health conditions. As a result, decision makers should be able to respond in an effective and timely fashion to popula- tion health needs.
Glossary: Selected Epidemiologic Concepts for GIS in Health

Note: Most of the following concepts have been taken or adapted from:
In addition to the Dictionary, some of these definitions are standardized by WHO.

**Active case finding**: The dynamic identification of the occurrence of a disease or health event under surveillance. (E.g. house visits by community workers to identify cases of tuberculosis).

**Active surveillance**: The dynamic and regular seeking of data from participants in the surveillance system.

**Attack rate**: The proportion of those exposed to an infectious agent who become (clinically) ill.

**Bias**
Deviations of results or inferences from the truth, or processes that lead to such deviation. Any trend in the collection, analysis, interpretation, publication or review of data that can lead to conclusions systematically different from the truth. The deviation of the truth may occur due to:

1. Systematic (one-sided) variation of measurements from the true values (syn: systematic error).
2. Variation of statistical summary measures (means, rates, measures of association, etc.) from their true values as a result of systematic variation of measurements, other errors in data collection, or errors in the design or analysis of the study.
3. The deviation of inferences from the truth as a result of flaws in the study design, data collection, or analysis and interpretation of results.
4. A tendency of the procedures (in the design of the study, data collection, analysis, interpretation, review or publication) to yield results or conclusions that depart from the truth.
5. Prejudice leading to the conscious or unconscious selection of procedures in the study that depart from the truth in a particular direction, or unilateral interpretation of results.
6. The term "bias" does not imply necessarily an imputation of prejudice or another subjective factor, such as the investigator's desire for a particular outcome. This differs from the conventional usage in which bias refers to a partisan point of view.

**Case**: A person who meets the case definition. The definition of a case will depend on what one is trying to describe. Infection can be clinical or sub-clinical. Both types of infection can lead to a carrier state.

**Case fatality ratio**: The proportion of people who die as a proportion of all cases. This will vary depending on the case definition used.

**Catchment area**: Region from which the clients of a particular health facility are drawn. Such a region may be well or ill defined.

**Classification**: (Syn: categorization) Assignment to predesignated classes on the basis of perceived common characteristics. A means of giving order to a group of disconnected facts. Ideally, a classification should be characterized by: 1) naturalness, the classes correspond to the nature of the thing being classified; 2) exhaustiveness, every member of the group will fit into one (and only one) class in the system; 3) usefulness, the classification is practical; 4) simplicity, the subclasses are not excessive; and 5) constructability, the set of classes can be constructed by a demonstrably systematic procedure.

**Classification of diseases**: Arrangement of diseases into groups having common characteristics. Useful in efforts to achieve standardization, and therefore comparability, in the methods of presentation of mortality and morbidity data from different sources. May include a systematic numerical notation for each disease entry. An example includes the *International Classification of Diseases* (ICD).

**Cluster**: Aggregation of relatively uncommon events or diseases in space and/or time in amounts that are believed or perceived to be greater than could be expected by chance. Putative disease clusters are often perceived to exist on the basis of *anecdotal evidence*, and much effort may be expended by epidemiologists and biostatisticians in demonstrating whether a true cluster exists.

**Cluster analysis**: A set of statistical methods used to group variables or observations into strong interrelated subgroups.

**Community trial**: Experiment in which the unit of allocation to receive a preventive or therapeutic regimen is an entire community or political subdivision. Examples include the trials of fluoridation of drinking water and or heart disease prevention in North Karelia (Finland) and California.

**Community surveillance**: Surveillance where the starting point is a health event occurring in the community and reported by a community worker or actively sought by investigators. This may be particularly useful during an outbreak and where syndromic case definitions can be used.
**Concentration index:** A summary measure of the distance between the concentration curve and diagonal of perfect equality. The concentration curve plots the cumulative proportion of a variable against the cumulative proportion of the population, ranking the population by socioeconomic status or another third variable. The concentration index is defined as twice the area between the concentration curve and the diagonal. It varies from +1 (inequality concentrated in the least disadvantaged population group) to -1 (inequality concentrated in the most disadvantaged population group).

**Confounding:**
1. A situation in which the effects of two processes are not separated. A distortion of the apparent effect of an exposure on the risk brought about by the association with other factors that can influence the outcome.
2. A relationship between the effects of two or more causal factors, as observed in a set of data, in such a way that it is not logically possible to separate the contribution of a causal factor to an effect.
3. A situation in which the measure of effect of an exposure on risk is distorted because of the association of the exposure with other factor(s) that influence the outcome under study.

**Contact:** An individual who has had contact with a case in a way that is considered to have cause significant exposure and therefore risk of infection.

**Continuous variables:** Characteristics or attributes with a potentially infinite number of values along a continuum. Some examples include age, height, weight, and blood pressure.

**Correlation:** The degree to which variables change together.

**Correlation coefficient:** A measure of association that indicates the degree to which two variables have a linear relationship. This coefficient, represented by the letter r, can vary between -1 and +1; when r = +1, there is a perfect positive linear relationship in which one variable varies directly with the other; when r = -1, there is a perfect negative linear relationship between the variables. The measure can be generalized to quantify the degree of linear relationship between one variable and several others, in which case it is known as the multiple correlation coefficient.

**Criterion:** A principle or standard by which something is judged.

**Data:** A collection of items of information. Note: the singular of *data* is *datum*; in this age of sloppy speaking and writing, the plural noun is often accompanied by a singular verb.

**Database:** An organized set of data or collection of files that can be used for a specified purpose.

**Data cleaning:** The process of excluding the information in incomplete, inconsistent records or irrelevant information collected in a survey or other form of epidemiologic study before analysis begins. This may mean excluding information that would distort the results if an attempt were made to edit and include it in the analysis, but it can also introduce biases. The fact that this step has been taken should be reported, along with the results of the study of analyzed data.

**Data processing:** Conversion of items of information into a form that permits storage, retrieval, and analysis. Data may be transferred to optical mark-sense cards, or directly into electronic files by computer. The term is loosely used to mean statistical analysis of data by means of a computer program.

**Demand (for health services):** Willingness and/or ability to seek, use, and in some settings, to pay for services. Sometimes further subdivided into *expressed demand* (equated with use) and *potential demand,* or *need.*

**Disability-adjusted life years (DALYs):** A measure of the burden of disease on a defined population and the effectiveness of interventions. DALYs are advocated as an alternative to QALYS and claimed to be a valid indicator of population health. They are based on adjustment of *life expectancy* to allow for long-term disability as estimated from official statistics. However, their use as currently expressed and calculated may be limited because the necessary data are not available or do not exist. Moreover, the concept postulates a continuum from disease to disability to death that is not universally accepted, particularly by the community of persons with disabilities.

**Disability-free life expectancy:** (Syn: active life expectancy) The average number of years an individual is expected to live free of disability if current patterns of mortality and disability continue to apply.

**Discrete variables:** Characteristics or attributes in which continuity cuts do not occur naturally; contrary to what happens in continuous variables. Some examples include the number of missing, decayed, and filled teeth, and the number of malaria cases in a given area.

**Early warning system:** In disease surveillance, a specific procedure to detect as early as possible any departure from usual or normally observed frequency of phenomena. For example, the routine monitoring of numbers of deaths from pneumonia and influenza in large American cities has been used as an early warning system for the identification of influenza epidemics. In developing coun-
tries, a change in children’s average weights is an early warning signal of nutritional deficiency.

**Endemic**: The constant presence of a disease within a given geographic area or population group.

**Environment**: All that which is external to the individual human host. Can be divided into physical, biological, social, cultural, etc., any or all of which can influence health status of populations.

**Environmental health impact assessment**: A statement of the beneficial or adverse health effects or risks due to an environmental exposure or likely to follow an environmental change. Statements may contain or refer to results of epidemiologic or toxicologic studies of environmental health hazards.

**Epidemic**: The occurrence of cases of an illness clearly in excess of expectancy. This is often referred to as an outbreak (more neutral). Endemic diseases are those that exist at higher rates over a prolonged period of time.

**Epidemiological case definition**: The definition of a case used for reporting to the surveillance system. The definition may be clinical, laboratory or both. It may relate to a specified disease (e.g. measles, yellow fever) or may identify a syndrome (e.g. meningitis, AFT).

**Epidemiological stratification**: It is the continuous process of investigation, diagnosis, analysis and interpretation of information, to methodologically categorize homogeneous geo-ecological areas and population groups according to their risk factors. An *epidemiological stratum* is a group of individuals or populations in defined geographic areas, sharing a similar hierarchy of main risk factors; thus, health services, programs and interventions to modify these factors are also similar.

**Error**
1. A false or incorrect result obtained in a study or experiment. In epidemiology various types of errors can occur, for example, due to bias.
2. Random error is the part of the variation in a measurement that has no apparent connection to some other measurement or variable, usually considered as due to chance.
3. Systematic error that usually has an identifiable source, for example, a defective instrument of measurement or pattern. It is consistently erroneous and in a particular direction.

**Exposed**: Someone who has met with an agent in a way that we know may cause disease.

**Gini coefficient**: This is a summary measure, based on the Lorenz curve, used to express inequalities. The Lorenz curve is a graphic representation of cumulative frequency that compares distribution of a variable with its uniform distribution in the population. A uniform distribution would be represented by a 45-degree diagonal line; the further the Lorenz curve deviates from this line, the greater the inequality. The Gini coefficient is the ratio of the area between the Lorenz curve and the diagonal line and ranges from 0 (total equality) to a maximum of 1 (total inequality).

**Geographic Information Systems in Health**: This concept encompasses the design, development and utilization of Geographic Information Systems (GIS) tools for the description of health situations, the epidemiological analysis and public health management. The abilities of GIS to integrate and process data contribute to its potential for application in different areas of public health. Some of the main application areas of GIS in health are: the spatial distribution of a health event; the identification of environmental and occupational risks; health situation analysis in a geographical area; analysis of health situation patterns and differences at various aggregation levels; identification of high-risk groups and critical areas; public health surveillance and monitoring; the generation of research hypothesis; the planning and programming of health activities; and, the evaluation of sanitary interventions.

**Geographic pathology**: (Syn: medical geography) The comparative study of countries, or of regions within them, with regard to variations in morbidity/mortality. The (implied) aim of such study is usually to demonstrate that the variations are caused by or related to differences in the geographic environment.

**Health event**: Any event relating to the health of an individual (e.g. the occurrence of a specific disease or syndrome, the administration of a vaccine or an admission to hospital).

**Hospital surveillance**: Surveillance where the starting point for a report is the admission of a patient with a particular disease or syndrome.

**Health indicator**: A variable, susceptible to direct measurement, that reflects the state of health of persons in a community. Examples include infant mortality rates, incidence rates based on notified cases of disease, disability days, etc. These measurements may be used as components in the calculation of a *health index*.

**Health risk appraisal (HRA)**: Syn: health hazard appraisal [HHA]. A generic term applied to methods for describing an individual’s chances of becoming ill or dying from selected causes. The many versions available share several common features. Starting from the average risk of death for the individual’s age and sex, a consideration of various lifestyle and physical factors indicates whether the
individual is at greater or less than average risk of death from the commonest causes of death for his age and sex. All methods also indicate what reduction in risk could be achieved by altering any of the causal factors (such as cigarette smoking) that the individual could modify.

**Health services**: Services that are performed by health care professionals, or by others under their direction, for purpose of promoting, maintaining, or restoring health. In addition to personal health care, health services include measures for health protection, health promotion and disease prevention.

**Health services research**: The integration of epidemiologic, sociological, economic, and other analytic sciences in the study of health services. Health services research outcomes of health services. The aim of health services research is evaluation; several components of evaluative health services research are distinguished:
- Evaluation of structure, concerned with resources, facilities and manpower.
- Evaluation of process, concerned with matters such as where, by whom, and how health care is provided.
- Evaluation of output, concerned with the amount and nature of health services provided.
- Evaluation of outcome, concerned with the results, i.e., whether persons using health services experience measurable benefits such as improved survival or reduced disability.

**Health statistics**: Aggregated data describing and enumerating attributes, events behaviors, services, resources, outcomes, or costs related to health, disease, and health services. The data may be derived from survey instruments, medical records, and administrative documents. Vital Statistics are a subset of health statistics.

**Health systems research**: The coordinated study of determinants of health (nutrition, housing, employment, education, etc.) as well as factors directly associated with health, such as use and functions of health services. A term popularized by the WHO.

**Incidence**: The number of persons who fall ill with a certain disease during a defined time period.

**Incidence rate**: It is the rate at which new events occur in a population. The numerator is the number of new events that occur in a defined period; the denominator is the population at risk of experiencing the event during this period, expressed sometimes as person-time. In a dynamic population, the denominator is the average size of the population, often the estimated population at mid-period, for example, if the period is one year, the annual incidence rate is calculated.

**Inequalities in health**: The virtually universal phenomenon of variation in health indicators (infant and maternal mortality rates, mortality and incidence rates of many diseases, etc.) associated with socioeconomic status. It has been observed since the vital statistics of England and Wales were examined by William Farr (1807-1883) and reported annually from 1840. If anything, the gap between best and worst health experience has widened in recent decades even in the rich industrial nations.

**Information system**: A combination of vital and health statistical data from multiple sources, used to derive information about the health needs, health resources, costs, use of health service, and outcomes of use by the population of a specified jurisdiction. The term may also describe the automatic release from computers to stored information in response to programmed stimuli.

**Integrated surveillance**: Common approach that provides a universal surveillance service using similar structures and techniques.

**Interaction**
1. The interdependent operation of two or more causes to produce or prevent an effect. The biological interaction represents the interdependent action of two or more causes in order to produce, prevent, or control the disease.
2. Differences in the effects of one or more factors according to the level of the remaining factors.
3. In statistics, the necessity for a product term in a linear model.

**International Classification of disease (ICD)**: The classification of specific conditions and groups of conditions determined by an internationally representative group of experts who advise the WHO, which publishes the complete list in periodic revisions. Every disease entity is assigned a number. There are 21 major divisions (chapters) and a hierarchical arrangement of subdivisions (rubrics) within each in the tenth revision. Some chapters are “etiologic,” e.g., Infective and Parasitic Conditions; others relate to body systems, e.g., Circulatory System; and some to classes of conditions, e.g., neoplasms, injury (violence). The heterogeneity of categories reflects prevailing uncertainties about causes of disease (and classification in relation to causes). The tenth revision of the manual (ICD-10) was published by WHO in 1990, after ratification in 1989.

**International statistical classification of diseases and related health problems**: The tenth revision, known in short as ICD-10, was approved by the International Conference for the Tenth Revision in 1989 and by the 43rd World Health Assembly in 1990. It is the latest in a series of international classifications dating back to the Bertillon classification i.e., the International List of Causes of Death,
1893); ICD-10 came into effect at the beginning of 1993, exactly 100 years after the original. The tenth revision has 21 chapters and uses an alphanumeric coding system in order to provide a larger coding frame than previously, leaving room for future expansion. The chapters of ICD-10 are as follows:

I  (A00-B99): Certain infectious and parasitic diseases
II  (C00-D97): Neoplasms
III (D50-D89): Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV  (E00-E90): Endocrine, nutritional, and metabolic diseases
V  (F00-F99): Mental and behavioral disorders
VI (G00-G99): Diseases of the nervous system
VII (H00-H59): Diseases of the eye and annexes
VIII (H60-H95): Diseases of the ear and mastoid process
IX  ((I00-I99): Diseases of the circulatory system
X  (J00—J99): Diseases of the respiratory system
XI (K00-K93): Diseases of the digestive system
XII (L00-L99): Diseases of the skin and subcutaneous tissue
XIII (M00-M99): Diseases of the musculoskeletal system and connective tissue
XIV (N00-N99): Diseases of the genitourinary system
XV (O00-O99): Pregnancy, childbirth, and the puerperium
XVI (P00-P96): Certain conditions originating in the perinatal period
XVII (Q00-Q99): Congenital malformations, deformations, and chromosomal abnormalities
XVIII(R00-R99): Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified
XIX (S00-T98): Injury, poisoning, and certain other consequences of external causes
XX  (V01-V99): External causes of morbidity and mortality
XXI (Z00—Z99): Factors influencing health status and contact with health services

Laboratory surveillance: Surveillance where the starting point is the identification or isolation of a particular organism in a laboratory. (e.g. surveillance of salmonellosis).

Mandatory surveillance: A surveillance where participants must report to the system. Notifiable diseases are one example of a mandatory system where reporting is by law. Another may occur where for example, a health authority requires that all public laboratories to report specified diseases.

Migrant studies: Studies that take advantage of migration to one country by those from other countries that have different biological and physical environments, different morbidity and/or mortality experience, and different cultural background and/or genetic makeup. Comparisons are made between the experiences of morbidity or mortality of the migrant groups against those of the current country of residence. Occasionally the experiences of a number of different groups that have migrated to the same country have been compared.

Monitoring: The performance and analysis of routine measurements, aimed at detecting changes in the environment or health status of populations. Not to be confused with surveillance. To some, monitoring also implies intervention in the light of observed measurements.

Mortality: Also known as the crude mortality rate, it is an estimate of the proportion of a population that dies during a specific period. The numerator is the number of persons who die during the period; the denominator is the size of the population, commonly considered as the mid-year population. When the rate is low, it is a good estimate of the cumulative mortality rate.

Multilevel analysis: Methods of analysis that explain individual outcomes in terms of both individual and environmental or aggregate variables, thus avoiding the ecological fallacy.

Needs: (Syn: health needs, perceived needs, professionally defined needs, unmet needs) This term has both a precise and an all-but-undefinable meaning in the context of public health. We speak of needs in precise numerical terms when we refer to specific indicators of disease or premature death that require intervention because their level is above that generally accepted in the society or community in question. For example, an infant mortality rate two or three times greater than the national average in a particular community. It should be clear that even in this seemingly precise usage, there are implied value judgments. It must be explicitly stated that “needs” always reflect prevailing value judgments as well as the existing ability to control a particular public health problem.
Notifiable disease: A disease that must be reported by law or ministerial decree.

Odds: The ratio of the probability of occurrence of an event to that of nonoccurrence, or the ratio of the probability that something is to the probability that it is not so. If 60 smokers develop a chronic cough and 40 do not, the odds among these 100 smokers in favor of developing a cough are 60:40, or 1:1.5; this may be contrasted with the probability that these smokers will develop a cough, which is 60/100 or 0.6.

Odds ratio: (Syn: cross-product ratio, relative odds) The ratio of two odds. The term odds is defined differently according to the situation under discussion. Consider the following notation for the distribution of binary exposure and a disease in a population or sample.

<table>
<thead>
<tr>
<th>EXPOSED</th>
<th>UNEXPOSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease</td>
<td>a</td>
</tr>
<tr>
<td>No disease</td>
<td>c</td>
</tr>
</tbody>
</table>

The odds ratio (cross-product ratio) is $a/d$. The exposure-odds ratio for a set of case control data is the ratio of the odds in favor of exposure among the cases ($a/b$) to the odds in favor of exposure among noncases ($c/d$). This reduces to $a/d$. With incident cases, unbiased subject selection, and a "rare" disease (say under 2% cumulative incidence rate over the study period), $a/d$ is an approximate estimate of the risk ratio.

Outbreak: The occurrence of two or more linked cases of a disease. An epidemic limited to localized increase in the incidence of a disease, e.g., in a village, town, or closed institution; upsurge is sometimes used as an euphemism for outbreak.

Outcomes: All the possible results that may stem from exposure to a causal factor, or from preventive or therapeutic interventions, all identified changes in health status arising as a consequence of the handling of a health problem.

Outliers: Observations differing so widely from the rest of the data as to lead one to suspect that a gross error may have been committed, or suggesting that these values come from a different population.

Passive surveillance: Surveillance where reports are awaited and no attempt made to actively seek reports from the participants in the system.

Perceived need: A felt need. The term usually refers to need for health care that is felt by a person or community concerned but which may not be perceived by health professionals.

Performance indicators: Specific agreed measurements of how participants are functioning within the surveillance system. These indicators may measure both the process of reporting, action taken in response to surveillance information and the impact of surveillance on the disease or syndrome in question.

Periodicity: The presence of a repeating pattern of excess cases. The repeater period can be in years, months or weeks.

Poisson distribution: A distribution function used to describe the occurrence of rare events or to describe the sampling distribution of isolated counts in a continuum of time or space (for example, sample counts of radioactive disintegration per minute). This distribution is utilized to model person-time incidence rates.

Prevalence: The number of persons who have a disease or other health-related event at a specific time.

Priority setting (prioritizing): Policymaking evaluation process to determine options on what topics to address and in what order. Epidemiology provides some of the evidence base (i.e. magnitude of problem, severity, trends, amenability for control, etc.) for evaluating the options and to make the best possible recommendations.

Public health surveillance: the ongoing and systematic process of observation, collection, analysis, interpretation and dissemination of information related to health events (diseases, outcomes, risk factors, etc.) and their determinants, for the purpose of designing, planning and organizing health services. It also includes the evaluation of sanitary and social interventions for user’s feedback and adjustments.

Quality-adjusted life years (QALY): An adjustment of the life expectancy that reduces the overall life expectancy by amounts which reflect the existence of chronic conditions causing impairment, disability, and/or handicap as assessed from health survey data, hospital discharge data, etc. In practice, numerical weights representing severity of residual disability are established by the judgement of patients and health professionals.

Quality of care: A level of performance or accomplishment that characterizes the health care provided. Ultimately, measures of the quality of care always depend upon value judgements, but there are ingredients and determinants of quality that can be measured objectively. These ingredients and determinants have been classified by Donabedian into measures of structure (e.g., manpower, facilities), process (e.g., diagnostic and therapeutic procedures), and outcome (e.g., case fatality rates, disability rates, and levels of patient satisfaction with the service).
Quality of life: The degree to which persons perceive themselves able to function physically, emotionally, socially. Contrast health status, which is an objective measurement. In a general sense, that which makes life free of impairment, disability, or handicap, as used in the expression quality-adjusted life years. Somewhere between these is an estimate of the utility of life—for instance, in clinical decision analysis, the utility of life that is impaired by a disabling degree of angina pectoris may be compared with that of a life which may be shorter in duration but free of disabling pain as a result of applying therapeutic procedures.

Quasi-experiment: A situation in which the investigator lacks full control over the allocation and/or timing of intervention but nonetheless conducts the study as if it were an experiment, allocating subjects to groups. Inability to allocate subjects randomly is a common situation that may be best described as a quasi-experiment.

Regression analysis: Given data on a dependent variable y and one or more independent variables x1, x2, etc., regression analysis involves finding the “best” mathematical model (within some restricted class of models) to describe y as a function of the x’s, or to predict y from the x’s. The most common form is a linear model; in epidemiology, the logistic and proportional hazards models are also common.

Relative risk
1. The ratio of risk of disease or death among the exposed to the risk among the unexposed; this usage is synonymous with risk ratio.
2. Alternatively, the ratio of the cumulative incidence rate in the exposed to that in the unexposed, the cumulative incidence ratio.
3. The term “relative risk” is also used as synonym of odds ratio and, in some biostatistical articles, it has been used for the ratio of forces of morbidity.

Reporting completeness: Proportion of all expected reports that were actually received (usually stated as “% completeness as of a certain date”).

Reporting system: The specific process by which diseases or health events are reported. This will depend on the importance of the disease and the type of surveillance.

Reporting timeliness: Proportion of all expected reports that were received by a certain due date.

Risk: The probability that an event will occur, e.g., that an individual will become ill or die within a stated period of time or age. Also, a nontechnical term encompassing a variety of measures of the probability of a (generally) unfavorable outcome.

Risk assessment: The process of determining risks to health attributable to environmental or other hazards. The process consists of four steps, as follows:
1. Hazard identification: Identifying the agent responsible for the health problem, its adverse effects, the target population, and the conditions of exposure.
2. Risk characterization: Describing the potential health effects of the hazard, quantifying dose-effect and dose-response relationships.
3. Exposure assessment: Quantifying exposure (dose) in a specified population, based on measurement of emissions, environmental levels of toxic substances, biologic monitoring, etc.
4. Risk estimation: Combining risk characterization, dose-response relationships and exposure estimates to quantify the risk level in a specific population. The end result is a qualitative and quantitative statement about the health effects expected and the proportion and number of affected people in a target population, including estimates of the uncertainties involved. The size of the exposed population must be known.

Risk factor: An aspect of personal behavior or life-style, an environmental exposure, or an inborn or inherited characteristic, which on the basis of epidemiologic evidence is known to be associated with health-related condition(s) considered important to prevent. The term risk factor is rather loosely used, with any of the following meanings:
1. An attribute or exposure that is associated with an increased probability of a specified outcome, such as the occurrence of a disease. Not necessarily a causal factor. A risk marker.
2. An attribute or exposure that increases the probability of occurrence of disease or other specified outcome. A determinant.
3. A determinant that can be modified by interventions, thereby reducing the probability of occurrence of disease or other specified outcomes. To avoid confusion, it may be referred to as a modifiable risk factor.

Routine surveillance: The regular systematic collection of specified data in order to monitor a disease or health event.

Scenario building: A method of predicting the future that relies on a series of assumptions about alternative possibilities, rather than on simple extrapolation of existing trends. Trend lines for demographic composition, morbidity and mortality rates, etc. can then be modified by allowing for each assumption in turn, or combinations of assumptions. The method is claimed to lead to greater flexibility in long-range health planning than simple forecasting that relies only on extrapolation of trends.

Sentinel surveillance: The surveillance of a specified
health event in a sample of the population at risk. The sample should be representative of a total population at risk.

**Situation analysis:** Study of a situation that may require improvement. This begins with a definition of the problem and an assessment or measurement of its extent, severity, causes and impacts upon the community and is followed by appraisal of interactions between the systems and its environment and evaluation of performance.

**Socioeconomic status (SES):** Descriptive term for a person’s position in society which may be expressed on an *ordinal scale* using such criteria as income, educational level attained, occupation, value of dwelling place, etc.

**Standardization:** A set of techniques used to remove as far as possible the effects of differences in age or other confounding variables when comparing two or more populations. The common method uses weighted averaging of rates specific for age, sex, or some other potential confounding variable(s) according to some specified distribution of these variables. There are two main methods, as follows:

*Direct method:* The specific rates in a study population are averaged, using as weights the distribution of a specified standard population. The directly standardized rate represents what the crude rate would have been in the study population if that population had the same distribution as the standard population with respect to the variable(s) for which the adjustments or standardization was carried out.

*Indirect method:* This is used to compare study populations for which the specific rates are either statistically unstable or unknown. The specific rates in the standard population are averaged, using as weights the distribution of the study population. The ratio of the crude rate for the study population to the weighted average so obtained is the standardized mortality (or morbidity) ratio, or SMR. The indirectly standardized rate itself is the product of the SMR and the crude rate for the standard population.

**Standardized mortality ratio (SMR):** The ratio of the number of deaths observed in the study group or population to the incident number that would be expected if the study population had the same incidence rate as a standard or other population for which the incidence rate is known; this ratio is usually expressed as a percentage.

**Stratification:** The process of or result of separating a sample into several subsamples according to specified criteria such as age groups, socioeconomic status, etc. The effect of confounding variables may be controlled by stratifying the analysis of results. Stratification is used not only to control for confounding effects but also as a way of detecting modifying effects.

**Surveillance:** Continuous analysis, interpretation, and feedback of systematically collected data, generally using methods distinguished by their practicality, uniformity, and rapidity rather than by accuracy or completeness. By observing trends in time, place, and persons, changes can be observed or anticipated and appropriate action, including investigative or control measures, can be taken.

**Surveillance predictive value:** The likelihood that an “outbreak” detected by a surveillance system is truly an outbreak.

**Surveillance sensitivity:** The ability of a surveillance system to detect an outbreak. (The proportion of all outbreaks that could have been detected by the system).

**Survey:** An investigation in which information is collected systematically. It is usually carried out in a sample of a defined population group and in a defined time period. Unlike surveillance, it is not ongoing though it may be repeated. If repeated regularly surveys can form the basis of a surveillance system.

**Trend:** A long-term movement in an ordered series, e.g., a time series. An essential feature is that the movement, while possibly irregular in the short term, shows movement consistently in the same direction over a long term. The term is also used loosely to refer to an association which is consistent in several samples or strata but is not statistically significant.

**Unusual event:** The occurrence of a disease or health in excess of the expectation. This expectation is either a static or dynamic threshold set by the system.

**Z score:** Score expressed as a deviation from the mean value, in standard deviation units; the term is used in analyzing continuous variables such as heights and weights of a sample, to express results of behavioral tests, etc.

**Zero reporting:** The reporting of zero cases when the participant has detected no cases. This allows the next level of the system to be sure that the participant has not sent data that has been lost or has forgotten to report.
References


(21) Special Program for Health Analysis (SHA). Surveillance of meningococcal meningitis in urban areas of Chile. Scientific paper No. 103. 2000(c).


CHAPTER 2:

Cartography, Geographic Information Systems and Spatial Analysis
2. CARTOGRAPHY, GEOGRAPHIC INFORMATION SYSTEMS AND SPATIAL ANALYSIS

Presentation

All human activities have a spatial reference. Human populations living in society, have constructed or modified natural habitats to ensure survival of the species, thus creating complex economic, political, and social networks. All human phenomena including health, occur within a geographical setting. These phenomena form the ecological and geographical boundaries of our times.

This chapter presents the principles, methods and functions of Geography, with emphasis on its public health applications. In addition, it presents the concepts and general components of Geographic Information Systems (GIS), as well as the basic concepts of spatial information, cartographic techniques, and spatial analysis.

Objectives

- To present basic concepts of Geography applied to health.
- To define some of the functions of Cartography and its applications to Public Health.
- To present the basic concepts of spatial information, Geographic Information Systems (GIS), and spatial analysis techniques.
- To present spatial analysis methods and techniques used in a Geographic Information System in health.

Expected results

After reviewing this chapter, the reader should understand the basic concepts of Geography, Cartography, and GIS, as applied to Public Health. This understanding should serve to promote the application of these concepts and technologies.

Summary

Medical geography and cartography simplify the description and spatial analysis of health and disease conditions of population groups. Different methods of cartographic representation facilitate the analysis of the absolute or relative frequency of a health situation, its time trends, as well as the behavior of related physical, environmental, and social factors. The health services response to health problems can be analyzed, for example, by mapping the distribution and type of health units in particular catchment areas. The recent technological advancements of geographic information systems have simplified data collection, entering, storage, and integration of large databases with public health data to support problem-solving and decision making in public health across geographic areas.
Geography and Cartography

Geography and cartography are devoted to the study of the description, distribution, and interaction of the physical, biological, and cultural characteristics of the earth's surface (Merriam-Webster, 1998). Geography deals with the arrangement and interrelations of phenomena such as climate, terrain, vegetation, soil, population, land use, etc. It also deals with the interpretation of the significance of differences and similarities of relevant phenomena.

Every natural and social event occurs in a specific place and time. Place (geographical space) is the object of study of geography. Place is the scenario where all human activities occur. It is therefore of central importance for the analysis of health-related events. Place includes areas where its components interact and create new spaces with specific characteristics needed for comprehensive health analysis.

 Territory is the concept most closely associated to the concept of space; it emphasizes the significance of spatial units. According to specific criteria, space may be defined as a territory. For example, an artificial borderline may be drawn to define clearly the political boundaries separating two countries, but it will not define as clearly the social relations, natural features, and influences between those two countries.

Maps are graphical representations of a territory on a flat surface (a bi-dimensional model), either of the whole world or a portion of it. Maps display the distribution, situation, magnitude, and relations of a variety of natural and social phenomena using conventional symbols. They are useful tools to recognize the presence of spatial patterns of events of interest.

The use of maps has been closely linked to tactical and strategic activities of countries. Throughout history, maps have been used in public health. One of the most notorious and early examples is the study conducted by John Snow, where he analyzed the distribution of the cholera epidemic from 1849 to 1854 in the Soho District of London, as described in the classic paper “Of cholera” in London in 1855 (Snow, 1885). Snow mapped the distribution of cholera deaths and pinpointed the locations where 83 of them occurred (Figure 2.1). He noticed that cholera deaths were geographically related with water supply pumps on Broad Street. In addition, he compared cholera cases with non-cases living in the same areas; after comparing cases with historical controls and considering alternative explanations, he identified the water supply company Southwark Vauxhall as the source of the epidemic. These findings established the association between cholera deaths and contaminated water supplies even before Robert Koch identified the agent in 1883. The evidence from Snow’s analyses led city officials to shut down the water pump on Broad Street, which caused a subsequent decrease in the number of cholera cases in the area.

Figure 2.1: Distribution of cholera deaths and water pumps in the Soho District, London, 1855

Cartography deals with the theory and construction of maps. Maps are also graphical tools used to provide a simplified display of events occurring in a given place. The main elements of a map are: scale, distance, elevation, orientation, location, and projection.
**Scale**

The appearance of an image will vary depending on how close or how far the observation point is (MacCarthy and Lindberg, 1974). A scale represents the number of times that an image has been reduced from its real size. On a map, the scale depends on the area to be represented, the level of detail required, and the amount of information displayed. For example, a local health level may require larger scales to provide greater detail of potential health risks, such as contaminated water sources, mosquito breeding sites, industrial pollution sources, etc. The national and regional levels may not require so much detail.

The scale of a map is the ratio between the values of distances on a map and corresponding distances on the terrain and is given by:

\[
\text{Map scale} = \frac{\text{Distance on map}}{\text{Distance on terrain}}
\]

Scales on maps may be represented numerically or graphically. The numerical scale represents the correlation between the distance unit on the map and the corresponding distance on terrain; it is expressed, for example, as 1:50,000. A measure unit on the map represents 50,000 units on terrain, and is read as "one to fifty thousand". The graphical scale is a straight line segment divided into sections, usually placed over the margin of the map. This segment represents a distance unit on the map that is proportional to the real distance and is measured in linear units, thus allowing its comparison with other segments on the map to estimate the real distance.

**Distance**

Distance is calculated by multiplying a segment in centimeters or millimeters on the map, by its numerical scale. The result will be the real distance on terrain. As an example, 3.2 cm. on a map with a scale of 1:50,000, represents a real distance of 3.2 cm X 50,000 = 160,000 cm, 1,600 m, or 1.6 km. Distance may also be measured comparatively using a graphical scale.

Distance calculations serve different purposes in health services, for example, to estimate travel time to get to a health care center from a given locality, or to calculate the amount of fuel needed for transportation from one place to another. It is also used for estimation of the flying range of vector mosquitoes and the location of breeding sites, estimation of the amount of air pollution, etc. Distance calculation procedures may be quite complex to perform, but they are currently available and simplified in GIS software packages.

**Elevation**

The lines found on some topographic maps represent curves of equal elevation (altitude) in meters above sea level (masl), known as isolines or vertical data. Any point over the line or close to it will have the same value, usually shown in numerical form along the curve. In the case of mountains, for example, the maximum elevation is noted on the summit or centroid (Figure 2.2), which is readily recognized as the smallest curve in the center of a series of concentric curves representing a mountain mass.

The elevation of a terrain is the difference between the reference curve data in masl and a higher curve data. Elevation measurements have been utilized in ecologic studies to determine the distribution of infectious disease vectors, since elevation is closely related to ecologic and climatic conditions. Another use of elevation measurements is to predict and adjust hemoglobin levels to determine the prevalence of anemia.

**Figure 2.2: Definition of isolines in relation to the elevation**

![Figure 2.2: Definition of isolines in relation to the elevation](image)
Orientation

Orientation is used for establishing a given direction with reference to points on a compass, such as, to trace routes for a vaccination team. The magnetic North is the conventional reference point.

On maps, the North point is indicated by an arrow (Figure 2.3); conventionally, the top border of a map is oriented to the North. Most maps have a reference grid where vertical lines represent the meridians and horizontal lines, the parallels.

A compass is used to determine routes or directions more precisely; to do this, a compass is placed on a map and the North point on the map is aligned with the magnetic North on the compass. Salient features on the terrain serve as landmarks to trace the route to the desired place.

Figure 2.3: Orientation on a map

Location

In general, four types of graphical elements are used to represent the location of spatial events: points, lines, surfaces, and volumes.

These elements are generally represented through planar Euclidean geometry and localized mathematically through a longitude and latitude Cartesian coordinate system (X, Y) (Figure 2.4). If altitudes or volumes are considered, points must be represented on a XYZ coordinate system, of longitude, latitude, and altitude. These systems are mainly used to measure and locate, with a preset precision, the spatial location of population attributes, life and health conditions, and environments. Other spatial properties may also be located, such as distances, altitudes, vicinities, rates or risks of a health event on three-dimensional models, etc.

Projection

A projection is a mathematical method that allows transforming the curved surface of the earth from a three-dimensional space to a flat or two-dimensional plane.

Depending on the location and shape of a given area on earth, transforming it to a two-dimensional surface may distort its features and therefore the analysis and interpretation of events occurring in the area, such as health events. To prevent distortion, several projection methods have been developed to represent areas on a plane with minimal distortion. Before using a given projection method, it is necessary to assess its effects on other areas (Figure 2.5). For example, tropical regions are represented on world maps using the Mercator projection, which adjusts the areas to the North and South of the Equator, as they become more distant from it.

There are three basic types of projection systems that have different reference planes: cylindrical, conical, and azimuthal or planar. Projections are important methods for map elaboration since they deal with the distortion of territories, and therefore with the perception of their dimensions.

Figure 2.5: Map projection systems
Elaboration of Maps

Maps are cartographic products of different types, classified according to their scale and content.

By type of scale maps are classified as:
- Large scale. Scale values of 1:250,000 or greater. This type of scale is used when greater detail is needed, for example, at the local level.
- Medium scale. Scale values between 1:1,000,000 and 1:250,000, are useful to display regional levels.
- Small scale. Values less than 1:1,000,000 are used to display a large region with little detail, such as for countries of a continent or the world.

By their content, maps may be general or thematic:
- General or base maps represent geographical outlines and boundaries; for example, the administrative boundaries of municipalities, surface relief, hydrographic rims, main localities, roads, telephone lines, urban areas, etc.
- Thematic maps display natural, social, and cultural events or themes. They are named after the theme they represent, for example, population density maps, cervical cancer maps, agriculture maps, urban population maps, household maps, election maps, climate maps, environmental monitoring maps, risk maps, etc.

General or Base Maps

Base maps are elaborated using manual or computerized techniques and procedures such as on-site methods using a compass, and remote methods using photographic pictures, satellite images, and global positioning systems (GPS).

Aerial Photography:

This technique consists of taking pictures of the earth's surface using special photography devices mounted on aircraft. These images are used to create large-scale maps through photogrammetry techniques. The cost of this procedure is lower than that of direct mapping on terrain. Photographs are advantageous in that the images provide an exact rendition of the terrain as well as a global view of all the observable geographic features.

Photographic images have interpretative and metric characteristics. In the former, the image is interpreted considering hues, colors, and textures, allowing the identification of details and features of the region by analysis and synthesis. In the latter, since the pictures are taken at known altitudes, their scale is used to estimate real distances.

Photographs offer many possibilities; their interpretative characteristics may help identify contaminated water reservoirs, locate areas at risk of flooding, identify types of crops, etc. Their metric characteristics are particularly useful to trace access routes to health services, evaluate the extent of damage after natural disasters, etc.

Remote Sensing:

It consists of obtaining information from objects on earth without physically contacting them. It uses sensors (radar, aerial scanning, radiometers, etc.), that are installed on airplanes, aerial balloons, helicopters, or satellites, and data are stored digitally on magnetic media. Data are then processed according to various techniques to identify and categorize characteristics of the terrain. Remote sensing is expensive but has several advantages; images taken from satellites have great area coverage; also, it is possible to gather data beyond the earth's surface, from geological sources, ocean depths and temperatures. Moreover, events may be tracked through time using sequential image animation.

Global Positioning Systems (GPS):

They are radionavigation systems developed and operated by the U.S. Department of Defense. GPS allow users on land, sea, and air, to determine their three-dimensional position, speed, and time, 24 hours a day, under any climate conditions, and in any part of the world. Their precision and accuracy far exceed those of any other existing radionavigation system.

GPS include spatial, control and user components. The spatial component consists of 24 satellites encircling the world in six orbits that constantly monitor position and time data. The control component con-
sists of one control station located in Colorado Springs, U.S.A., five monitoring stations, and three terrestrial antennas placed at different locations around the world. These monitoring stations and antennas contact the GPS satellites, receive information from them, and relay it to the control station, where precise satellite orbits are calculated. This information is then used to feedback navigational instructions to each satellite, through the ground antennas. The user component consists of receptors, processors, and antennas to allow users on earth, air, or sea, gather GPS information.

Users may determine their position on earth using reference data from satellites that are constantly transmitting exact position and time signals. The user’s receptor measures the time it takes the signal to reach it. Simultaneous measures collected from four satellites are processed to provide the three dimensions of position, speed, and time. With the deployment of new satellites and the declassification of strategic information, GPS yield estimates within 10 m error ranges, which is precise enough for health applications in GIS.

Thematic Maps

Feature representation on thematic maps follows international standards to provide map designers with a universal way of communicating spatial phenomena and objects. Cartographic representation methods have been developed to visualize quantitative and qualitative information. Some of the most common methods that may be applied in public health are presented below. As a reminder, all thematic maps must include a brief and self-explanatory title (emphasizing time, place, and person components), a legend (a graphical synthesis of information on the theme, classification method, and representation symbols) and, the scale and orientation that were used (optional).

Cartogram Method

These maps are also known as choropleth or color-pattern maps (Figure 2.6). They indicate ranges of magnitude of an event within defined spaces. Magnitudes are expressed quantitatively, such as number of cases, mortality rates, essential needs index, etc. Ranges are expressed in different ways (quartiles, equal intervals, standard deviations, equal number of members, etc.) and the number of intervals may vary (usually no more than eight). Each geographical space includes a value within a range interval. The interval is represented using colors and patterns for differentiation. For example, the map shown in Figure 2.6 represents population density (population/area) by state in Venezuela. It shows that states with highest population density (dark colors), where more health resources would be targeted, are those in the North. Additional examples could show the proportion of the population covered by local health services, mortality rates due to accidents by region, rodent infestation by territorial unit, etc.

Figure 2.6: Cartogram Method. Population density in Venezuela

Dot Density Method

These maps represent the value of an event with a number of dots proportional to that value (Figure 2.7). Dots are randomly distributed across a territory or an area and each dot has a value of one or more units. Note that dots indicate a value within a region and not a specific place (geographical point). Occasionally, random grouping of dots may show a "clustering effect" (as in the case of epidemics) when actually there is none. To avoid this effect, the use of cartograms is recommended. The map in Figure 2.7 shows the distribution of total population across regions of Nicaragua. Each dot represents 10,000
people living within a region. It is readily apparent that people are concentrated in the Western regions of the country.

Figure 2.7: Dot density method. Population distribution in Nicaragua

Graduated Symbols Method
Maps created with this method use symbols of size proportional to the magnitude of the value of the event to be represented (thus the term "graduated"). Graduated symbol maps are useful to display geographic data referenced to one point, such as major city sizes in the Region (Figure 2.8). Different methods are utilized to determine the size of the symbols, such as quantiles, ranges of equal number of mem-

bers, logarithms, square roots, constant ranges, etc. Different types of symbols may be related to a specific event; for example, a hospital symbol may represent the number of primary health care units in the region, or a mosquito symbol may be used to represent malaria incidence rates by localities on a map.

Cartogram Method
These maps use diagrams to represent the magnitude of variables in series, by associating diagrams to each territorial unit (Figure 2.9). Cartograms are used to express the absolute or relative frequency of two or more variables in each geographical area. The most commonly used diagrams are bar and pie graphs. The example shown in the inset presents literacy levels of males and females and percentages of total population by regions of Ecuador. Other examples are the distribution of diarrheal disease cases by age by municipality; the distribution of AIDS cases by disease type by province, etc.

Figure 2.9: Cartogram method. Male and female literacy by province in Ecuador

Qualitative Background Method
These types of maps, also called individual value maps, represent events with a compact and uniform distribution in a territory, such as the major vegetation formations in Brazil (Figure 2.10). An all or nothing assumption is made for the presence or absence of the phenomenon in the geographical unit, and it is not possible to distinguish degrees of intensity within the unit. Thus, the information displayed using this
method has a qualitative or categorical nature. This method may be used to create maps of crops, maps of insect species relevant to health, etc.

**Figure 2.10: Qualitative background method. Types of vegetation in Brazil.**

![Vegetation Map](image)

**Diffusion Maps**

These maps are useful to locate and measure the spread of events and to identify their dynamics in place and time. The modeling of the availability of hospital beds in Honduras is an example of spread (Figure 2.11). Also, epidemics may be represented on maps displaying the spread of disease across timeseries, with color or pattern gradients that become more intense in areas of greater occurrence of cases.

**Figure 2.11: Diffusion map. Availability of hospital beds in Honduras.**

![Hospital Bed Availability Map](image)

A common way to represent these phenomena - to make empirical predictions - is through conventional vector models as a function of the susceptible population density as well as the dissemination pathways. This model allows the identification of the three main steps in the evolution of an epidemic: dissemination, saturation, and decline. Also, it is possible to estimate the direction, intensity, extent, and velocity of the event of interest.

**Topologic Maps**

These maps are not proportional to the actual terrain distance and areas; their purpose is to provide a comparative view of the order and continuity of a given event. Areas on these maps are proportional to the magnitude of the event, rather than to the surface where it occurs. In this regard, they may be similar to cartograms when the geographical unit is used to represent magnitude.

**Isoline Maps**

This method is particularly useful to display an event undergoing very uniform change and uninterrupted dissemination. Events are represented by isolines, which are lines that connect points of equal numerical value and should be drawn equidistantly at a preset distance.

These maps are created using a technique known as interpolation. It consists of a comparative analy-

**Figure 2.12: Isoline map. Digital elevation model**

![Digital Elevation Model](image)
sis to represent the probability of occurrence of an event in a defined time period. Examples include the representation of a terrain or surface relief through isolines of elevation values above sea level (isohypsal), maps of temperatures (isothermal), pluvial precipitation, concentration of air gases and dusts, time required to travel to a locality and health centers (isochronal), etc.

Some Geographic Information Systems offer tools or accessory modules capable of transforming isoline data values to generate digital elevation models (DEM) (Figure 2.12). This is accomplished through the calculation of trends and relations among points, using a pattern grid (contour). The grid is an interpolated and smoothed depiction of data. This generates a "continuous surface" of geographical information to visualize trends across areas on maps. Multiple data layers may be mathematically presented to explore and solve complex spatial relations.

Geographic Information Systems

Geographic Information Systems (GIS) consist of an organized set of computerized technology (hardware, software, digital geographic data, methods and techniques) and personnel, designed to capture, store, retrieve, manage, display, and analyze geographically referenced data. GIS integrate technology to support comprehensive decision making on problems occurring in a given geographical space.

Following the GIS model presented in Chapter One (Figure 1.3), GIS components are summarized in figures 2.13, 2.16, 2.17 and 2.22.

Input:

GIS handle two types of data: spatial and attribute data (figure 2.13). Spatial data, also called geographical or location data, are the graphical objects of a map, such as the political and territorial boundaries, roads, water networks, cities, health centers, disease occurrence sites, etc. Spatial data may have explicit geographical references such as latitude and longitude coordinates, or implicit references such as address zip code, census tract, forest area, identification number, etc. An automated GIS function called geocoding allows converting explicit geographic references from implicit references. These geographic references are used to locate characteristics and events on the Earth's surface for analysis.

Attribute data are variables that represent characteristics of spatial data and geographical entities, such as disease-specific medical records, the number of patients seen at a health care unit, children mortality rates in a city, etc. Attributes require a unique geographic reference component to relate to spatial data; that is, they require sharing a variable with the name or code (identifier) of the region, city, health care unit, or place of occurrence of a specific event. The chapter on Relational Database Systems and Geographic Information Systems presents the properties and management of attributes in greater detail.

Other maps

Other graphical elements represented on maps are areas or point locations. Examples of defined areas or boundaries include maps of forest areas, maps of water supplies, maps of flooding areas, etc. The objects on specific map locations may be represented by symbols related to health services, access roads, water supplies, fire stations, etc.

Maps in traditional format (on paper) have several restrictions for handling, displaying and analyzing information: they are static, difficult to manage and update, their elaboration is time-consuming, the simultaneous display of multiple variables is sometimes not possible, the work areas are limited to the paper size and scale, and their construction requires some expertise. Today, Geographic Information Systems digital formats for automated processing make map construction more efficient.
The two most common formats utilized to represent spatial entities are: vector and raster.

The vector format consists of strings of coordinates and uses three types of graphical elements to represent the geographical objects of a map: points (nodes), lines (segments), and areas (polygons) (Figure 2.14). A point is represented by a pair of x,y, Cartesian coordinates, usually geographically referenced; i.e., referenced to a real location on the Earth's surface, such as a city on the map of a country. A line is represented by a set of x,y, coordinate pairs, linked sequentially by a line, forming segments (to-
pology); for example, it may represent a road between two cities, a river, a street, etc. An area is represented by a set of pairs of x, y coordinates linked in a similar fashion as lines, but in this case the last pairs of coordinates are joined to the first pair, thus closing a polygon that encloses a surface (not only its perimeter). Examples of areas are a state on a country map, the types of land use in a region, flooding areas along riversides, etc. Both in lines and areas, each pair of coordinates forms a node, joined by lines to make up the boundaries of vector objects. The vector format may be generated through digitizing or through global geopositioning (using GPS receivers) processes.

The raster or grid format stores spatial data on a matrix generated by dividing a graphical image into a regular grid of cells (Figure 2.15). Each element of the matrix stores an attribute that identifies each cell (or pixel). The position of the cells in the matrix provides information on the location of spatial data. Moreover, each matrix cell stores additional information on the attributes of individual spatial data. The raster format may be generated through digitizing maps, aerial photographs, or digital information obtained from satellite remote receivers. Examples of raster formats are satellite images and digital photographs. It should be noted that raster images may be transformed into vector formats.

![Figure 2.15: Raster format](image)

The vector format is commonly used to describe discrete entities, such as a state or a municipality of a country, and less used to describe continuous entities, such as soil types, access to a health care unit, etc. (see the glossary at the end of the chapter on Epidemiology for definitions of types of variables). The raster format is commonly used for continuous variables.

Attribute data are variable values or variable categories associated to a setting. They may come from data sources already available (census registries, public health surveillance registries, vital statistics, surveys, etc.), or from newly collected data registries (surveys, field studies, etc.). These data are entered into GIS through data capturing and importing. Capturing may be carried out through keyboard entry or through automated devices such as optical readers or scanners. Importing data involves conversion of data from other digital sources such as database, spreadsheet, text-delimited files (ASCII), etc., from either local or remote systems. Access to remote systems may be through local networks or by modem. One of the most commonly used remote systems is the Internet.

**Storage:**

Spatial and attribute data are commonly stored using relational database models available as part of the database management system included in GIS software. These models are widely used to store data as sets of tables that share fields with different tables, allowing their linkage (Figure 2.16). Database management systems keep spatial and attribute data unambiguously linked. Data may be stored either in magnetic media (diskettes or tapes), or in optical media (CD-ROM).

**Processing:**

Data for a specific project or purpose may require modifications to make them compatible with GIS (figure 2.17). Modification of data structure is a process to achieve compatibility, through data conversion to a compatible format. For example, if the identifier is a string variable, and the variable of the
attribute table that will relate this table to the cartographic database is numeric, then this numeric variable will have to be converted into a string variable. Also, this process may be used to add new variables and to select the variables that should be indexed, to make the linkage of tables more efficient.

Occasionally, it is necessary to process simple variables to generate a more complex one; for example, a variable containing mortality rates by geographical areas would be generated from variables corresponding to deaths (numerator) and population (denominator), multiplied by a constant. Complex variables such as the basic needs index (BNI) may be generated using multiple variables and mathematical operators; it requires calculating the mean and standard deviation of each of the variables conforming the index as well as the values of normalized standard deviations with respect to the expected values.

The standardized values of each variable are added to generate the BNI variable.

After entering spatial data it may be necessary to modify or edit the cartogram, to correct errors, or to update changes in the political and organizational composition, as well as regional changes due to natural phenomena. In addition, it may be necessary to edit cartographic databases to combine or divide territories into new units for administrative or analytical purposes, a process known as redistricting.

GIS data query functions allow the retrieval of spatial and attribute data, using either spatial data retrieval tools or Structured Query Language (SQL). SQL is a standard programming language to conduct queries within relational database systems. This subject is reviewed in detail in the chapter Relational Database Systems in Geographic Information Systems.
Figure 2.17: GIS operations framework: Processing

- Input
  - Storage
    - Manipulation
      - Modification of data structure
        - Transformation of data types
        - Generation of variables
        - Indices of Tables
      - Data Editing
        - Generation of complex variables
        - Modification of cartography
      - Selection
        - Spatial
        - By criteria
    - Data queries
      - Generation of new tables
        - Virtual
        - Permanent
      - Grouping
        - Determination of frequency
        - Calculation of summary measures
      - Overlay
        - Integration of multiple variables
      - Analysis
        - Determination of buffer zones
        - Distance measurement
      - Distance
      - Proximity
        - Identification of clusters
        - Interpolation
  - Output
Queries using spatial tools may be performed directly on maps. For example, in an area with greater incidence of disease, additional information on potential determinants (population, socioeconomic level, etc.) may be obtained from data stored in tables associated to that particular map (Figure 2.18). SQL queries use criteria corresponding to analytical procedures (described below). For example, to identify areas with greater health risks, a query is made to select areas having values greater than or lower than a given parameter (criterion), such as incidence values above the country mean incidence or mean BNI values. These criteria may be combined using logic operators such as "and", "or", and "not" (described in detail in the Appendix).

SQL queries generate virtual output tables stored temporarily with necessary table data in the computer RAM space; they may be saved as permanent tables for later use. Summarized information from geographical units in a table may be obtained through GIS SQL data aggregation functions. Summary measures include frequencies, sums, minimum and maximum values, means, standard deviations, etc.

Data analysis, one of the basic functions of GIS, includes spatial analysis and attribute analysis. Due to their importance for health applications, emphasis is made on three spatial analysis concepts: overlay, distance, and proximity.

Attribute analysis is particularly useful to answer simple questions, such as "Which municipalities have higher gross national product (GNP)?", or more complex ones, such as "Which municipalities have higher GNP, more than 5,000 inhabitants, and infant mortality rates two standard deviations above the mean for the country?" It is also possible to do a combined analysis of spatial and attribute data, to answer more complex questions, such as "How many inhabitants are located within 11 km. around a health center, in communities with an annual parasite index (API) of Plasmodium falciparum malaria greater than 10 cases per 1,000 inhabitants, having at least one malaria episode during the year, and living in households with no mosquito screens or bednets?"

GIS data are organized in information layers or coverages, geographically referenced with spatial and attribute elements. Each of these layers represents a theme with elements, which may be base or thematic. For example, base elements may be roads, political boundaries, and location of health services. Thematic elements are, for example, children mortality rates by geographic unit, vegetation types, and population density. The organization of a collection of overlapped geographically referenced themes makes up a map (Figure 2.19).

GIS layer organization capabilities greatly simplify spatial analysis. Intersecting functions further enable the integration of multiple variables from different layers to generate a synthesis (as a new layer formed by overlapping others). This is accomplished mainly through two procedures: inclusion and overlaying (Figure 2.20).

Inclusion identifies an object from a layer "A" that includes (entirely) an object from a layer "B" (or vice versa). For example, consider a layer with the location of mosquito breeding sites and flying ranges.
(buffer zones) and another layer with the location of households in a community; inclusion functions would allow the identification of households at higher risk of mosquito infestation, i.e., the households within the flying range of mosquitoes.

Overlying combines objects not entirely contained within two layers, to create a new layer. Overlaying operations are junctions and intersections. Junctions combine objects from two layers and create a new layer that contains them. For example, consider a layer with information on the extent and distribution of measles vaccination coverage by private healthcare providers and a similar layer for public healthcare providers; junction operations could create a new layer for total vaccination coverage. In contrast, intersection would “join” only those objects or portions of them that share the same geographical reference in both layers. For example, the assessment of risk for lung cancer in a region where different risk factors are present (asbestos pollution and plutonium radioactivity), can be carried out by intersecting a layer of asbestos exposure levels with a layer of plutonium radiation levels. A new layer containing both risk factors would be created, where it would be possible to analyze exposure interactions and independent effects. Epidemiologic analysis would yield the frequency of disease for each separate exposure as well as the frequency of combined exposures. If the observed frequencies are greater than the expected frequencies, a higher risk is present due to interaction.

**Figure 2.20: Intersection process**

GIS distance analysis tools include functions to determine buffer zones around an object and distance measurements between two or more points or objects (figure 2.21). A buffer zone of an object is calculated considering that an object A is within or beyond a distance D from object B. This will yield the buffer zone around object B; all boundary points of this zone have equal distances D from B. If object B is a point, the resulting buffer zone will be a circle with a distance D radius. If object B is a straight line, the buffer zone will be a rectangle with rounded corners. If object B is an irregular line or a polygon, the buffer zone will be an enlarged version of object B, with distance D from its boundary to the boundary of B. As an example, a study conducted in Brazil shows buffer zones delimited within a 0.5 km radius around health disease cases among schoolchildren. Schools within that radius were further studied to identify more cases and determine risk factors for disease.
The measurement of Euclidean distance is the shortest distance between two points (A and B) on a plane and is useful to determine distances between points of interest; for example, the actual distance in kilometers (or other measurement units) between two communities.

The spatial distribution of events or objects may have homogeneous or heterogeneous patterns that may suggest spatial aggregation, defined as the proximity of two or more object in time and/or space within a relatively small area, forming groups or clusters (as occurs in epidemics). For public health purposes, it is important to determine whether the presence of temporal-spatial aggregation is due to chance or to a risk factor associated with the cluster. This may be further tested using statistical methods, such as spatial auto-correlation (Chou, 1997) to assess the spatial configuration of patterns and the probability of clustering by chance.

GIS proximity assessment tools are useful to estimate and predict values of continuous measurements by interpolation, an operation that transforms discrete data into continuous data to compare spatial patterns of discrete variables with those from other entities; in other words, available data for some points are used to impute likely values for intermediate points or locations. Different interpolation methods are available; their description is quite complex and beyond the scope of this document. A brief review of interpolation methods may be found in the Glossary (for example, see “Kriging”). Interpolation is useful to elaborate contour or isoline maps.

**Output:**

GIS generate high-quality products, including maps (with graphical components such as legends, scale, orientation, text, etc.), graphs (bar, line, pie, scatterplot, etc.), and tables (reports) (figure 2.22). The main output products of GIS are maps of different types and content. The quality of maps depends on geographic precision, color and pattern representation of objects, and display definition of objects and events. The layout and final presentation of results depend on the purpose and information to be conveyed. It should be noted that the same cartograph-
ic data may be represented in different versions (color, pattern, classification method, etc.) and settings, to emphasize characteristics and display data that will convey different information and lead to different conclusions. Recommendations to organize data presentation for the visual display of quantitative data are available (Tufte, 1997). Presentation formats have greater impact when they combine maps, graphs and tables.

GIS make such integration possible and offer additional enhancements, such as three-dimensional maps, that provide further insight of data. Some of these tools are presented in the Appendix.

GIS outcomes may also be organized to be analyzed and displayed in electronic formats and not only in paper, like in a situation room. Possibilities range from stand-alone presentations to client-server applications that can be distributed over the Internet. An Atlas of Health Inequalities in the Americas represents a stand-alone application that is prepared using HTML tools. The PAHO Core Health Data GIS is an example of such client-server applications. Both may be seen on PAHO’s Web page (http://www.paho.org)
GIS software programs of different types and complexity are available for public health applications. Selecting a particular program depends on the purpose and available resources (equipment, financial, and personnel). Table 2.1 shows the recommendations issued by SHA/PAHO to select a particular GIS for local, regional, and central levels.

One of the most commonly used programs is EpiMap 2.0, developed by the United States Centers for Disease Control and Prevention (CDC), in collaboration with the World Health Organization (WHO). EpiMap is a simple and user-friendly program that allows creating cartographic bases without additional special equipment (i.e. a digitizer tablet). In addition, EpiMap allows the user to create simple databases and a variety of thematic maps quickly and with modest computer equipment. New hypertext tools allow the linkage of maps, tables, and text generated by EpiMap, with EpiInfo, a software program for epidemiologic data processing and analysis. Furthermore, compatibility is extended to several other data sources and formats, such as dBase (DBF), Lotus 1-2-3 (WK*), etc. EpiMap's flexibility, low computer requirements, and user-friendliness make it one of the most attractive tools for GIS applications.

Recently, WHO and the United Nations Children’s Fund (UNICEF) started a program on data management and mapping for public health (HealthMap). As part of their activities, they developed a new information and mapping tool to support planning and decision-making at both micro and macro levels, called HealthMapper 1.0. This is a GIS software comprising three components: a standardized geographic database, a mapping interface and a data manager. HealthMapper simplifies the use of geographic information systems and mapping, and provides a user-friendly interface for public health data analysis (WHO, 1999).

### Table 2.1: PAHO’s recommendations for resources required by GIS applied to health, for different health system levels.

<table>
<thead>
<tr>
<th>Requirements to set up GIS at each level</th>
<th>Level of organizational development of health services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local (Basic)</td>
</tr>
<tr>
<td>GIS software</td>
<td>EpiMap 2.0, EpiMap 2000</td>
</tr>
<tr>
<td></td>
<td>HealthMapper 1.0</td>
</tr>
<tr>
<td>Cost</td>
<td>Free distribution</td>
</tr>
<tr>
<td>PC computer equipment</td>
<td>512K to RAM, 16MB</td>
</tr>
<tr>
<td>Operating system</td>
<td>MS-DOS, Windows</td>
</tr>
<tr>
<td>Type of training</td>
<td>Simple, self-training</td>
</tr>
<tr>
<td>Time of training</td>
<td>Days</td>
</tr>
</tbody>
</table>

* Distributed by the HealthMap program at no charge
In response to the needs of public health managers regarding an affordable appropriate tool to carry out epidemiological analysis more efficiently, SHA/PAHO has developed the SIG-Epi software (Figure 2.23). This software introduces simplified methods and procedures for analysis in public health, including GIS functionality coupled with analytic epidemiological and statistical capacities. In this regards, it is the first to include spatial analysis methods and techniques oriented to health not yet available in the most frequently used commercial software.

SIG-Epi was developed as a result of many workshops and consultations with PAHO’s GIS Collaborating groups on SIG-Epi and other health professionals and experts. However, its intended target audience is epidemiologists, decision makers, policy makers in health at the national, regional, community or local levels, who have limited access to commercial GIS software and require simplified procedures and methods for analysis.

SIG-Epi features basic GIS functionality for spatial and attribute data storage, handling and processing and more specific epidemiological analytical procedures. These include: description of frequency distributions; summarized statistic measures; box plot graphs and maps; single and multiple linear regression; calculation, standardization and spatial smoothing of rates; identification of priority areas; calculation of compound health indexes; tests for time-space clusters of cases; analysis of association between exposure-effects in epidemiological studies based on individual data; exploratory spatial data analysis; global and local indexes of spatial autocorrelation; index and significance index maps; Moran’s scatterplot and scatterplot map; and, spatial lag map, among others.

SIG-Epi is a platform independent GIS software. The version with basic capacities will be distributed free of charge. It will be a Multilanguage software in Spanish, English, Portuguese and French.

For those interested in further details on the various software developed by CDC, WHO and PAHO, a recent paper comparing functionality features has been prepared (Schenck-Yglesias et al., in press). It is also important to indicate that recently, different institutions, such as WHO, CDC, Health Canada, ATSDR, and ESRI, and PAHO initiated activities in order to coordinate the development of software, training materials, applications, methods, etc. The proceeding from the first meeting on GIS in Public Health, held in Atlanta, Georgia in October, 2000, will be available soon. A CD-ROM® has been prepared (Schenck-Yglesias, 2000) and other formats (paper and through the Web of the respective agencies) will follow.

GIS programs useful for epidemiologic analysis of spatial data include MapInfo®, ArcView®, MapTitude®, ArcInfo®, GIS+, GRASS®, IDRISI®. Most of them are powerful commercial software
packages that require considerable computer resources. Some of them are designed for Windows and in general, they require Windows 3.1 or higher, a minimum of 16 Mb Ram, (64 Mb recommended), hard disk space of 40 Mb or more, additional storage capacity for cartographic bases and databases, and a VGA or higher resolution video card.

These data management systems offer a wide variety of symbols, greater analytical possibilities, high quality tables and graphs associated to maps, and high definition and precision output.

Cartographic bases should be as precise as possible, with referenced geographic coordinates (latitude and longitude) of vertices and geographical objects. High quality data may be obtained from national and international organizations that handle large databases, including data on censuses, cadastres, morbidity and mortality, population, climate, soils, vegetation, resources, living conditions etc. Academic, research and other institutions are also good sources of quality data for making high quality maps.

In summary, the ability of GIS to manage spatial data allows the creation of diverse "scenarios", an indispensable function for strategic analysis to inform decision making. Spatial analysis tasks should be performed by multi-disciplinary teams to decrease bias and uncertainty of results. The ultimate goal is to expand the capabilities for action, diagnosis, and planning of health systems at their different organization levels.

Figure 2.23: SIG-Epi Software developed by PAHO
Glossary of selected geographic and cartographic terms

Accessibility. An aggregate measure of how reachable locations are from a given location. It is a function of the distance between locations and an empirically derived distance decay parameter.

Accuracy. Degree of conformity with a standard. Accuracy relates to the quality of a result and is distinguished from precision which relates to the quality of the operation by which the result is obtained.

Adjustment. Process designed to remove inconsistencies in measured or computed quantities by applying derived corrections to compensate for random or accidental errors.

Aerial mosaic. Assembly of aerial photographs whose edges usually have been torn or cut selectively and matched to the imagery on adjoining photographs to form a continuous representation of a portion of the Earth’s surface.

Aerial photograph. A photograph, usually taken from an airplane, as a means of remotely recording ground level events. Not to be confused with satellite remote sensing which produces digital images, aerial photography provides black and white, color and infrared photographs on film. These can be taken at either vertical or oblique angles, depending on the phenomenon in question and the desired application. Aerial photography differs from satellite imagery in that the results are almost instantaneous and require only developing, as opposed to images which must undergo a great deal of processing before electromagnetic signals resemble real world features.

Algorithm. A set of rules for solving a problem that must be specified before the rules can be written in a computer language.

Altimeter. Instrument for measuring altitudes or elevations with respect to a reference level, usually mean sea level. The most common type is an aneroid barometer. A radar altimeter determines the height of an aircraft above the terrain by measuring the time required for an electromagnetic pulse to travel from aircraft to the ground and back.

Application. A process (or a series of processes) that uses data or performs some function on a computer system.

Areal interpolation. Methods for determining data values at all points from values at a sample of points. In the software, Spherekit, for example, the available interpolation methods are: Inverse distance weighting; Triangulation; Kriging; Multiquadric; and Thin plate spline.

Attribute. A characteristic of a geographic feature described by numbers, characters, images and CAD drawings, typically stored in tabular format and linked to the feature by a user-assigned identifier (e.g., the attributes of a well might include depth and gallons per minute). A column in a database table.

Autocorrelation. Is a term referring to the degree of relationship that exists between two or more spatial variables, such that when one changes, the other(s) also change. This change can either be in the same direction, which is a positive autocorrelation, or in the opposite direction, which is a negative autocorrelation. For example, soil type and vegetation may be highly correlated, either positively or negatively depending upon the type of soil and vegetation under examination. Popular measures of spatial autocorrelation with spatial analysis are Moran’s I and Geary’s C

Azimuth. Horizontal direction reckoned clockwise from the meridian plane.

Band. One layer of a multispectral image representing data values for a specific range of the electromagnetic spectrum of reflected light or heat (e.g., ultraviolet, blue, green, red, near-infrared, infrared, thermal, radar, etc.). Also, other user-specified values derived by manipulation of original image bands. A standard color display of a multispectral image shows three bands, one each for red, green and blue. Satellite imagery such as LANDSAT TM and SPOT provide multispectral images of the Earth, some containing seven or more bands.

Bathymetry. Science of measuring water depths (usually in the ocean) to determine bottom topography.

Buffer. A zone of a specified distance around coverage features. Both constant- and variable-width buffers can be generated for a set of coverage features based on each feature’s attribute values. The resulting buffer zones form polygons-areas that are either inside or outside the specified buffer distance from each feature. Buffers are useful for proximity analysis (e.g., find all stream segments within 300 feet of a proposed logging area).

Cadastral survey. Survey relating to land boundaries, made to create units suitable for title transfer or to define the limitations of title. Derived from “cadastre” meaning a register of land quantities, values, and ownership used levying taxes, the term may properly be applied to surveys of a similar nature outside the public lands, such surveys are more commonly called “land surveys” or “property surveys.”

Cartography. Science and art of making maps and charts. The term may be taken broadly as comprising all the steps needed to produce a map: planning, aerial photography,
field surveys, photogrammetry, editing, color separation, and multicolor printing. Mapmakers, however, tend to limit use of the term to the map-finishing operations, in which the master manuscript is edited and color separation plates are prepared for lithographic printing.

**Census block.** The smallest geographical area, bounded by visible boundaries, for which census data are collected.

**Census tract.** A small, permanent statistical subdivision of an area, normally with homogeneous characteristics.

**Centroid.** The term given to the center of an area, region, or polygon. In the case of irregularly shaped polygons, the centroid is derived mathematically and is weighted to approximate a sort of “center of gravity.” Centroids are important in GIS because these discrete X-Y locations are often used to index or reference the polygon within which they are located. Sometimes attribute information is “attached,” “hung,” or “hooked” to the centroid location.

**Chart.** Special-purpose map designed for navigation or to present specific data or information. The term “chart” is applied chiefly to maps made primarily for nautical and aeronautical navigation, and to maps of the heavens, although the term is sometimes used to describe other special-purpose maps.

**Contour.** Imaginary line on ground, all points of which are at the same elevation above or below a specific datum.

**Contour interval.** Difference in elevation between two adjacent contours.

**Coordinates.** Linear and (or) angular quantities that designate the position of a point in relation to a given reference frame. Pairs of numbers expressing horizontal distances along orthogonal axes, or triplets of numbers measuring horizontal and vertical distances, or n-numbers along n-axes expressing a precise location in n-dimensional space. Co-ordinates generally represent locations on the earth’s surface relative to other locations.

**Coordinate system.** A recognized reference system for the unique location of a point in space. The Cartesian coordinate system and the system of latitude and longitude of the earth are examples of co-ordinate systems based upon Euclidean geometry.

**Depth curve.** Line on a map or chart connecting points of equal depth below the datum.

**Digital elevation model (DEM).** A digital representation of a continuous variable over a two-dimensional surface by a regular array of z values referenced to a common datum. Digital elevation models are typically used to represent terrain relief. Also referred to as ‘digital terrain model’ (DTM).

**Digitization.** Conversion of map data from graphic to digital form.

**Elevation.** Vertical distance of a point above or below a reference surface or datum.

**Exploratory data analysis (EDA).** An approach used in the analysis of data, in which as little prior structure as possible is used to expose patterns within the data. Particularly atypical data is then further explored in order to investigate the underlying relationships within the data. EDA techniques recognize the relationship between computational displays and human cognition, and thus allow the user to manipulate views of the data analysis, through such means as histograms, or box plots, in order to further investigate perceived patterns. Exploratory spatial data analysis (ESDA) is a subset of EDA.

**Geocode.** The process of identifying the coordinates of a location given its address. For example, an address can be matched against a TIGER street network to determine the location of a home. Also referred to as address geocoding.

**Geographic data.** The locations and descriptions of geographic features. The composite of spatial data and descriptive data.

**Geographic database.** A collection of spatial data and related descriptive data organized for efficient storage and retrieval by many users. Geographic data sets include coverages, grids, DBMS tables, tins, images, lattices, and CAD drawings.

**Geographic feature.** A user-defined geographic phenomenon that can be modeled or represented using geographic data. Examples of geographic features include streets, sewer lines, manhole covers, accidents, lot lines, and parcels.

**Geographic information system (GIS).** An organized collection of computer hardware, software, geographic data, and personnel designed to efficiently capture, store, update, manipulate, analyze, and display all forms of geographically referenced information.

**Georeference.** To establish the relationship between page coordinates on a planar map and known real-world coordinates.

**Georelational model.** A geographic data model that represents geographic features as an interrelated set of spatial and descriptive data.

**Global positioning system.** A system of satellites and receiving devices used to compute positions on the Earth with a high degree of accuracy given a suitable GPS receiver. GPS is used in navigation, and its precision supports cadastral surveying. The network of satellites is owned by the US Department of Defense, and as such,
the accuracy of the signal is intentionally degraded for non-US military users.

**Graticule.** Network of parallels and meridians on a map or chart. System of coordinates of latitude and longitude used to define the position of a point on the surface of the Earth with respect to the reference spheroid.

**Grid.** Network of uniformly spaced parallel lines intersecting at right angles. When superimposed on a map, it usually carries the name of the projection used for the map—that is, Lambert grid, transverse Mercator grid, universal transverse Mercator grid.

**Grid cell.** A two dimensional object that represents an element of a regular or nearly regular tessellation of a surface. The term grid cell is often used to refer to a single element of a raster data structure.

**Hydrography.** Science that deals with the measurement and description of the physical features of the oceans, seas, lakes, rivers, and their adjoining coastal areas, with particular reference to their use for navigation.

**Hydrology.** Scientific study of the waters of the Earth, especially with relation to the effects of precipitation and evaporation upon the occurrence and character of ground water.

**Hypsography.** Topography referred to the national geodetic vertical datum of 1929. The science or art of describing heights of land surfaces with reference to this datum.

**Hypsometry.** Science or art of determining terrain relief, by any method.

**Imagery.** Visible representation of objects and (or) phenomena as sensed or detected by cameras, infrared and multispectral scanners, radar, and photometers. Recording may be on photographic emulsion (directly as in a camera or indirectly after being first recorded on magnetic tape as an electrical signal) or on magnetic tape for subsequent conversion and display on a cathode ray tube.

**Infrared scanner (thermal mapper).** Instrument that detects infrared radiation and converts the detected energy to an electrical signal for recording on photographic film or magnetic tape.

**Interpolate.** To estimate the value of an attribute at an unsampled point from measurements made at surrounding sites.

**Isoline.** A line on a surface connecting points of equal value for any of the characteristics used in the representation of the surface.

**Kriging.** A method for estimating the prevalence of a variable of interest at a given place using data from the surrounding region that incorporates the spatial structure of the variable. ‘Ordinary kriging’ is based on ‘the intrinsic hypothesis’ which states that the difference in value of variables between two positions depends only on the distance between them. The measure of this distance dependence is the “semivariogram function”. A key property of kriging is that the semivariogram function can be used to estimate the value of the process at unrecorded places from the neighboring sampling values. The local variance is known as ‘the nugget variance;’ ‘the sill’ represents the degree of spatial autocorrelation; and ‘the range of influence’ represents the distance over which the autocorrelation is found to extend. Generally, a map is obtained by estimating the value at each node of a regular grid superimposed over the area of interest and then applying a contouring program to draw isolevel curves.

**Landsat.** A series of satellites that produce images of the earth. The Landsat remote sensing satellite program was developed by NASA (National Aeronautics and Space Administration). Landsat data are provided in .BIL (band interleaved by line) or .BIP (band interleaved by pixel) formats.

**Land use classification system.** Coding system of categories and subcategories designed for use on a map to designate land or water use.

**Latitude.** Angular distance, in degrees, minutes, and seconds of a point north or south of the Equator.

**Linear feature.** A geographic feature that can be represented by a line or set of lines. For example, rivers, roads within a health services delivery area, and electric and telecommunication networks are all linear features.

**Longitude.** Angular distance, in degrees, minutes, and seconds, of a point east or west of the Greenwich meridian.

**Map.** Graphic representation of the physical features (natural, artificial, or both) of a part or the whole of the Earth’s surface, by means of signs and symbols or photographic imagery, at an established scale, on a specified projection, and with the means of orientation indicated. There are several, more common, types of maps.

1. **Base map.** Map on which information may be placed for purposes of comparison or geographical correlation. The term “base map” was at one time applied to a class of maps now known as outline maps. It may be applied to topographic maps, also termed “mother maps” that are used in the construction of other types of maps by the addition of particular data.

2. **Bathymetric map.** Maps delineating the form of the bottom of a body of water, or a portion thereof, by the use of depth contours (isobaths).
3. **Cadastral map.** Map showing the boundaries of subdivisions of land, often with the bearings and lengths thereof and the areas of individual tracts, for purposes of describing and recording ownership. It may also show culture, drainage, and other features relating to land use and value.

4. **Choropleth map.** Thematic map in which areas are colored, shaded, dotted, or hatched to create darker or lighter areas in proportion to the density of distribution of the theme subject.

5. **Hypsometric map.** Map showing relief by any convention, such as contours, hachures, shading, or tinting.

**Map Overlay.** The superimposition of two or more geographic data layers and the ability to make boolean queries with respect to the attributes of polygons.

**Map projection.** Orderly system of lines on a plane representing a corresponding system of imaginary lines on an adopted terrestrial or celestial datum surface. Also, the mathematical concept for such a system. For maps of the Earth, a projection consists of 1) a grid of lines representing parallels of latitude and meridians of longitude or 2) a grid.

**Map query.** The process of selecting information from a GIS by asking spatial or logical questions of the geographic data. Spatial query is the process of selecting features based on location or spatial relationship (e.g., select all features within 300 feet of another; point at a set of features to select them). Logical query is the process of selecting features whose attributes meet specific logical criteria (e.g., select all polygons whose value for AREA is greater than 10,000 or select all streets whose name is ‘Main St.’). Once selected, additional operations can be performed, such as drawing them, listing their attributes or summarizing attribute values.

**Map scale.** The reduction needed to display a representation of the Earth’s surface on a map. A statement of a measure on the map and the equivalent measure on the Earth’s surface, often expressed as a representative fraction of distance, such as 1:24,000 (one unit of distance on the map represents 24,000 of the same units of distance on the Earth). Map scale can also be expressed as a statement of equivalence using different units; for example, 1 inch = 1 mile or 1 inch = 2,000 feet.

**Map units.** The coordinate units in which a geographic data set (e.g., a coverage) is stored in a geographic information system. Map units can be inches, centimeters, feet, meters, or decimal degrees.

**Meridian.** Great circle on the surface of the Earth passing through the geographical poles and any given point on the Earth’s surface. All points on a given meridian have the same longitude.

**Metric system.** Decimal system of weights and measures based on the meter as a unit length and the kilogram as a unit mass.

**Model.** A representation of reality used to simulate a process, understand a situation, predict an outcome, or analyze a problem. A model is structured as a set of rules and procedures, including spatial modeling tools available in a geographic information system.

**Multispectral scanner.** Device for sensing radian energy in several channels of the electromagnetic spectrum.

**Overlay.** The process of superimposing two or more maps, through registration to a common coordinate system, such that the resultant maps contain the data from both maps for selected features.

**Orientation.** Establishing correct relationship in direction with reference to points of the compass; the state of being in correct relationship in direction with reference to the points of the compass.

**Orthophotograph.** Photograph having the properties of an orthographic projection. It is derived from a conventional perspective photograph by simple or differential rectification so that image displacements caused by camera tilt and terrain relief are removed.

**Parallel of latitude.** A circle, or approximation of a circle, on the surface of the Earth, parallel to the Equator, and connecting points of equal latitude; a circle of the celestial sphere parallel to the ecliptic, and connecting points of equal celestial latitude.

**Photogrammetry.** Science or art of obtaining reliable measurements or information from photographs or other sensing systems.

**Pixel.** A contraction of the words picture element. Pixel refers to the smallest unit of information available in an image or raster map.

**Precision.** In geographic information systems, it refers to the number of significant digits used to store numbers, and in particular, coordinate values. Precision is important for accurate feature representation, analysis and mapping.

**Raster.** A cellular data structure composed of rows and columns for storing images. Groups of cells with the same value represent features.

**Raster data.** An abstraction of the real world where spatial data is expressed as a matrix of cells or pixels, with spatial position implicit in the ordering of the pixels. With the raster data model, spatial data is not continuous but divided into discrete units. This makes raster data partic-
Geographic Information Systems in Health

particularly suitable for certain types of spatial operation, for example overlays or area calculations. Unlike vector data however, there are no implicit topological relationships.

**Relief.** Elevations and depressions of the land or sea bottom.

**Remote sensing.** Process of detecting and (or) monitoring chemical or physical properties of an area by measuring its reflected and emitted radiation.

**Resolution.** Resolution is the accuracy at which a given map scale can depict the location and shape of geographic features. The larger the map scale, the higher the possible resolution. As map scale decreases, resolution diminishes and feature boundaries must be smoothed, simplified, or not shown at all. For example, small areas may have to be represented as points.

**Satellite image.** A picture of the earth taken from an earth-orbital satellite. Satellite images may be produced photographically or by on-board scanners.

**Scale.** Relationship existing between a distance on a map, chart, or photograph and the corresponding distance on the Earth.

**Semivariogram.** A figure relating the variance of the difference in value of an attribute at pairs of sample points to separation distance.

**Small Area Variation.** The examination of differences in disease rates in small administrative areas. Used for the purposes of health service planning, disease surveillance and the identification of areas of poor health.

**Smoothing.** A technique used to remove or reduce local noise or high frequency signal within spatial data, and therefore reveal the global pattern or trend. Smoothing is a technique used in both spatial analysis and digital image processing using a variety of methods, but normally based upon a matrix or filter that passes over the image. Certain interpolation techniques, particularly the approximate interpolation methods, are also used to for smoothing images.

**Space-time clustering.** Clustering in time and space is a marker for contagion. Tests for space-time clustering measure the proximity of case pairs in space and time. The paucity of good data, cross-classified by time and space, has hampered the development of methods in this area.

**Spatial analysis.** Analytical techniques associated with the study of locations of geographic phenomena together with their spatial dimensions and their associated attributes. Spatial analysis is useful for evaluating suitability, for estimating and predicting, and for interpreting and understanding the location and distribution of geographic features and phenomena. Spatial analysis is the process of modeling, examining, and interpreting model results. There are four traditional types of spatial analysis: topological overlay and contiguity analysis, surface analysis, linear analysis, and raster analysis.

**Spatial data.** Information about the location and shape of, and relationships among, geographic features, usually stored as coordinates and topology.

**Spatial data models.** Systems for encoding, storing and manipulating spatial data. There are two kinds of spatial data models: field models and object models.

**Spatial Filters.** The geographical distribution of cases of a disease can be generalized at different spatial scales by computing rates for different sized areas. Different terms have been given to this process. Kernel density estimation; spatial smoothing. Unlike ground temperature or rainfall, both of which are continuous in space, the number of new cases of a disease is not measurable at any location, since all of the cases in an area are artificially gathered at one point for statistical and/or administrative purposes. The resulting densities at any point on the surface are usually interpreted to mean the expected rate that would be observed at that point if one were to collect information around the point for a sufficient length of time for the density to be estimated correctly. Such maps have many advantages in comparison with mapping methods that provide an indication of the level of a disease by area. They are not constrained by the borders of geographic units, and sudden transitions between levels of two neighboring areas are avoided. Most commonly, spatial filters are equal in area but in some applications the filters have equal numbers of people at risk. Spatial filters may have irregular boundaries when the data within them are from aggregated zones whose centroids lie within the filter area.

**Spot elevation.** Point on a map or chart whose height above a specified datum is noted, usually by a dot or a small sawbuck and elevation value. Elevations are shown, on a selective basis, for road forks and intersections, grade crossings summit of hills, mountain

**Standard mortality rate.** The ratio of disease count observed in an area to that expected, based on the age and sex structure of the area and the age- and sex-specific death-rates of a standard population.

**Surface.** A geographic phenomenon represented as a set of continuous data, such as elevation or air temperature over an area. Surfaces can be represented by models built from regularly or irregularly spaced sample points on the surface.

**Surface model.** Digital abstraction or approximation of a surface. Because a surface contains an infinite number of points, some subset of points must be used to represent
the surface. Each model contains a formalised data structure, rules, and X,Y,Z point measurements that can be used to represent a surface, for example, see TIN.

**Table.** A set of data elements that has a horizontal dimension (rows) and a vertical dimension (columns) in a relational database system. A table has a specified number of columns but can have any number of rows. A table is often called a relation. Rows stored in a table are structurally equivalent to records from flat files in that they must not contain repeating fields.

**Tessellation.** The subdivision of a 2-dimensional plane (or 3-dimensional volume) into polygonal tiles (polyhedral blocks) that completely cover a plane or volume. Planes or volumes can be either divided into regular or irregular tessellations.

**Tesseral.** A gridded representation of a plane surface into disjoint polygons. These polygons are normally either square (raster), triangular (TIN), or hexagonal. These models can be built into hierarchical structures, and have a range of algorithms available to navigate through them.

**Thematic map.** A map depicting selected kinds of information relating to one or more specific themes. Examples are soil type, land classification, population density and rainfall maps.

**Theme.** A user-defined perspective on a geographic dataset specified, if applicable, by a name and feature class or dataset name, attributes of interest, or data classification scheme.

**Thiessen polygons.** Polygons whose boundaries define the area that is closest to each point relative to all other points. Thiessen polygons are generated from a set of points. They are mathematically defined by the perpendicular bisectors of the lines between all points. A tin structure is used to create Thiessen polygons.

**Triangulated irregular network (TIN).** A form of the tesseral model based on triangles. The vertices of the triangles form irregularly spaced nodes. Unlike the grid, the TIN allows dense information in complex areas, and sparse information in simpler or more homogeneous areas. The TIN dataset includes topological relationships between points and their neighboring triangles. Each sample point has an X,Y co-ordinate and a surface, or Z-value. These points are connected by edges to form a set of non-overlapping triangles used to represent the surface. Tins are also called irregular triangular mesh or irregular triangular surface model.

**Topological overlay.** An analysis procedure for determining the spatial coincidence of geographic features.

**Topology.** The spatial relationships between connecting or adjacent coverage features (e.g., arcs, nodes, polygons, and points). In digital data, topological relationships such as connectivity, adjacency and relative position are usually expressed as relationships between nodes, links and polygons. For example, the topology of an arc includes its from- and to-nodes, and its left and right polygons. Topological relationships are built from simple elements into complex elements: points (simplest elements), arcs (sets of connected points), areas (sets of connected arcs), and routes (sets of sections, which are arcs or portions of arcs). Redundant data (coordinates) are eliminated because an arc may represent a linear feature, part of the boundary of an area feature, or both. Topology is useful in GIS because many spatial modeling operations don’t require coordinates, only topological information. For example, to find an optimal path between two points requires a list of the arcs that connect to each other and the cost to traverse each arc in each direction. Coordinates are only needed for drawing the path after it is calculated.

**Topography.** Configuration (relief) of the land surface; the graphic delineation or portrayal of that configuration in map form, as by contour lines; in oceanography the term is applied to a surface such as the sea bottom or surface of given characteristics within the water mass.

**Union.** A set operation. The outcome of the union of sets is a set with all the elements of all the input sets. The set operation union can be used in a query languages such as SQL, or can be used as an operator for an overlay operation with spatial data.

**Universal Transverse Mercator (UTM) grid.** Military grid system based on the transverse Mercator projection, applied to maps of the Earth’s surface extending from the Equator to 84 Degrees north and 80 degrees south latitudes.

**Vector data.** An abstraction of the real world where positional data is represented in the form of coordinates. In vector data, the basic units of spatial information are points, lines and polygons. Each of these units is composed simply as a series of one or more co-ordinate points, for example, a line is a collection of related points, and a polygon is a collection of related lines.

**Zoom.** To enlarge and display greater detail of a portion of a geographic data set.
References


(12) INEGI/SPPMANUAL: Elementos de cartografía detenial. p. 29

(13) Joly Fernand. La cartografía Ed Ariel. Barcelona. 1982


CHAPTER 3:

Relational Database Systems in Geographic Information Systems
3. RELATIONAL DATABASE SYSTEMS IN GEOGRAPHIC INFORMATION SYSTEMS

Presentation

Data processing, analysis, and interpretation are the most complex tasks prior to informed decision making by health managers. These tasks are often conducted by statistics and epidemiology personnel lacking the proper training, instruments, and methods to use information systems.

Epidemiologic data are usually stored in different electronic formats, such as spreadsheet (Lotus, Excel), database (Dbase, Access, EpilInfo), and client/server (mainframe or Unix) database (MS SQL Server, Oracle, Sybase, Informix, Adabas, etc.) formats. Prior to analysis, data must be subject to validation, standardization, and normalization.

GIS include management functions for spatial and non-spatial data, based mainly on the Relational Model (RM). This model simplifies data integration from different sources by establishing relationships among entities (objects and their characteristics), using a programming language called Structured Query Language (SQL).

This chapter presents an introduction to GIS RM and reviews its capabilities for efficient and effective data management.

Objectives

- To present RM basic concepts on database design and management applied to GIS.
- To describe the structure of SQL query clauses for epidemiologic analysis using GIS.
- To apply SQL query clauses for data processing and analysis.

Expected Results

The user is expected to understand the importance of SQL to process spatial and non-spatial data using GIS. Also, he will be able to design simple databases using RM to process data using SQL query clauses. Finally, the user will be able to apply SQL for epidemiologic and health data analysis using GIS.

Summary

This section will review RM with particular reference to data structure, integrity, and manipulation.

The structure of data has two basic characteristics: domains and relations. Domains are the possible values of a variable (e.g., gender may take the female, male, and no data values). Relations represent the breakdown of an object or entity in attributes or characteristics, presented as rows (tuples) and columns (attributes) on a table. In spatial analysis, rows usually represent the geographical unit and columns its variables (e.g., population, socioeconomic status, mortality, etc.) A special type of attribute - the primary key or identifier - is required to identify uniquely each row; i.e., no two rows are equal.

Data integrity establishes rules that are followed by entity values of relations. There are two basic rules: the entity integrity and the referential integrity rules. The former establishes that the key identifier cannot take null values, so that the table rows are uniquely referenced; the latter establishes that the values of relation attributes between two tables must correspond to the values of the key identifier.

Data manipulation involves three types of functions: relational allocation, relational algebra operations, and data processing language (such as SQL) functions. Relational allocation functions are used by RM to represent relationships among tables. These relationships are classified by the type of links established among entities (e.g., one-to-one, one-to-many, many-to-many). Relational algebra operations are functions that allow the manipulation of relationships, such as union, intersection, difference, product, restriction, projection, join, and division. Data processing using the RM is frequently done with SQL, based on relational algebra that consists of a series of clauses to perform data processing queries.

RM are powerful tools to represent and abstract reality through the design of a database. Their ability to facilitate data processing and analysis is greatly expanded by GIS, since it allows management of both spatial and non-spatial data.
Introduction to Relational Database Systems

Informal Description of the Model

This model views reality as the sum of elements made up of objects, which are held together by relationships. Objects have particular characteristics that define them in specific contexts and establish relations within them.

Objects are also called entities. An entity is an object that consists of attributes or specific characteristics that distinguish them from other objects. A relationship is a specific association or link between two or more entities.

Data may be structured as a table with attributes placed in columns and the specific values of each entity in rows. For example, the number of inhabitants in each of Malaria island's regions\(^\text{16}\) may be presented as shown in Figure 3.1.

This type of presentation is called a table; it graphically displays the attributes of a group of entity values; in this case the Malaria island regions.

Using a table that shows the number of malaria cases by region (Figure 3.2), we could model the relationship with the previous table (Figure 3.3). The name of the region in the table of cases must correspond to the name of the region in the table of the population. In relational algebraic notation, this relationship would be stated as:

**cases.region=population.region.**

After establishing this relationship, a virtual table will include data on malaria cases and population by region (Figure 3.4). Two or more separate data sources in a table may be related by a common variable (in this case the variable region).

---

16. A hypothetical country used in PAHO training materials
If a table represents attributes of more than one entity, some attribute values may be repetitive. For example, Figure 3.5, Table 1, shows a relation that includes population attributes, number of disease cases by region, population data, and number of cases by municipality. Note that region values are repeated for each municipality of a given region (dotted circles). This is because the table combines data from two different entities; regions and municipalities. Presenting data in this manner may be inconsistent and redundant. Then, it is better to organize data in two tables (one for region and one for municipality) (Figure 3.5, Tables 2 and 3) using a master-child (one-to-many) relationship, presenting the regions (all, master) that have one or more municipalities (part, child). The table of regions should have one column or a combination of columns (primary key) that uniquely identify the region. The table of municipalities should have one or more columns (foreign key, shown in Figure 3.5, Table 3, column Region), with reference data for the region to which each municipality be-
longs; this column should correspond to primary key data in the table of regions. The same process can be applied when a table has to be divided into more than two tables. This is a very important relationship, where one table represents the parts or members of the elements of another table.

For example, if data for the localities within each municipality are available, they may be stored as a third table having an all-part relationship with the table of municipalities.

Usually, data presented as crosstabs (n x m) of two variables that will store the values of a third variable are organized as in a spreadsheet (Excel, Lotus) or as a text table in which the values of variable A are the row titles, the values of variable B are the column titles, and the values of variable C are in the row-column cells.

For example, population data of Malaria regions by year (1980 and 1985) could be displayed as shown in Figure 3.6. To include data for an additional year (1990), a new column would have to be added. If the database shown in Figure 3.6 is the original one, adding a column implies modifying the database structure.

**Figure 3.6: Crosstabs of two variables**

<table>
<thead>
<tr>
<th>Region</th>
<th>1980</th>
<th>1985</th>
<th>1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Coast</td>
<td>38,263</td>
<td>43,492</td>
<td>58,409</td>
</tr>
<tr>
<td>South Coast</td>
<td>149,521</td>
<td>215,376</td>
<td>331,585</td>
</tr>
<tr>
<td>South East(highl.)</td>
<td>206,372</td>
<td>264,821</td>
<td>352,344</td>
</tr>
<tr>
<td>East (mountains)</td>
<td>95,518</td>
<td>120,364</td>
<td>155,676</td>
</tr>
<tr>
<td>North West (jungle)</td>
<td>78,651</td>
<td>105,785</td>
<td>122,386</td>
</tr>
</tbody>
</table>

This may involve unwanted effects on data entering formats, reports, etc., which would have to be modified. One of the main problems of this type of design is that the values of the attribute year (1980, 1985, 1990) have been modeled as if they were the actual attribute and not the value of the attribute, which makes data processing more difficult. For example, if one wanted to know the maximum population value in any Malaria region, each one of the year columns would have to be processed to find that value. The RM offers an alternative to represent the attributes in the columns and the attribute values as tuples, to design the table as shown in Figure 3.7, where population data for a new year (1990) can be added without having to modify the database structure.

**Figure 3.7: Representation of attribute values as tuples**

<table>
<thead>
<tr>
<th>Region</th>
<th>Year</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Coast</td>
<td>1980</td>
<td>38,263</td>
</tr>
<tr>
<td>North Coast</td>
<td>1985</td>
<td>43,492</td>
</tr>
<tr>
<td>North Coast</td>
<td>1990</td>
<td>58,409</td>
</tr>
<tr>
<td>South Coast</td>
<td>1980</td>
<td>149,521</td>
</tr>
<tr>
<td>South Coast</td>
<td>1985</td>
<td>215,376</td>
</tr>
<tr>
<td>South Coast</td>
<td>1990</td>
<td>331,585</td>
</tr>
<tr>
<td>South East(highl.)</td>
<td>1980</td>
<td>206,372</td>
</tr>
<tr>
<td>South East(highl.)</td>
<td>1985</td>
<td>264,821</td>
</tr>
<tr>
<td>South East(highl.)</td>
<td>1990</td>
<td>352,344</td>
</tr>
<tr>
<td>East (mountains)</td>
<td>1980</td>
<td>95,518</td>
</tr>
<tr>
<td>East (mountains)</td>
<td>1985</td>
<td>120,364</td>
</tr>
<tr>
<td>East (mountains)</td>
<td>1990</td>
<td>155,676</td>
</tr>
<tr>
<td>North West (jungle)</td>
<td>1980</td>
<td>78,651</td>
</tr>
<tr>
<td>North West (jungle)</td>
<td>1985</td>
<td>105,785</td>
</tr>
<tr>
<td>North West (jungle)</td>
<td>1990</td>
<td>122,386</td>
</tr>
</tbody>
</table>

Data collection and management usually involves coding the specific values of the variables. For example, the attribute gender of the entity person may take the values female and male, which could be coded as F for female and M for male, or as 0 and 1. This association between the code and the descriptor of the attribute can be modeled using RM, as will be described below.

In summary, RM are used to represent entities and attributes as relations or tables; i.e., for each group of entity values there is a relation. Relationships are established through values of common attributes that allow displaying related tables as one single table. It should be emphasized that the technical terms relationship and relation are different; the former refers to an association or connection between two objects, and the latter to a specific set of data organized and presented on a table. Since most GIS programs are based on the RM, understanding the design and purpose of the basic aspects of RM is essential to profit from GIS.
Relational Model (RM)

The Relational Model (RM) has been established as the main model available for commercial data processing applications (Data Base Management Systems (DBMS)) to increase the independence and consistency of data by means of strict definitions of data structure and manipulation. Three important aspects of the RM are data structure, integrity, and processing/manipulation.

Relational Data Structure

The central notion of the RM is that a database (DB) is conceived as a set of relations that is represented as one or more tables. Entity attribute values are stored in tuples or records, Relationships among entities are established by common columns. Each relation (or table) corresponds to one entity. Each table must have a primary key that uniquely identifies its rows.

Data structure includes two basic characteristics: domains and relations. Domains are the possible specific values a variable may take (e.g., gender may take the values female, male, and no data). Relations represent the breakdown of an object or entity in attributes or characteristics in a table with rows or tuples and columns or attributes; therefore, a relation is represented by a table.

In the context of spatial analysis and GIS, rows usually represent the geographical unit and columns represent the variables related with that geographical unit (e.g., population, socioeconomic status, mortality, etc.).

The number of tuples is called the cardinality and the number of attributes the grade or arity. The primary key or identifier is the attribute (or combination of attributes) that uniquely identifies each tuple of a relation, or in other words, each row of a table. The primary key has two important properties:

1) No two tuples in a relation have the same values in the same attribute or combination of attributes that identify said relation; this is, no two tuples in a relation have the same primary key.

2) A primary key has no superfluous components; no attribute that is part of a primary key can be discarded. Formally, the primary key is that with the smallest number of attributes of all the possible keys in a table.

Table 3.1 summarizes the formal and informal relational data structure terms and provides an example of each one.

Elements of a relation:

- A tuple is a row and an attribute is a column in a table.
- The primary key is a column or combination of columns that uniquely identifies the table.
- A domain is the set if possible values that an attribute (column) can take.
- The atomic value is the smallest semantic information unit without an internal structure (it cannot be disassembled without becoming meaningless), in the model’s perspective. For example, a person’s name becomes meaningless if its letters are scattered, but the name can be broken into first name and last name.

<table>
<thead>
<tr>
<th>Formal term</th>
<th>Informal term</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
<td>Table</td>
<td>Registry of notifiable diseases</td>
</tr>
<tr>
<td>Tuple</td>
<td>Row or record</td>
<td>Unit data (municipality X)</td>
</tr>
<tr>
<td>Cardinality</td>
<td>No. of row</td>
<td>No. of municipalities</td>
</tr>
<tr>
<td>Attribute</td>
<td>Column or field</td>
<td>Variables (name of municipality, week, cases)</td>
</tr>
<tr>
<td>Grade or Arity</td>
<td>No. of columns</td>
<td>No. of variables</td>
</tr>
<tr>
<td>Primary key</td>
<td>Unique identifier</td>
<td>Municipality code</td>
</tr>
<tr>
<td>Domain</td>
<td>Possible legal values</td>
<td>Set of all the municipality names</td>
</tr>
</tbody>
</table>
Properties of relations

Relations have four inherent properties, as follows:

- **There are no repeat tuples.** Table rows cannot have the same exact values as table columns. At least one of the column values must differ between rows.

- **Tuples are not ordered from top to bottom.** This is a property of relations, and not necessarily of tables, which are physical representations of relations.

- **Attributes are not ordered from left to right.** Same as above, but applied to the sequential order of table columns.

- **All attribute values are atomic.** Attributes or columns represent units that may not or cannot be divided into smaller parts; this is, a column cannot be divided into two or more columns. For example, the table column full name may not be an atomic value, since it can be divided into last name and first name columns. For example, the column full name may not be atomic since it can be divided into at least two columns, one for last name and one for first name.

It should be noted that in the conceptual model there is no ordering, but the physical representation of the relation as a table requires that rows and columns are ordered. This order does not affect the conceptual model.

Types of relations

A relational system includes different types of relations (Date, 1994):  

- **Base relations**, also called real relations: This is an independent relation that has a name. Base relations are so important for the given application that the database designer has given them a name, stored them in physical media, and included them in the database, unlike relations of a temporary nature, such as query results.

- **Views**, also called virtual relations: A view is a relation with a name, represented in the system just by definitions of other base relations. Unlike base relations, views have no proprietary data and are just different versions of base relations.

- **Query results**: They are the end results of queries that may receive a name or not. Query results are not permanently stored in the database.

- **Intermediate results**: Relations (usually with no name) resulting from relational expressions nested within a bigger relational expression. A relation resulting from a query within another query is an intermediate result. Similar to query results, they are not permanently stored in the database.

Keys

In Relational Databases, keys are the attributes that identify the tuples of a relation. A key is the column or combination of columns that uniquely identifies the values of an entity (tuples) in a relation. The key types are (Date 1994):

- **Primary key**: It uniquely identifies each tuple of a relation; no two tuples in a table can have the same primary key (Figure 3.8, bold text). If more than one key is present, the primary key will be the one with the smallest number of columns.

- **Foreign key**: It is an attribute or combination of attributes in a relation A that establishes a relationship with a relation B. The foreign key of relation A is often the primary key of relation B and not necessarily the primary key of relation A (Figure 3.8, underlined text).

Keys can be simple or concatenated:

- **Simple keys**: They are formed by one and only one attribute (Figure 3.9, underlined text).

- **Concatenated keys**: They are formed by two or more attributes (Figure 3.9, bold text)

**Integrity Rules (Codd 1981)**

- **Entity integrity rule**: Primary key components in a base relation cannot take null values (empty data). A base relation of a relational database can only record data that can be identified, thus, the row values must be different.

- **Referential Integrity Rule**: The database should not contain foreign key values without concordance. A foreign key value in a relation should have the same value in the primary key of the other relation. For...
example, if a record in table B refers to a record value in table A, the value in table A has to exist; for example, a state has to exist to contain municipalities. It is important to note that if primary key data are changed, then all foreign key data using that primary key should also be changed in order to maintain referential integrity.

*Observing Integrity Rules:* Most database systems that support the relational model (MS Access\(^{18}\), Corel Paradox\(^{19}\), Oracle\(^{20}\), Informix\(^{21}\), SQL Server\(^{22}\), Sybase\(^{23}\), etc.) include functions to declare and confirm integrity rules, thus allowing the user to define limitations (constraints) required to maintain the consistency and integrity of the database.

The theory of functional dependencies and the normalization method provide the grounds to analyze data and assure proper data modeling while observing the integrity rules. A detailed description is beyond the scope of this document, so the interested reader is referred to the specialized literature on the subject.

*Data Manipulation and Processing*

Three aspects of data manipulation and processing are considered: Relational allocation, data processing language (reviewed in detail below), and relational algebra operations (reviewed briefly in the Appendix).

---

Relational Assignment

Relational assignment refers to the RM representation of a relationship between tuples (rows) of one relation (table) and tuples (rows) of another relation (table), through the primary and foreign keys. The possible cardinalities of relational assignment are four:

- One-to-one (1-1): A tuple in relation A is associated with one and only one tuple in relation B, and a tuple in relation B is associated with one tuple in relation A (Figure 3.10). This is, each row in table A is associated with one and only one row in table B.

- One-to-many (1-∞): Also called composition; a tuple in relation A associates with a number of tuples in relation B; however, a tuple in relation B can be associated with one and only one tuple in relation A (Figure 3.11). This is, a row in table A can correspond with many rows in table B, but rows in table B can only correspond with one row in table A. Figure 3.11 shows a state composed of many municipalities, represented as two tables. The State identification row (ID-State) in the State table has as many corresponding rows in the municipality table as municipalities need to be represented and stored for that State.

- Many-to-one (∞-1): A tuple in relation A is associated with no more than one tuple in relation B; however, a tuple in relation B can associate with more than one tuple in relation A. This is a symmetrical view of the one-to-many relationship (a Municipality can be associated with no more than one State, but a State can be associated with many Municipalities.)

- Many-to-many (∞-∞): Also called association; a tuple in relation A is associated with any number of tuples in relation B, and a tuple in relation B is associated with any number of tuples in relation A. This is usually represented using an intermediate table that has a one-to-many relationship with table A and a many-to-one relationship with table B (Figure 3.12). For example, a many-to-many relationship occurs between values of the variable gender (male and female) of the table Gender, and the variable age group (<1, 1-4, 5-24, 25, 45-64, 65+) of the table Age Group; in this case, a gender value (e.g., male) may correspond with many values of age group, and a value of age group, for example, 5-24, may cor-

Figure 3.10: One-to-one relationship

<table>
<thead>
<tr>
<th>Provincia</th>
<th>Municipio</th>
<th>Municipio</th>
<th>casos_ac</th>
<th>muetes_ac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pinar del Rio</td>
<td>CANDELARIA</td>
<td>CIEGO DE AVILA</td>
<td>38</td>
<td>5</td>
</tr>
<tr>
<td>Matanzas</td>
<td>CIENAGA DE ZAPATA</td>
<td>CONTRAORDERE</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Matanzas</td>
<td>CARDENAS</td>
<td>CAMAGUEY</td>
<td>72</td>
<td>16</td>
</tr>
<tr>
<td>Matanzas</td>
<td>COLON</td>
<td>CERRO</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Cienfuegos</td>
<td>CUMANCAYAGUA</td>
<td>CHAMBAS</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Villa Clara</td>
<td>CADDARIEL</td>
<td>CALIXTO GARCIA</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Ciego de Avila</td>
<td>CHAMBAS</td>
<td>CUETO</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Ciego de Avila</td>
<td>CIEGO DE AVILA</td>
<td>CALIETE</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Las Tunas</td>
<td>COLOMBIA</td>
<td>COLON</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>Huquin</td>
<td>CALIXTO GARCIA</td>
<td>CABAQUAN</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Granma</td>
<td>CAMPECHEULA</td>
<td>COTORRO</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>La Habana</td>
<td>CAMILO</td>
<td>CABARION</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Matanzas</td>
<td>CALIETE</td>
<td>CIENFUEGOS</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Villa Clara</td>
<td>CARRILLO</td>
<td>CIENFUEGOS</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>Cienfuegos</td>
<td>CIENFUEGOS</td>
<td>CARLOS MANUEL</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Cienfuegos</td>
<td>CIERNES</td>
<td>CACICUNA</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Villa Clara</td>
<td>CIFUENTES</td>
<td>COLOMBIA</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Villa Clara</td>
<td>CAMAJUAN</td>
<td>CARDENAS</td>
<td>18</td>
<td>6</td>
</tr>
</tbody>
</table>
respond with the two values of the variable gender. This is presented in the table Gender Age Group (Figure 3.13), which consists of rows that combine the different values of the variable gender and age group.

The different types of relations and relationships have the purpose of modeling reality. The underlying principle poses that if entity x depends on entity y, then x is said to be dependent upon the existence of y. In practice, this means that if y is suppressed, x is also suppressed. Entity y is called the dominant entity and entity x the subordinate entity.

The section below presents a review of the benefits of the RM. Earlier database design paradigms (mainly the hierarchical model and the network model) had data consistency and redundancy problems. The RM is based on a robust formal mathematical theory that overcomes those problems and provides a solid foundation for database design. It should be noted that the RM paradigm is being constantly reviewed to adjust it to the demands posed by new data types, such as multimedia images, audio, video, etc., in addition to traditional numeric and string data.

**Benefits of the Relational Model**

Benefits include assurance of data consistency, elimination of data redundancy, optimization of computer memory usage, and simplification of data processing.

- **Assurance of data consistency.** Storage of the same data in different forms may occur when all the tuples with the same data values are not updated in a consistent fashion. For example, to represent data about the population of localities of a municipality in a table, one could do it as shown in Figure 3.14 (table structure) and 3.15 (table values). Representing data in this way has several problems, because data are repeated (redundancy) in the table; if the locality name is modified (in this case Cienfuegos), all the records
with this value of the variable locality will have to be modified accordingly. Otherwise, there will be two different values for the same data, which will result in inconsistency. This problem may be avoided by using RM to represent these data as shown in Figures 3.16 and 3.17.

- **Elimination of data redundancy.** Repetitive attributes in tables should be eliminated. For example, when recording a patient's symptoms on a table, as many attributes as number of symptoms are created, i.e., symptom1, symptom2, ..., symptomn. This can be avoided by constructing an additional table of symptoms that can be related through the patient's identifier data by means of a many-to-one relation with the patient's table.

- **Optimization of computer memory usage.** Adequate database design without inconsistency and redundancy warrants proper memory allocation and computer resources.

- **Simplification of data processing.** Figures 3.14 and 3.15 show tables that have not undergone normalization and Figures 3.16 and 3.17 show the same tables after normalization. At first sight it may seem that the model has been complicated, rather than simplified; however, the original table (Figure 3.14) shows redundancy and thus a greater chance of inconsistency; a close look at the normalized tables shows that redundancy has been eliminated. In addition, data updating is easier and modeling of reality more precise. Note that in the original table it is difficult to determine relations among entities. Also, if necessary, one may access the original table after normalization, as presented in the next section.

---

**Figure 3.15: Table values before redesign**

<table>
<thead>
<tr>
<th>ID_Estado</th>
<th>Nombre</th>
<th>Municipio_Suf</th>
<th>ID_Localidad</th>
<th>Localidad_Suf</th>
<th>Sexo_Descripción</th>
<th>Edad_Descripción</th>
<th>Población</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>&lt;1</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>1-4</td>
<td></td>
<td></td>
<td>120</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>5-9</td>
<td></td>
<td></td>
<td>130</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>10-15</td>
<td></td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>16-19</td>
<td></td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>20-24</td>
<td></td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>25-44</td>
<td></td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>45-64</td>
<td></td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>65+</td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Total</td>
<td>Total</td>
<td></td>
<td></td>
<td>163</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>&lt;1</td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>1-4</td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>5-9</td>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>10-15</td>
<td></td>
<td></td>
<td>150</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>16-19</td>
<td></td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>1 Cienfuegos</td>
<td>200</td>
<td>1</td>
<td>12 Femenino</td>
<td>20-24</td>
<td></td>
<td></td>
<td>170</td>
</tr>
</tbody>
</table>
Figure 3.16: Table structure and relationships after redesign using RM

Figure 3.17: Table values after redesign using RM
## Processing Databases

**Query Languages**

A query language is a computer programming language that allows a user to ask questions that will elicit answers through data processing. These languages are classified into procedural and non-procedural. Procedural languages require that users provide a precise sequence of instructions (procedure) and operations for the system to compute the query.

This section presents the Structured Query Language (SQL), a non-procedural language that has become the standard for relational model query languages. It is based on relational algebra and calculus (see further details of relational algebra in the Appendix).

**Structured Query Language (SQL)**

SQL is utilized to create, manage, and query DBs. SQL queries allow data processing necessary for data analysis and are reviewed in this section.

What is the difference between using the RM and SQL instead of other tools such as spreadsheets, statistical software and similar systems? First, the RM provides guidance for database design, allowing the construction of a model closer to reality, with internal consistency and integrity, and with less problems than other types of designs. Second, the RM and SQL are not exclusive features of a given product, but rather industry standards that are utilized by a wide variety of database systems. Therefore, learning RM and SQL methods, enables users to understand different database programs. In addition, middleware24 programs have been developed, that allow one user or group of users in a client/server environment access and put together data from different sources. Windows-based middleware most commonly used is Open Database Connectivity (ODBC), a program that allows accessing remote databases regardless of their type. Since different SQL versions are available, the user should review the DB program documentation to check the SQL version and its capabilities.

A hypothetical example to review the utilization of SQL is presented below:

Consider having three files (paper, not electronic) of mortality records: one with cancer and neoplasm as the general basic causes of death (BCD), one with malaria as the BCD, and a third one with AIDS as BCD. Below you will find several strategies that will allow you to select different records from those files.

Try to answer the questions presented along the example before proceeding and then check the answers. The answers may differ a little from your answers; if so, think of the possible reasons (after the question you will see the SQL query syntax in boldface; always use the English syntax).

1. **Select all the records where AIDS was BCD.**
   This is a simple query, but will be described in detail.
   1.1 From which file(s) will the records be retrieved? (FROM)
   Answer (A): From the third file, which is the one that includes AIDS as BCD.
   1.1 What will be selected? (SELECT).
   A: All the records of AIDS as BCD:
   It is evident that by answering the previous questions the query will be fulfilled.

2. **Select the records where leukemia was the BCD.**
   2.1 From which file(s) will the leukemia records be retrieved? (FROM)
   A: From the first file, the one that includes cancer as BCD.
   2.2 What will be selected? (SELECT)
   A: Records of leukemia as BCD.
   2.3 Which conditions must be met to select leukemia cases from among all other cancer cases? (WHERE)
   A: The cases selected must be only leukemia cases.

---

24. Middleware: software that connects an application program and a network. It allows the interconnection among incompatible programs access heterogeneous computer platforms. An example is the Object Request Broker (ORB), a software that allows communication among objects (object-oriented technology). [http://www.instantweb.com/foldoc/foldoc.cgi?query=middleware](http://www.instantweb.com/foldoc/foldoc.cgi?query=middleware) In the computer industry, middleware is a generic term for any program that “glues together” two programs. A common application of middleware is writing programs to access databases from other databases. [http://whatis.techtarget.com/WhatIs_Search_Results_Exact/1,282033,,00.html?query=middleware](http://whatis.techtarget.com/WhatIs/Search_Results_Exact/1,282033,,00.html?query=middleware)
3. Select the records of female cases.
3.1 From which file(s) will female cases be retrieved? (FROM)
   \( A: \text{From the three files, since there may be female cases of cancer, malaria, and AIDS.} \)
3.2 What will be selected? (SELECT)
   \( A: \text{Records of female cases.} \)
3.3 Which conditions must be met to select female cases from among all others cases? (WHERE)
   \( A: \text{That the selected cases are only female cases.} \)

After gaining some experience with SQL questions, all similar questions follow the same structure and the process becomes automatic. However, the novice user should describe the questions explicitly to make sure that the query clauses are constructed adequately.

**Basic structure**

Basic SQL queries consist of three clauses: SELECT, FROM and WHERE.

- The SELECT clause lists the attributes (columns) that are requested in a query. (This relational algebra operation is known as projection).
- The FROM clause lists the relations (tables) that will be examined during the query. (This relational algebra operation is known as product).
- The WHERE clause consists of a condition that checks the attributes of the relations included by the FROM clause, to select those tuples (rows) that meet the given condition. (This relational algebra operation is known as the selection predicate).

A typical SQL query takes the form:

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Informal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>SELECT A_1, A_2, ..., A_n</code></td>
<td>Selects the next list of attributes (necessary)</td>
</tr>
<tr>
<td><code>FROM t_1, t_2, ..., t_m</code></td>
<td>From the relations (necessary)</td>
</tr>
<tr>
<td><code>WHERE C</code></td>
<td>Where this (these) condition(s) are met (optional)</td>
</tr>
</tbody>
</table>

Each \( A_i \) represents an attribute (column) or mathematical expression, where it is possible to perform operations such as addition, subtraction, multiplication, division, exponentiation, etc. and combination of them, to construct an expression.

Each \( t \) represents a relation that will provide the data for the query. \( C \) is a single or multiple condition to be met by the rows that will be selected for the query.

The WHERE clause may be omitted, but SELECT and FROM are indispensable. The list of attributes \( A_1, A_2, ..., A_n \) may be substituted by an asterisk (*) to select all the attributes from all the relations selected by the FROM clause. The order of the clauses is important: specify SELECT first, FROM second, and WHERE and other optional clauses afterwards.

A SQL query result is a new result relation. If duplicate records need to be eliminated, use the DISTINCT keyword after SELECT. For example, suppose we have a table with data on hospital cases and their diagnosis and we want to know the different diagnosis. Using DISTINCT in the column of the variable diagnosis will generate a table showing the name of each diagnosis.

In general, when the FROM clause is indicated in more than one table, the WHERE clause should include the relationship condition, to state which condition(s) must be met to establish the relationship between tables. Also, in these cases dot notation name-table.name-attribute is used to define the attributes, to make clear which is the source table.

Some examples of different query types are presented below. The heading of each example is the name of a query type in the nomenclature of relational databases. The examples use the tables Region, Population and Cases below, and their relationships:
Simple retrieval
Select the units from all municipalities
   SELECT IDUNIT
   FROM CASES

Simple retrieval (SELECT *)
Select the complete population data from municipalities
   SELECT *
   FROM POPULATION

Conditional retrieval
Select the units where more than ten cases have occurred.
   SELECT IDUNIT
   FROM CASES
   WHERE CASES > 10

Regarding GIS software program capabilities, MapInfo supports SQL and RM with some restrictions. For example, it does not allow using combined keys among spatial tables and data tables. In some cases, when complex expressions are given using the SELECT clause, a syntax error message may be displayed when in fact there is no error. MapInfo also supports communication with a relational database management server (RDBMS) through ODBC, and allows using the server for more complex queries and operations.

ArcView software may also be connected with a SQL server through ODBC to get results using the SQL CONNECT window, which allows using the basic SQL clauses SELECT, FROM, and WHERE on a database server. However, ArcView does not use SQL to query its data and spatial tables; rather, it uses relational assignation and query sub-language available under the QUERY option. It only supports one-to-one (1-1), also called “join” and one-to-many (1-∞), also called “link” relationships. When a link across the tables is established, if a row from the first table is selected, all the linked rows of the second table are automatically selected and no result table is created. Note that more than two tables may be joined (JOIN) or linked (LINK) at the same time, hence allowing representation of many-to-many (∞-∞) relationships. The query sub-language available though the query builder includes criteria to select comparative and logical operators of attributes and possible values (domains) (known in relational algebra as restriction).

Operations in a query
SQL includes the common set theory operations UNION, INTERSECT and DIFFERENCE, that allow working with two relations to obtain a new result relation. As an example Figure 3.18 shows the results applying these operators to regions.

Figure 3.18: Set theory operations
In MapInfo, these operations are substituted by geographical operators, as follows:

- CONTAINS: Object X contains object Y if the centroid of Y is within the limit of X. For example, Figure 3.18 shows that region C contains region A since the centroid of A is within the limits of region C.

- CONTAINS ENTIRELY: Object X contains entirely object Y if the limits of Y are entirely within the limits of X. Figure 3.18 shows that region A contains region B entirely, since the limits of region B are within the limits of region A.

- WITHIN: Object X is within object Y if its centroid is within the limits of Y. Figure 3.18 shows that region A is within region C, since the centroid of A is within the border of region C.

- ENTIRELY WITHIN: Object X is entirely within object Y if the limits of X are completely within the border of Y. Figure 3.18 shows that region B is entirely within region A, since the limits of region B are entirely within region A.

- INTERSECTS: Object X intersects object Y, if these two objects share at least one common point. Figure 3.18 shows that region A intersects region C and region C intersects region A, since they share a common area.

The software ArcView utilizes the following operations to perform spatial queries.

- ARE COMPLETELY WITHIN: Object X is completely within object Y if the limits of X are completely within the limits of Y. Figure 3.18 shows that region B is completely within region A since the limits of region B are within the limits of region A.

- COMPLETELY CONTAINS: Object X completely contains object Y if the limits of Y are completely contained within the limits of X. Figure 3.18 shows that region A completely contains region B, since the limits of region B are completely within the limits of region A.

- HAVE THEIR CENTER IN: The centroid of object X is within object Y. Figure 3.18 shows that region C contains region A, since the centroid of region A is within the limits of region C.

- CONTAIN THE CENTER OF: Object X contains the centroid of object Y. Figure 3.18 shows that region A is within region C, since the limits of region C contain the centroid of region A.

- INTERSECT: Object X intersects object Y if these two objects share at least one common point. Figure 3.18 shows that region A intersects region C and region C intersects region A, since they share a common area.

- ARE WITH DISTANCE OF: Object X and object Y are with distance D. ArcView works also with two types of spatial joins: INSIDE and NEAREST (ESRI 1994).

**Conditions and logical operators**

SQL uses dot notation for prefixing table names to column names, such as name-table.name-attribute, to avoid ambiguity when attributes appear in more than one table.

SQL also uses the logical operators, AND, OR, and NOT; comparative operators < (less than), > (greater than), = (equal to), <> (different from), <= (less than or equal to), and >= (greater than or equal to). SQL also allows using the arithmetical expression + (addition), - (subtraction), * (multiplication), / (division), and ^ (exponential), on constants or attribute values.

Many SQL include the comparative selection operator BETWEEN. This operator simplifies WHERE clauses to specify conditions where a variable can take values less than or equal to a given lower value. A corresponding comparison operator NOT BETWEEN is also available.

Note that condition prefixing is case-sensitive i.e., uppercase characters are different from lowercase characters.

The operator LIKE is available to indicate that an attribute must correspond with another attribute or with a constant. Some SQLs include the operator EXACTLY LIKE, which allows defining character-to-character correspondence. The operator NOT LIKE would indicate non-correspondence.
SQL allows using matching character string models to make comparisons. These models are described through two special characters (wild-card):

- Per cent (%): Equal to any group of characters (sub-string)
- Underline (_): Equal to any other character.

For example, "Healt%" is equal to any string starting with the characters "Healt", such as: "Health", "Healthy", "Healthier". In turn, "Healt_" is equal to any string starting with "Healt" and followed by any one character; from the previous examples, "Health" is the only alternative.

Some examples of queries using these conditions are presented below. The heading of each example corresponds to the name given to that particular query in the nomenclature of relational databases models.

The same tables from a previous example are utilized:

---

**Simple equijoin**

Get data on regions, their municipalities, and populations.

```sql
SELECT REGION.*, POPULATION.*
FROM REGION, POPULATION
WHERE REGION.IDMUNICIP = POPULATION.IDMUNICIP
```

Comparison operators other than "=" can be utilized. By definition, equijoining should generate two identical columns. If one of those columns is eliminated, the result will be what is known as a natural join, which is the most useful type of join since it allows integrating normalized tables in two or more tables.

**Join with an additional condition**

Get data on regions and their municipalities, that have a population under 5,000 inhabitants.

```sql
SELECT REGION.*, POPULATION.*
FROM REGION, POPULATION
WHERE REGION.IDMUNICIP = POPULATION.IDMUNICIP
AND POPULATION < 5000
```

---

**Retrieving specific fields of a join**

Get the name of the regions and their municipalities

```sql
SELECT REGION REGION, POPULATION.MUNICIPALITY
FROM REGION, POPULATION
WHERE REGION.IDMUNICIP = POPULATION.IDMUNICIP
```

**Ordering the presentation of records**

SQL allows some control over the order in which table rows are presented. The ORDER BY clause orders the records selected by a query. By default, SQL lists data in ascending order. Data can be ordered using DESC for descending order and ASC for ascending order.

**Syntax:** [ORDER BY attribute] [,attribute] ... [attribute]|25

Also, data can be ordered by levels, this is, if the result needs to be ordered using various attributes, data will be first ordered by the first attribute of the ORDER BY clause, then by the second attribute while keeping the order of the first one, and so on, until ordering by all attributes is accomplished. Ordering a great number of records may be costly in terms of computer resources and time, thus it should be used sparingly.

**Retrieving with order**

Get the units of "Las Golondrinas" municipality ordered from smallest to largest number of cases.

```sql
SELECT IDUNIT
FROM CASES
WHERE IDMUNICIP="LAS GOLONDRINAS"
ORDER BY CASES
```

**Grouping, Aggregation Functions, and Group Conditions.**

SQL allows grouping records that share the values of one or more variables. The GROUP BY clause will form groups according to the specified attributes.

**Syntax** [GROUP BY attribute] [,attribute] ... [,attribute]

The records that have the same value as the attribute specified in GROUP BY are placed in the

---

25 All syntax notations show the exact clauses in bold type and uppercase, variables in normal type, and optional clauses between brackets [ ]. Remember that only FROM and SELECT are necessary clauses.
same group. When several attributes are specified, the records have the same values across all the attributes.

SQL offers functions to calculate the mean (AVG), minimum (MIN), maximum (MAX), total (SUM), and count (COUNT). These operations are aggregation functions that work with groups of registers. They are commonly utilized in the SELECT clause to calculate functions. The result of an aggregation function is a single value for each group, presented in a column of the result table.

Keeping duplicate values is important to calculate mean values; however, in some cases duplicate values need to be eliminated before calculating an aggregation function. Duplicate values can be eliminated by using DISTINCT in the aggregation expression.

Syntax: [Aggregation Function ([DISTINCT] attribute)]

COUNT is frequently used to count the number of records on a table and is expressed as COUNT(*).

Also, the SELECT clause can include expressions that combine mathematical operators with aggregation functions, for example, to calculate rates, as will be described below.

Sometimes it is useful to state group conditions, to work with groups rather than with records. A given condition is applied to each of the groups constructed by the GROUP BY clause. The SQL HAVING clause will express this type of query; it is applied after groups are formed and may be combined with aggregation functions.

It is also possible to treat the whole table as a single group. In that case, HAVING allows selecting the groups just as WHERE is utilized to select rows. HAVING should produce a single value for each group, so that the comparison operators check for the specified conditions.

Syntax: [HAVING group condition]

If a query clause includes both WHERE and HAVING, apply first the condition of the WHERE clause. The records meeting that condition will be grouped by the GROUP BY clause. Then, the HAVING clause will be applied to each group. The groups that meet the condition of HAVING are utilized by the SELECT clause to generate records for the result of the query. If HAVING is not utilized, the whole set of records that satisfy the WHERE clause will be treated as a single group.

MapInfo supports the clause GROUP BY and the aggregation functions, but sometimes, when dealing with complex queries, it sends message errors when there is none. It does not support the clause HAVING. This problem will be reviewed later. Figure 3.19 shows MapInfo's SQL Select dialog box.

**Figure 3.19: MapInfo SQL dialog box**

ArcView utilizes the Summary Table Definition to perform GROUP BY and aggregation functions. Result tables can be obtained for the following aggregation functions: AVERAGE, SUM, MINIMUM, MAXIMUM, STANDARD DEVIATION, VARIANCE, FIRST, LAST, and COUNT.

Some examples of grouping, aggregation functions, and group conditions are shown below. The heading of each example corresponds to the nomenclature of relational databases. These examples are given with the tables utilized in previous examples:
Aggregation Functions using Select:
Get the total number of units
SELECT COUNT(*)
FROM CASES

Aggregation functions using Select with Distinct:
Get the total number of municipality units that have different numbers of cases.
SELECT COUNT(DISTINCT CASES)
FROM CASES

Aggregation functions using Select with a condition:
Get the total number of cases from "LAS GOLONDRINAS" municipality
SELECT CASES
FROM CASES
WHERE IDMUNICIP="LAS GOLONDRINAS"

Get the total number of cases from "LAS GOLONDRINAS" municipality
SELECT SUM(CASES)
FROM CASES
WHERE IDMUNICIP="LAS GOLONDRINAS"

Using group By:
SELECT MUNICIPALITY, SUM(CASES)
FROM CASES
GROUP BY MUNICIPALITY

Using Having:
Get the mean number of cases by municipalities
having means lower than five cases.
SELECT MUNICIPALITY, AVG(CASES)
FROM CASES
GROUP BY CASES
HAVING AVG(CASES) < 5

SQL data types
The standard SQL supports a limited set of data types, as follows:
- CHAR: String characters with a fixed length, specified by the user.
- NUMBER: Number with a fixed comma, its precision is specified by the user.
- INTEGER: Whole number (a finite subset of machine-dependent integers).
- SMALLINT: Small integer (a subset of integer type domain, also machine-dependent).
- DECIMAL AND DOUBLE: Number with floating comma and two-fold precision floating comma, also machine-dependent.

Different database systems support different data types and the range of numeric values depends on computer power available, (PC, minicomputer, main frame, etc.).

MapInfo supports the following data types:

<table>
<thead>
<tr>
<th>DATA TYPES</th>
<th>DESCRIPTION AND RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMALLINT</td>
<td>Integers between -32,767 to +32,767 (inclusive).</td>
</tr>
<tr>
<td>INTEGER</td>
<td>Integers between -214,748,364 to +214,748,364 (inclusive).</td>
</tr>
<tr>
<td>FLOAT</td>
<td>Floating point numbers.</td>
</tr>
<tr>
<td>STRING</td>
<td>Strings of varying lengths, up to 32768 bytes in size.</td>
</tr>
<tr>
<td>STRING * LENGTH</td>
<td>Fixed length strings (where length is the length of bytes in the string). Note: fixed length strings are filled with blank spaces at the end.</td>
</tr>
<tr>
<td>LOGICAL</td>
<td>Can take the values: TRUE or FALSE.</td>
</tr>
<tr>
<td>DATE</td>
<td>Date (MM/DD/YYYY).</td>
</tr>
<tr>
<td>OBJECT</td>
<td>Graphical object POINT, REGION, LINE, POLYLINE, ARC, RECTANGLE, ROUNDED RECTANGLE, ELLIPSE, TEXT, or LAYOUT FRAME.</td>
</tr>
<tr>
<td>ALIAS</td>
<td>Column name.</td>
</tr>
<tr>
<td>PEN</td>
<td>Pen style (line).</td>
</tr>
<tr>
<td>BRUSH</td>
<td>Brush style (fill).</td>
</tr>
<tr>
<td>FONT</td>
<td>Font (text).</td>
</tr>
<tr>
<td>SYMBOL</td>
<td>Symbols (Point marking).</td>
</tr>
</tbody>
</table>
Object-based programs like ArcView include the FIELD object to convert SQL data into objects. Table data types include:

<table>
<thead>
<tr>
<th>DATA TYPES</th>
<th>DESCRIPTION AND RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>#FIELD_BYTE</td>
<td>Integers between 0 and 255 (inclusive) with length of one byte.</td>
</tr>
<tr>
<td>#FIELD_CHAR</td>
<td>Fixed length strings.</td>
</tr>
<tr>
<td>#FIELD_DATE</td>
<td>Date (YYYYMMDD).</td>
</tr>
<tr>
<td>#FIELD_DECIMAL</td>
<td>Real number in decimal format.</td>
</tr>
<tr>
<td>#FIELD_DOUBLE</td>
<td>Real number with double precision.</td>
</tr>
<tr>
<td>#FIELD_FLOAT</td>
<td>Real number with single precision.</td>
</tr>
<tr>
<td>#FIELD_ISODate</td>
<td>Date and hour in string format (YYYY-MM-DDhh:mm:ss).</td>
</tr>
<tr>
<td>#FIELD_ISOTime</td>
<td>Hour in string format (hh:mm:ss).</td>
</tr>
<tr>
<td>#FIELD_LOGICAL</td>
<td>Can take the values: TRUE or FALSE.</td>
</tr>
<tr>
<td>#FIELD_LONG</td>
<td>Long integer.</td>
</tr>
<tr>
<td>#FIELD_MONEY</td>
<td>Real number in string format with two decimal digits.</td>
</tr>
<tr>
<td>#FIELD_SHAPELINE</td>
<td>Graphic object. Polyline.</td>
</tr>
<tr>
<td>#FIELD_SHAPEMUTIPOINTER</td>
<td>Graphic object. Multiple points.</td>
</tr>
<tr>
<td>#FIELD_SHAPEPOINT</td>
<td>Graphic object. Point.</td>
</tr>
<tr>
<td>#FIELD_SHAPEPOLY</td>
<td>Graphic object. Polygon.</td>
</tr>
<tr>
<td>#FIELD_SHORT</td>
<td>Short integer.</td>
</tr>
<tr>
<td>#FIELD_UNSUPPORTED</td>
<td>Unsupported data type.</td>
</tr>
<tr>
<td>#FIELD_BLOB</td>
<td>Object in binary format up to two gigabytes. It is used for multimedia formats (images, audio, video, etc.).</td>
</tr>
<tr>
<td>#FIELD_VCHAR</td>
<td>Variable length string character.</td>
</tr>
</tbody>
</table>

**System Catalog**

Also called data dictionary. It could be considered a database in itself (a system database, not a user database). Its contents represent data about the data (metadata), this is, definitions of other objects in the system and not raw data. Catalog structures differ among DB systems. Date (Date, 1994) proposed that a minimal version of a catalogue is formed by:

- SYSTABLES. The set of tables that form the database (or project). It has a row for each named table (base table or view).
- SYSCOLUMNS. It includes the attributes of all the tables that form the database and that are mentioned in SYSTABLES. It provides a name for each of the columns (attributes), as well as the name of the table to which it belongs, and the type of data in the column.
- SYSINDEX. It includes all the database indices. It provides a name for each index in each row, as well as the name of the indexed table, and the index code.

MapInfo does not include a system catalogue but provides a workspace that serves similar purposes.

ArcView provides a system catalogue through projects (project) that contain information on the following metadata: View, Tables, Chart, Layout, and Script.

In summary, data processing, analysis and interpretation are the most complex tasks to be performed before decision-making. Prior to analysis, data must be subject to validation, standardization and normalization. The RM are powerful tools to facilitate data processing and analysis. These capacities are enhanced by GIS, since they allow management of both spatial and non-spatial data.
Additional examples of applications of the SQL in epidemiology

Some examples of SQL applications to epidemiology are given below, using the same tables from previous examples:

Get the number of cases by Region:
SELECT REGION.IDREGION, SUM(CASES.CASES)
FROM REGION, CASES
WHERE REGION.IDMUNICIP = CASES.IDMUNICIP
GROUP BY REGION.IDREGION

Calculate prevalences by municipality:
SELECT POPULATION.IDMUNICIP, SUM(CASES.CASES) / POPULATION.POPULATION
FROM REGION, CASES
WHERE REGION.IDMUNICIP = CASES.IDMUNICIP
GROUP BY REGION.IDMUNICIP

Calculate the variance of the number of cases by municipality:
SELECT IDMUNICIP, SUM((CASES - AVG(CASES))^2) / COUNT(CASES)
FROM CASES
GROUP BY IDMUNICIP

This example uses a valid SQL expression that is too complex for MapInfo to process. The solution is to divide the query into three queries, following a logical order, as follows:

First query:
SELECT IDMUNICIP, AVG(CASES) "Average"
FROM CASES
GROUP BY IDMUNICIP
INTO TABLE AVG

If using MapInfo, the name in quotation marks after SELECT will be the name of that column in the result table. In standard SQL, the word AS should be included in addition to the name in quotation marks. The name after INTO will be the table name.

Second query:
SELECT CASES.IDMUNICIP, SUM((CASES.CASES - AVG.AVERAGE)^2) "NUMERATOR"
FROM CASES, AVG
WHERE CASES.IDMUNICIP = AVG.IDMUNICIP
GROUP BY IDMUNICIP
INTO TABLE NUM

Third query:
SELECT IDMUNICIP, NUM.NUMERATOR / COUNT(CASES) "VARIANCE"
FROM CASES, NUM
WHERE CASES.IDMUNICIP = NUM.IDMUNICIP
GROUP BY IDMUNICIP

Calculate the percentage of cases by region:
First, calculate the total sum of all cases from all the units, municipalities, and regions, using the query:
SELECT SUM(CASES) "TOTAL"
FROM CASES
INTO TOTCASES

The next step is to generate a table with aggregated data by region for all the regions and to calculate the percentage by adding the total number of cases of a given region, dividing it by the total number of cases from the previous query, and multiplying it by 100. This is accomplished by:
SELECT REGION.IDREGION, (SUM(CASES.CASES)/TOTCASES.TOTAL)*100
FROM REGION, CASES, TOTCASES
WHERE REGION.IDMUNICIP = CASES.IDMUNICIP
GROUP BY REGION.IDREGION
Glossary

Set
A collection of entities, called the elements or members of the set. A set may be an element of another set. A set with no elements is called the null or empty set. Two sets are equal if and only if they have the same elements (Halmos, 1965).

System
A group of elements that interact to accomplish a common goal (Senn, 1989). These elements form a complex whole; each of them has its own function and may be part of another system. The absence of any one of these interdependent elements may alter the functionality of the whole. A system is always part of a larger system.

Database
A collection of data organized for ease and speed of search and retrieval. It intends to abstract a portion of reality (Wertz, 1993).

Database Management System (DMBS): It consists of a set of interrelated data (DB) and a set of computer software programs to manage data (Date, 1994).

Data model (DM)
A descriptive representation of data structure (Korth, 1991). Data models are represented using different tools, such as diagrams and other types of visual display of data and their relations.

Relational Model (RM)
A paradigm of DBMS to model data, based on a fixed format record structure; it consists of fields or attributes that usually have a fixed length. It represents data and their relations, by means of a set of tables with columns that take unique names (Korth, 1991).

Relational DataBase (RDB)
A set of relations (tables) that receive a unique name (Korth, 1991).
Figure 3.20

Figure 3.21
Appendix: MapInfo functions

The tables below include the functions that can be utilized in MapInfo SQL clauses, presented according to the type of data they handle.

**Mathematical functions:**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs( num )</td>
<td>Returns the absolute value of a number.</td>
</tr>
<tr>
<td>Cos( num )</td>
<td>Returns the cosine of a number; num is in radians.</td>
</tr>
<tr>
<td>Int( num )</td>
<td>Returns the whole portion of a number.</td>
</tr>
<tr>
<td>Maximum( num, num )</td>
<td>Returns the bigger of two numbers.</td>
</tr>
<tr>
<td>Minimum( num, num )</td>
<td>Returns the smallest of two numbers.</td>
</tr>
<tr>
<td>Round( num1, num2 )</td>
<td>Returns a number (num1), rounded to the nearest value of num2 (ex. if num2 is 10, then num1 is rounded to the whole number closest to 10).</td>
</tr>
<tr>
<td>Sin( num )</td>
<td>Returns the sine of a number; num is radians.</td>
</tr>
<tr>
<td>Tan( num )</td>
<td>Returns the tangent of a number; num is radians.</td>
</tr>
</tbody>
</table>

**Date functions:**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CurDate( )</td>
<td>Returns the current day.</td>
</tr>
<tr>
<td>Day( date )</td>
<td>Returns the day of the month (1 - 31) part of the date.</td>
</tr>
<tr>
<td>Month( date )</td>
<td>Returns the month (1 - 12) part of the date.</td>
</tr>
<tr>
<td>Weekday( date )</td>
<td>Returns the day of the week (1-7) part of the date; 1 is Sunday.</td>
</tr>
<tr>
<td>Year( date )</td>
<td>Returns the year part of the date. (Ex. 1994).</td>
</tr>
</tbody>
</table>

**String character functions:**

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr$( num )</td>
<td>Returns a character corresponding to the indicated num character code (ex. Chr$(65) returns the character “A”).</td>
</tr>
<tr>
<td>DeformatNumber$(str)</td>
<td>Undoes the effect of the function FormatNumber$ and returns a string character that does not separate hundredths.</td>
</tr>
<tr>
<td>Format$( num , str )</td>
<td>Returns a string character representing a number with a given format. Ex. Format$(12345.678, &quot;$,.##&quot;) returns &quot;$12,345.68&quot;.</td>
</tr>
<tr>
<td>FormatNumber$( num )</td>
<td>Returns a string character representing a number with split hundredths. This simplifies the function Format$, but control over the format is lost.</td>
</tr>
<tr>
<td>InStr( num , str1 , str2 )</td>
<td>Searches the string character str1 starting at the character that has the num position, and checks the presence of the string character str2. Returns the position where str2 was located and zero if not found. To search from the beginning use a num of value one (1).</td>
</tr>
<tr>
<td>Lcase$( str )</td>
<td>Returns a string character that is the lowercase version of the string character str.</td>
</tr>
</tbody>
</table>
String character functions (continued):

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left$( str , num )</td>
<td>Returns the first num characters of the str string, taken from left to right.</td>
</tr>
<tr>
<td>Len( str )</td>
<td>Returns the number of characters of the str string.</td>
</tr>
<tr>
<td>Ltrim$( str )</td>
<td>Deletes blank spaces to the left of the str string and returns the resulting string.</td>
</tr>
<tr>
<td>Mid$(str, num1, num2)</td>
<td>Returns a portion of the string character str starting with the character in position num1 and taking the next num2 characters.</td>
</tr>
<tr>
<td>Proper$( str )</td>
<td>Returns a string character where each word begins with an uppercase letter.</td>
</tr>
<tr>
<td>Right$( str , num )</td>
<td>Returns the last num characters of the str string.</td>
</tr>
<tr>
<td>Rtrim$( str )</td>
<td>Deletes blank spaces to the right of str string to the end and returns the resulting string.</td>
</tr>
<tr>
<td>Str$( expr )</td>
<td>Transforms an expr expression into a string character.</td>
</tr>
<tr>
<td>Ucase$( str )</td>
<td>Returns the uppercase version of string character str.</td>
</tr>
<tr>
<td>Val( str )</td>
<td>Returns the numeric value of a string; ex. Val(&quot;34&quot;) returns the number 34.</td>
</tr>
</tbody>
</table>

Spatial object functions:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area( obj , str )</td>
<td>Returns the area of obj object. The parameter str specifies the area unit, for example, &quot;sq mi&quot; or &quot;sq km&quot;.</td>
</tr>
<tr>
<td>CentroidX( obj )</td>
<td>Returns the coordinate x of the centroid of obj object.</td>
</tr>
<tr>
<td>CentroidY( obj )</td>
<td>Returns the coordinate y of the centroid of obj object.</td>
</tr>
<tr>
<td>Distance(num_x, num_y, num_x2, num_y2 , str )</td>
<td>Returns the distance between two points. The first two parameters specify the value of x and the value of y of the first point, the next two parameters give the value of x and the value of y of the second point. The parameter str indicates the distance unit, ex. &quot;mi&quot; or &quot;km&quot;.</td>
</tr>
<tr>
<td>ObjectLen( obj , str )</td>
<td>Returns the length of obj object. The value of str specifies the distance unit in &quot;mi&quot; or &quot;km&quot;. The only objects of length equal to zero are lines, polylines and arcs.</td>
</tr>
<tr>
<td>Perimeter( obj , str )</td>
<td>Returns the perimeter of obj object. The value str specifies the distance unit in &quot;mi&quot; or &quot;km&quot;. Only regions, ellipses, and rectangles have perimeters greater than zero.</td>
</tr>
</tbody>
</table>
Appendix: Relational Algebra (RA)

RA consists of a set of eight high-level operators that handle the relations. Each operator takes two relations as input and generates a new relation as a result. These operators are presented in the graphic below.

**Restriction**
It retrieves the specified tuples (rows) in a given relation, this is, it restricts the relation by selecting only the tuples that meet a given condition.

**Projection**
It retrieves the specified attributes in a given relation, this is, it selects only those columns that will be part of the result relation.

**Product**
Using two specific relations, it creates a relation having all the possible combinations of tuples of each of the tuples of the two relations.

**Union**
It creates a relation formed by tuples occurring in any of two specified relations.
**Intersection**
It creates a relation formed by tuples occurring in two specified relations.

**Difference**
It creates a relation formed by all the tuples in the first relation that are not in the second relation, from two input relations.

**Reunion**
From two specified relations, it creates a relation having all the possible combinations of tuples, one from each of the two relations, so that the two tuples in a combination meet a given condition. This is, the tables will be linked through the comparison of column values of each table.

**Division**
It takes two relations; a binary (two attributes) and a unary (one attribute) relation, to create a relation formed by all the values of one attribute of the binary relation that correspond (in the other attribute) with all the values of the unary relation.

Note that the result of each operation is another relation, thus, a result may be the operand in another operation. This means that it is possible to write nested relational expressions.
References


(4) Codd EF. “A Relational Model of Data for Large Shared Banks”. *CACM (13)* 6. 1970


(6) Halmos PR. *Teoría Intuitiva de los Conjuntos*. CECSA 1965


(8) Shakuntala A. *Técnicas de Bases de Datos*. Trillas. 1988


(10) ESRI. *User’s Manual ArcView (Training Course)*. 1994