

Introduction to GIS

VÍCTOR OLAYA

Introduction to GIS

Foreword by GRETCHEN PETERSON

Introduction to GIS

Text: Copyright ©2018 Víctor Olaya

Foreword: Copyright ©2018 Gretchen Peterson

Cover Image: The Art Journal The Industry of All Nations Illustrated Catalogue
(London, England: Bradbury and Evans, 1851)

Last update: June 13, 2018

This book is distributed under a Creative Commons Attribution license.

FOREWORD

I first met Victor when we were working at the same company. During that time I learned a few interesting things about him. Like how he works at such a rapid pace that if you blink you might find that he's written a new plugin for QGIS or even that he's written a book like this one. Aside from these great qualities, the thing I most remember about him is that he helped direct me to the last packet of hot chocolate in the office kitchen, after a day full of meetings when I needed it the most. It's helpful things like that which make a difference to people. And in this book you will find so many helpful things, akin to that hot chocolate but for Geographic Information Systems (GIS), organized in a thoughtful manner which will help you get through that sometimes-long GIS slog.

This book is an excellent reference text regarding the history and basics of GIS. It includes clear examples of concepts illustrating choices the geospatial professional must make in design and layout and how those choices affect a map product. The reader can literally see how decisions about line, color, shape, and other qualities will render a map that is the most useful and the most aesthetic. It also includes important information about the various

ways in which GIS data is obtained, how it is stored, and a great overview of GIS software.

The book begins with the history of GIS and proceeds into sections that discuss and define such topics as spatial analysis, data visualization, web mapping and data sources, among many others. I envision the book being used as a teaching tool, both in a formal setting and for self-learners. Additionally, for more experienced geospatial professionals, this book can be used in the initial ideation phase of creating a map, reminding us of the elements we need to consider and prioritize to meet the objectives for a particular map or analysis. It is really a digital pocket guide to GIS.

Victor is generously making his book available to all, free, for users. Knowing the hours of work that go into any book, I appreciate his attitude of community and contribution to the field of GIS. Learning and continually revisiting the fundamentals is paramount for success in our field. So pour yourself a good cup of hot chocolate and get started.

GRETCHEN PETERSON
Co-author of QGIS Map Design

PROLOGUE

When in 2005 I started writing my book “Sistemas de Información Geográfica”, I did it for two reasons: first, because no books on GIS theory had been published in Spanish since the early 90’s; second, because there were no free books about GIS, except those related to free GIS software, which included little theoretical content.

It took me five years to write the book, which ended up being a complete reference book with almost a thousand pages. Knowing that its size and its level of detail could be intimidating, and that many people would prefer a shorter version, in 2015, I wrote “Introducción a los SIG”. The book you are reading now is the English translation of that shorter work.

Unlike what happens in Spanish, there are many good books on GIS theory written in English, and new editions are published constantly to update them with the latest changes in the field of GIS. However, no free book (that is, no book that can be freely copied, printed and distributed) on this topic had been published yet.

I believe this book will be of great use for current GIS users and for anyone wanting to start in this fascinating field of GIS. If you have any suggestions or comments, you can contact me at: voleyaf@gmail.com.

INTRODUCTION TO GIS

WHAT IS GIS?

Most of the information that we use nowadays is georeferenced. That is, it is information to which a geographical position can be assigned, and it is thus information that has some ancillary information related to its location.

A **Geographical Information System** (GIS) is a tool to work with georeferenced information. In particular a GIS is a system that allows the following operations:

- **Reading, editing storing**, and, generally speaking, **managing** spatial data.
- **Analyzing** those data. This includes everything from simple queries to complex models, which can be performed using the **spatial component** of the data (the location of each value or element), the **thematic component** of the data (the value or element itself), or both.
- Generating **documents** such as maps, reports, plots, etc.

GIS is a step beyond traditional maps. A map represents a rendering of a set of spatial data, and while this rendering has great importance within GIS, it is but one of its many components. GIS includes not only data and their rendering, but also all the operations that can be performed on them which are part of the system, also.

GIS is a flexible and versatile tool and most disciplines today use GIS in one way or another. One of the main reasons for this is the integrative nature of GIS. The following are some of the main contexts in which GIS plays this integrative role.

- **GIS as a tool to integrate information.** A common link between most disciplines is that they study something which can be located. This allows for combining and getting results from a joint analysis. In this context, GIS provides the framework on which information from different disciplines can be added and we can work with it.
- **GIS as a tool to integrate technologies.** A large part of the technologies that have appeared in the last several years (and most likely those that will appear in the near future) are based on using spatial information and are connected to some extent to GIS to extend their capabilities and their reach. Due to its central position in this group of technologies, GIS plays an important role in linking them and allowing them to communicate around its own functionalities.
- **GIS as a tool to integrate technologies.** GIS functionalities cover a broad range of users, most of whom would not have such a well-defined framework if it were not for GIS itself. Consequently, there is better coordination among them.
- **GIS as a tool to integrate theoretical areas.** We can understand GIS as the sum of two disciplines: geography and computer science. However, a more detailed analysis reveals that GIS incorporates elements from many different scientific fields, such as those related to technology and data management (computer science, database design, digital image analysis), those that study the Earth from a physical point of view (geology, oceanography,

ecology) or from a social and human one (anthropology, geography, sociology), those that study human behavior and understanding (psychology), or those that have themselves traditionally integrated knowledge from different fields, such as the already mentioned geography.

The term **geomatics**, derived from *geography* and *informatics*, frequently refers to the array of scientific areas related to GIS.

Therefore, we see that GIS integrates technology, informatics, people and geographical information, of which the main purpose is to capture, analyze, store, edit and visualize georeferenced data.

From a different point of view, a GIS can be considered as composed of five main elements:

- **Data.** Data is needed for the rest of the components to make sense and be able to serve a given purpose. Geographical information, the core of GIS, lives in the data, and a detailed knowledge of the data that we use, its quality, its origin, its characteristics, and how to manage and store it is paramount to correctly understand GIS itself.
- **Analysis.** Analysis is one of the main strengths of GIS, and one of the reasons why the first GIS were developed. Most GIS include analysis capabilities. They include methods that were already used with traditional cartography, others that existed but were not feasible to use without computers, and new approaches that were developed specifically after GIS appeared.
- **Visualization.** All types of information can be represented graphically which makes it easier to interpret it. In the particular case of geographical information, visualizing it is not only a different way of working with that

information, but indeed the main one, since it is the one to which we are more accustomed.

While maps are graphical entities, in GIS, we work with raw alphanumeric data. In order to have the same capabilities of a printed map, GIS must be able to create visual representations from that data, including map-like ones. The same cartographic principles that apply when designing a printed map are also valid when rendering geographic data within GIS, and GIS users must be familiar with them.

- **Technology.** This includes both the GIS software and the hardware that runs it. Additional elements that are common when working with GIS data, such as peripherals used for entering data or for creating printed cartography, are included here.
- **Organization.** This includes the elements that ensure a proper coordination between people, data and technology. As GIS gets more complex, managing the relations among its elements becomes more important.

In the following chapters, we will describe these elements in detail.

HISTORY OF GIS

GIS has experienced a huge development since its early days. With the popularization of GIS technologies, and thanks to the help of all other disciplines that use GIS and rely on it, the field of GIS has been redefined and expanded, especially in the last years.

We can locate the origins of GIS in the sixties, when the first GIS applications appeared. The two main reasons for this were the **increasing need of geographical information** and the **appearance of the first computers**.

The theoretical foundation of GIS was laid a few years before, with the development of new approaches in the field of cartography, such as **quantitative cartography**, which seemed to predict the future needs that the use of computers and geographical data would bring.

The first relevant experience that combined computers and geography can be found in 1959, when Waldo Tobler defined the principles of a system called MIMO (map in-map out), with the purpose of applying computers to the field of cartography. He defined the basic ideas for creating, encoding, analyzing, and rendering geographical data within a computer system.

The first GIS was the CGIS (Canadian Geographical Information System). It was developed in Canada in the early sixties by Roger Tomlinson, who is popularly known as the “father of GIS”.

In the mid-sixties, two applications, SYMAP and GRID, laid out the theoretical foundation for the analysis of **raster** and **vector** data, the two main approaches for encoding and storing geographical information (we will explain them in detail in the upcoming chapters). The main ideas for performing analysis in raster GIS were defined by Dana Tomlin with his **map algebra**.

During the sixties, the field of GIS starts developing itself from those seminal works. GIS is not anymore an experimental tool, and it starts to become an important part of the cartographic world.

From this moment, GIS evolves through several different periods, moving very fast thanks to the influence of many external factors. This evolution affects the discipline of GIS itself, the technology it involves, the data, and also the theories and techniques it is built on.

THE EVOLUTION OF GIS AS A DISCIPLINE

At first, GIS was just a combination of ideas from quantitative cartography, and the computer systems that existed at that time. It was basically the work of cartographers and geographers who tried to adapt their knowledge and their needs to a technology that looked promising. Since then, a large number of other disciplines have contributed to the field of GIS and their contributions are as important, or in some cases even more so, than those of cartography and geography.

More or less at the same time, society was becoming more concerned about the environment and the effect of

human actions on it. This influenced GIS which was becoming a fundamental tool for all tasks related to environmental management (land-use planning, environmental monitoring, etc.), and boosted its development.

At the beginning of the seventies, once it was clear that GIS had a great future ahead, the field of GIS started to shape its identity and to become a solid discipline. The first conferences and symposiums about GIS took place and GIS was already included in University *curricula*. Specialized journals and forums appeared in the eighties and helped spread GIS to a wider audience. The industry of GIS consolidated itself in the seventies. **ESRI** (Environmental Systems Research Institute), pioneer and current leader of the GIS market, was founded in 1969, and its products have played a key role in the popularization of GIS. The first open-source GIS, **GRASS** (Geographic Resources Analysis Support System), appeared in 1985.

The beginning of the 21st century marks a turning point in the history of GIS, as it reaches non-professional audiences. Cartography services such as **Google Maps** allow users with little or no technical GIS knowledge to interact with a GIS application and use it. **GPS navigators**, which include both analysis and rendering capabilities that come from GIS, are another good example of this.

THE EVOLUTION OF TECHNOLOGY

The evolution of computers has affected GIS. Three are the main areas that have had a major influence in shaping GIS as we know it now.

- **Graphical outputs.** The capabilities of computers to generate graphical outputs have greatly improved since their beginnings, and they are still evolving. GIS has

followed this evolution closely, both for screen rendering and for the case of printed outputs.

- **Data access and storage.** The size of GIS datasets has increased enormously, and using these large datasets would not be possible without the corresponding improvements in both data storage and data access.
- **Data input.** In the early days of GIS, data were manually digitized. Nowadays, creating data that can be used in a GIS is a completely different process, and it uses specific hardware such as high-resolution scanners, or specific software such as the one used for automatic digitalization of pattern recognition based on images, all of which generate ready-to-use data.

Along with this, software has changed following the evolution of computers themselves, from mainframes to personal computers, and more recently, to other platforms such as tablets or mobile phones.

By the end of the eighties, cartography can be efficiently produced in personal computers, with a comparatively low cost, without the need of expensive and dedicated large mainframes.

Nowadays, the combination of positioning systems such as GPS with mobile platforms is playing an important role in the development of GIS, in areas such as data collection.

The Internet also changed GIS, much like it changed every other field, whether scientific or not. In 1993, *Xerox PARC*, the first **map server** to distribute cartography over the Internet, was created. The first digital on-line atlas, the Canadian National Atlas, has been available since 1994. More recently, the ideas of the Web 2.0 are adapted to the field of GIS and contribute to the development of what is now known as **Web Mapping**.

THE EVOLUTION OF DATA

The first geographical datasets used in GIS contained just **scanned maps** and **digitized features** obtained from them. Since then, new data sources have been constantly appearing, with formats that are better adapted to GIS, and with GIS itself adapting to them as well. As a consequence of that, the amount, precision, and quality of data that is now available to be used in a GIS has dramatically increased.

The launching of the first **earth observation satellites** represents a key advance. The techniques that were already in use for aerial photography, developed mostly during the First World War (although the discipline goes back to the second half of the 19th century, when photos were taken from hot air balloons), are applied on a global scale when the first satellites are created. SPOT Image, the first commercial company to distribute satellite images that cover the entire globe, was created in 1982.

Positioning technologies are another important data source for GIS. In 1981, the GPS system became completely operative, and in 2001, its accuracy for civil use was increased.

As it happened with GIS software, digital geographical data becomes more popular and receives more attention. In 1976, the United States Geological Service (USGS) publishes its first **Digital Elevation Models** (DEM), in response to the high relevance that this type of data now had in the context of geographical analysis. In 2000, elevation data from the *Shuttle Radar Topographic Mission*(SRTM) is released to the public, covering 80% of the Earth's surface with a resolution of one arc second (about 30 meters).

The development of techniques such as **LiDAR**, which can be used to get elevation data with much more detail,

opens a large array new possibilities for areas such as terrain analysis.

The evolution of data is not just technical, but also **social and organizational**. As the amount of data increases, it becomes clear that new strategies must be developed for managing those data. So-called **Spatial Data Infrastructures** are developed as a result of this. The most relevant of them is the United States National Spatial Data Infrastructure (NSDI), created in 1994. In Europe, the INSPIRE directive serves a similar purpose.

Many of these activities and developments follow the specifications set up by the **Open GIS Consortium** (OGC), and international consortium founded in 1994, which works to **homogenize and standardize** the use and distribution of geographical data.

THE EVOLUTION OF THEORIES AND TECHNIQUES

Once the first GIS was implemented and could respond to the data management and analysis needs for which they were created, new techniques and approaches began to be developed.

Spatial analysis is a comparatively recent field. In 1854, **John Snow** performed what is usually considered one of the first examples of analytical cartography, when he used a map to determine the source of a cholera outbreak in London.

In his book *Design with Nature* (1969), Ian McHarg defined the basic ideas about **map overlays**, which, as we will later see, are fundamental for the analysis and visualization of geographical data **layers** within a GIS.

Terrain analysis is another field that has experienced a huge qualitative change thanks to GIS. Traditional terrain analysis, mostly based on geology and geomorphological

analysis, developed into a quantitative science focused on the morphometric analysis of relief.

Along with the analytical component, cartography also evolved in the context of GIS. In 1819, Pierre Charles Dupin created the first **choropleth map**. With the arrival of GIS, this type of map will become very popular.

The advances in Computer-Assisted Design (CAD) applications and in-screen rendering techniques helped in defining a new discipline: computational geometry. GIS vector analysis is based on it.

FUNDAMENTALS OF CARTOGRAPHY AND GEODESY

Since GIS inherits concepts and ideas previously used to create printed maps, it is mandatory to know them in order to correctly use the tools included in a GIS. The fundamental concepts from cartography and geodesy are the most important ones. Without them, it is not possible to understand GIS.

BASIC CONCEPTS OF GEODESY

The main property of georeferenced information is that it has a **location**, and more particularly, a location on the earth. This location is given with **coordinates** that define it, which requires a reference system for the coordinates.

Geodesy is the science that provides the theoretical framework for this, and it studies the **Earth's shape**. Geodesy, through its different branches, provides methods and concepts that allow defining and using precise and rigorous coordinates to locate elements and phenomena that take place on Earth.

Geodesy is needed due to the fact that the Earth is not flat, and when the area that is studied is large enough, the

effect of the Earth's curvature cannot be ignored. For this reason, GIS implement the required elements to manage geographical information, taking into account the ideas and principles of geodesy.

One of the main purposes of geodesy is to establish a reference system and define a set of points (known as **geodesic vertices**), whose position is known with a high level of accuracy. Based on those points, which form a **geodesic network**, coordinates for any point on the Earth's surface can be computed.

Reference surfaces

To accomplish this, geodesy defines two basic reference surfaces: **reference ellipsoid** and **geoid**.

Earth has a spherical shape. However, it is not a perfect sphere, but is instead what is called an **ellipsoid**. In an ellipsoid, the radius is not constant and depends on the location over its surface. Using an ellipsoid to define the Earth's shape is more precise than assuming it has a spherical shape, and is needed to create accurate cartography, especially when the represented surface is not too large.

The ellipsoid provides a theoretical expression of the Earth's shape, and the next step is to determine the parameters that define it. In the case of a sphere, the only parameter needed is the radius. In the case of an ellipsoid, two parameters have to be determined: the length of semi-major and semi-minor axis.

For historical reasons, many ellipsoids exist, all of them derived from the work of geodesists in different times and places. The first general ellipsoids, which can be used for representing any place on Earth's surface, appeared approximately a hundred years ago, created as an international

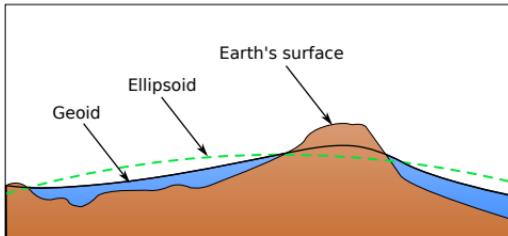


Figure 3.1: Comparison of the three fundamental surfaces: Earth's surface, geoid and ellipsoid.(Adapted from Wikipedia).

reference that can be used for creating cartography in different areas of our planet. The **WGS-84** ellipsoid is one of the most popular currently, and it is used by the GPS positioning system.

The other reference surface is the geoid, defined as the three-dimensional surface where every point have the same gravitational attraction. It is an equipotential surface that results from assuming average ocean levels and extending them under the Earth's surface.

As in the case of ellipsoids, there are several geoids as well. These are not constant and evolve to adapt to the changes that take place on the Earth's surface.

Figure 3.1 shows a comparison of the three surfaces: Earth's surface, geoid and ellipsoid.

In a **general ellipsoid**, both the location of its center of gravity and its equatorial plane match those of the Earth. In a **local ellipsoid**, this does not have to be true, and the ellipsoid by itself is not enough, since we do not know how to place it relative to the real Earth's surface.

The concept of **datum** solves this problem. A **datum** is the combination of a reference surface (the ellipsoid) and a point in which it is linked to the geoid. That point is called the **fundamental point**, and the ellipsoid is tangent

to the geoid there. At the fundamental point, a line perpendicular to the geoid is identical to a line perpendicular to the ellipsoid.

Coordinate reference systems

Once we have a model to define the Earth's shape, we can establish a system to code any position over its surface and assign a corresponding coordinate to it. The combination of a coordinate system and a datum is called a **coordinate reference system** (CRS).

Regarding the coordinate system, we have two main alternatives: using the elements of **spherical geometry** using the concepts of **plane geometry**. In the latter, we need a **projection system** to place the elements on the surface of the ellipsoid into a plane.

Geographical coordinates use a spherical coordinates system in which the location of every point is defined by two angular values: **latitude** and **longitude**. Lines of equal latitude are called **parallels**, while lines of equal longitude are called **meridians**.

Geographical coordinates are of great utility, especially when working with large regions. However, it is not a cartesian system, and **it is difficult to perform tasks such as measuring distances or areas**. To simplify operations like those, we need cartesian coordinates. To assign a plane coordinate to every point on the Earth's surface (which is not a plane), we must use a **cartographic projection**.

Earth's surface is not **developable**. That is, it cannot be flattened without distortion. For this reason, we need a methodology for converting points on this surface into points on a plane. Figure 3.2 shows this idea.

In the case depicted in the figure, point are projected directly onto the plane. Another alternative is to project

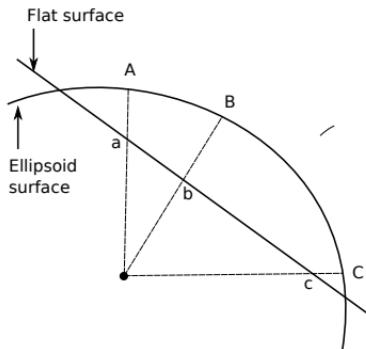


Figure 3.2: Graphical explanation of a projection. Points A , B and C on the surface of the ellipsoid are converted into their equivalent points a , b and c on a plane.

them onto a surface that, unlike the surface defined by a sphere or an ellipsoid, can be developed (that is, it can be flattened later without distortion). The most usual surfaces for that are the cylinder and the cone. The corresponding projections are called **conical projections** and **cylindrical projections**.

It can be seen in the figure that projecting points introduces distortions. For instance, the distance between points A and B is not the same as the distance between points a and b . All projections introduce some sort of distortion, regardless of their properties. Depending on the metric properties that are preserved undistorted, we have **equal-area projections** (which preserve area), **conformal** (preserve angles and shapes) o **equidistant** (preserve distances).

Depending on the context and the purpose of our data, we might use one or another type of projection.

One of the most widespread projections nowadays is the **Universal Transverse Mercator**, which is the basis for the **UTM coordinate system**. This system is not just a

projection, but a complete system of many of them. Earth's surface is divided in rectangular regions, and for each of them a different projection and a different set of geodetical parameters are used. It uses a single ellipsoid: WGS-84.

In the UTM system, coordinate are not expressed as absolute coordinates, but instead the are refered to the corresponding rectangle, as relatives coordinates within it.

The UTM grid contains 60 zones, each 6° of longitude in width. Zone 1 is located between 180° and 174° West, and numbering increases eastward.

Each zone is segmented into 20 latitude bands, ranging from 80° South to 84° North. These are coded with letters from C to X, excluding I and O due to their similarity to the numerals one and zero. Each band has 8° of latitude in height, except the X band, which has 12.

A UTM rectangle is therefore defined by **a number and a letter**, and the coordinates that are used to locate a given point on the Earth's surface are referred to the zone to which it belongs. Coordinates are expressed in meters and represent the distance between the point and the origin of the UTM rectangle. The origin is located at the intersection between the meridian passing through the center of the zone and the equator.

To avoid negative numbers, the origin is assumed to have an X coordinate of 500000 meters and a Y coordinate of 10000000 meters, causing all coordinates referred to it to have only positive values.

Coordinate conversion and transformation

It is common when working with GIS to have layers in **several different coordinate systems**, or in the same coordinate system but using different parameters (such as a different datum). In order to be able to use those layers

together, we have to work in a single coordinate systems, and at least some of those layers will have to be converted to it. That is known as **coordinate conversion**. If the origin and destination coordinate systems have a different datum, coordinate conversion is called **coordinate transformation**.

In a GIS, conversion and transformation capabilities allow to generate new layers that use a different CRS. Also, GIS include the ability to perform them *on-the-fly* when layers are rendered, so we can create a map with layers that do not share the same CRS. These are correctly represented on the map and “match” one with another, since the GIS is automatically performing the corresponding changes to their coordinates to have them in a common CRS.

To facilitate the use of coordinate reference systems, there are initiatives that organize and code them so each system can be easily identified by a unique code, (called a **Spatial Reference System Identifier (SRID)**). The most common coding system is the one created by the European Petroleum Survey group (**EPSG**).

BASIC CARTOGRAPHIC CONCEPTS

Among the fundamental concepts of cartography that any GIS user has to know, **scale** is the most important one. The scale of a map represent the **size ratio** between the “map” that would be obtained by developing the real surface we are representing (the Earth’s surface in this case), and the scale of our smaller map. Knowing this ratio, we can know the real measures of the elements that are included in the map, since we can convert the measurements that we make on it into real-world measures. It’s important to keep in mind that these measures are not so “real”, since the projection might have distorted them, but they are,

nonetheless, measures at the original scale of the object that is measured.

Scale is usually expressed as a quotient between the distance measured in a map and the distance that this measure represents in reality. For instance, a 1:50000 scale means that 1 centimeter in a map is equivalent to 50000 centimeters in reality, that is 500 meters. This value is known as the **numeric scale**.

Regardless of the projection used, scale is completely true only at certain points in the map. In the rest of them, scale changes. The relation between the scale in those points and the numeric scale is known as the **scale factor**.

Although scale is traditionally understood as a concept related to the data representation, geographical data has an inherent scale not related to its representation, but to the level of detail with which the data was captured on the field. It's more correct to understand scale as something related to the data **resolution**, that is, related to the **minimum mapped size**.

The resolution of the human eye is 0.2mm. With that value, and the scale we want to use for a map, we can know the level of detail that we need to use when capturing data to be used as part of that map.

In GIS, as we have seen, data does not contain its visualization. That means we can visualize it at any scale (and that is very easy to do with the zoom tools that are found in most GIS software). However, data has been captured at a given level of detail. However, data has been captured at a given level of detail, which defines the scale meant to be used and represents a limitation of that data. It will not be correct to represent it beyond that scale. In other words, to create a map at a more detailed scale, we will need more detailed data. It's easy to forget that when using a GIS.

Raster layers, as we have seen, have a **cell size** which defines the resolution of the layer and is related to scale.

Related to the concept of scale, we find the so-called **cartographic generalization**. It means to express an idea or information in a more succinct manner, so it can be more useful in a given context. In a GIS, generalization is needed to represent data at a smaller (less detailed) scale than its inherent one, mainly because of the limitations imposed by rendering devices. For instance, if we have a layer with the roads of a given country, it makes no sense to use it in a map that represents the whole planet. We will get a mass of lines, and whoever uses that map will not be able to differentiate among them. Also, rendering all those lines will consume a lot of processing power. A much more interesting option would be to just use the main highways and motorways, and to not paint the rest of them. The map will be clearer and more useful, and the screen rendering will be much faster.

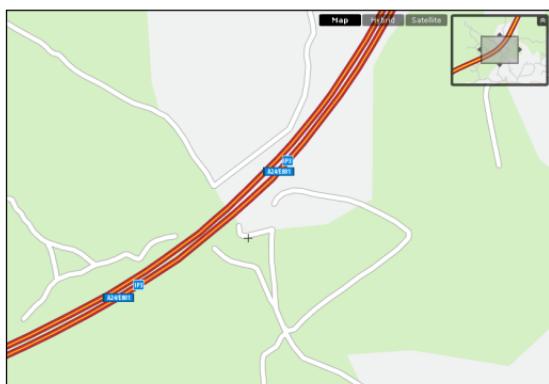


Figure 3.3: Generalization by aggregation. Two roads that are almost parallel are represented as two separate elements in the map, but in the overview map of the upper-left corner, with a smaller level of detail, they are generalized as one single element. (Taken from Yahoo Maps.)

Generalization in the context of a GIS also includes other changes that are made to improve the quality of the map and improve the way it conveys information. This might not imply reducing the level of detail, but just altering the data for purely cartographic purposes.

For instance, if we are representing the road layer in a map that covers the entire world, we should not paint the roads at their real size. They would be too thin and almost invisible. We will paint them much thicker, thus creating a map that might be less correct (we are distorting the real size of those roads), but is much more useful.

Generalization is, therefore, a process whose main purpose is **to produce a cartographic image more legible and expressive**, selecting and adjusting the elements contained in a the map. It emphasizes the important ones, while it omits the least important ones.

Some of the most relevant operations in cartographic generalization are **simplification** (representing an element that is less complex), **aggregation** (representing several elements as just one —Figure 3.3—), **exaggeration** (representing elements with a larger size) and **displacement** (representing elements at a different location, to ensure legibility).

In GIS, generalization can be implemented as part of the visualization mechanism. That is, when rendering a layer, it is modified at the same time according to the cartographic scale and other factors. This is a time-consuming procedure, and it usually does not yield good results, mainly because of the complexity of the process, which is hard to automate.

An alternative solution is to use a multi-scale approach (Figure 3.4). Information for a given study area is prepared at different scales (using generalization based on a single one or just using layers with a different origin), and the

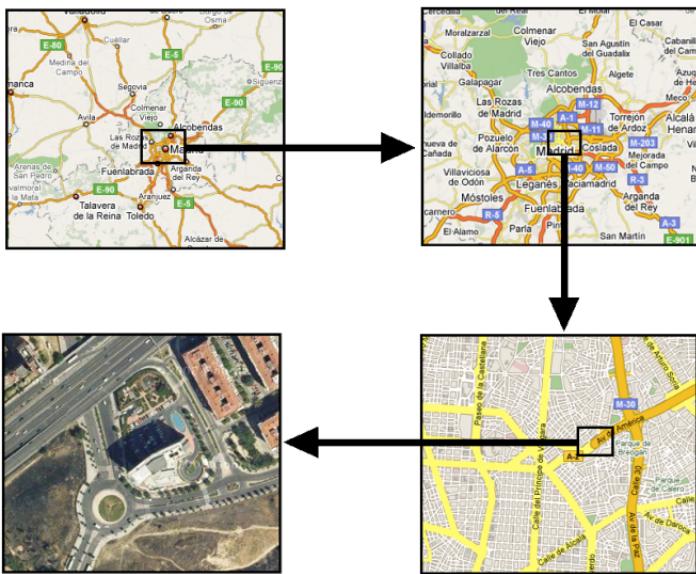


Figure 3.4: In a GIS, it is common to use information at different scales. The current cartographic scale will define which part of it rendered and which one is not.

most convenient one is used in each case depending on the current scale. This is the equivalent of having several printed maps at different scales.

The concept of **layer**, which we will see in the next chapter, is key for this multi-scale approach.

In the case of images, the approach implies creating so-called **pyramids**. Instead of using a single image, we have a set of them with different cell sizes. To optimize data handling and minimize the amount of data to be processed when rendering the image, the layer in the pyramid that best fits the current scale is used.

GEOGRAPHICAL DATA

From all GIS subsystems, data is probably the most important one. It is also the more interrelated, as it is linked to all of them and all depend on it to a certain extent. Data is the fuel that drives GIS.

DATA AND INFORMATION. TYPES OF INFORMATION.

There is a big difference between the concepts of **data** and **information**. A GIS is a Geographical *Information* System, but it uses geographical *data*.

Data is a **set of values or elements used to represent something**. For instance, the string 502132N is data.

We can interpret that data as being a geographical reference, in which case it could be a latitude value, in particular $50^{\circ}21' 32''$ North. If we interpret it as being a reference to an identity card (such as a driver's license) associated with a person, the information that we get is completely different. It is the same data, containing six digits and a letter, but the information that we extract from it is different, since we understand and interpret it differently.

Information is, therefore, the result of data and **its interpretation**. and in many cases, working with data means just trying to extract from it all the information that it might contain.

Understanding the meaning and the differences between data and information allows us to understand, for instance, why the ratio between the size of a given data and the amount of information it contains is not constant. The strings 502132NORTH and FIFTY TWENTY ONE THIRTY TWO NORTH are longer than 502132N, but they contain the same information (as long as we interpret them as the latitude component of a coordinate).

Geographical information has two separate components: **spatial** and **thematic**. The spatial component contains the position, referred to a given reference system, and it answers the question *where*. The thematic component answers the question *what* and it defines the characteristics of the phenomenon or feature that occurs at the location indicated by the spatial component.

While the spatial component is usually a numerical value (most coordinate systems use just numbers), the thematic component can be **numeric** or **alphanumeric**. A numeric variable can be itself of four different types: **nominal**, **ordinal**, **interval** or **ratio**.

The operations that can be performed on a certain geographical data are defined by the types of variables contained in its thematic component.

The different approaches for representing and storing geographical information, which we will consider later in this same chapter, depend on the type of variables that we are working with.

An important concept to consider related to geographical information is the **dimension**. The elements that we

store range from simple points (0D), to three-dimensional volumes (3D) (Figure 4.1).

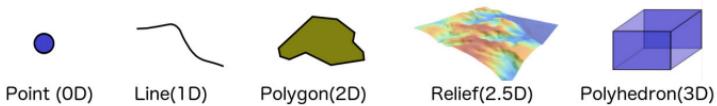


Figure 4.1: Dimensions of the spatial component of geographical data.

SUBDIVISION OF INFORMATION. LAYERS

In a GIS, the information about a given study area **is divided into several levels**. Even if it refers to the same location, the information about different variables is stored separately. That is, a set of different blocks of information exists for the same area, each of them containing a particular variable or set of elements. Each of these blocks is called a **layer**.(Figure 4.2).

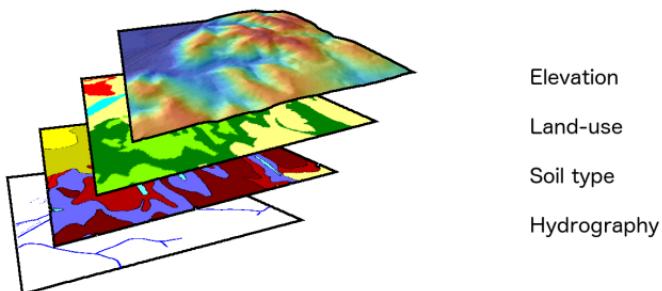


Figure 4.2: A graphical explanation of the concept of *layer*.

The concept of layer is fundamental to understand GIS and helps to correctly structure and manage geographical information. All the information that we will use in a GIS

will be in the form of layers. Each one of them can be used independently or along with others.

With traditional cartography, it is not possible (or it is complex and not accurate) to combine different types of information, such as the type contained in a topographic map and the type from a land-use map. In the case of a GIS, the different layers in which that information is contained can be combined in an easy and clean way.

The concept of layer as the fundamental unit for geographical information in GIS has a huge relevance, as it constitutes the basic framework for most operations. For instance, we saw in the section about cartographic generalization how in a GIS we can use different versions of the data corresponding to a given area, and show one or another depending on the current scale. These versions will be stored as independent layers. Layers are, thus, not just the fundamental unit for a given area, but also for a given scale, and allow us to optimally separate and organize geographical information.

Layers also help avoid data redundancy, since each layer just contains information about a particular variable or type of feature. A traditional map always contains a set of different variables, not just a single one. Some of them are used to provide a general context, such as the names of the main cities or the main roads, and these appear in most maps. In a GIS, they exist independently, and the user can add and combine them with other layers whenever is needed. Therefore, working with layers provides a more efficient approach and a **more atomized** organization of the data, with the advantages that it has for storage, management and use.

Apart from dividing geographical information in layers depending on its content, it is also divided considering purely spatial criteria, cutting it in smaller parts that cover

a smaller area. This is similar to what happens with traditional cartography, divided into **map sheets**.

The main feature of a GIS to transparently integrate data corresponding to different areas and create a seamless mosaic is the **separation between the data and its visualization**. Data is required for visualization, but these two elements constitute different parts of a GIS, with a clear separation between them. That means that data is used to create a visual output, but data itself does not contain any value related to its rendering and visualization.

Thus, it is possible to combine data and then represent that combination together as a whole. Something like that is not possible with a printed map, since it contains also the visualization elements (colors, line thickness, label placements, etc.) and even some additional cartographic ones (legend, scale, North arrow, etc.). Even if printed maps can be combined, information contained in them does not fuse to create a single new map. In a GIS, on the other hand, visualization of several blocks of data can be identical to the one that would be obtained if that data were stored as a single block.

GEOGRAPHICAL INFORMATION MODELS

The process of converting a given geographical area and the information about it in data that can be used within GIS can be divided into three different phases.

- Establishing a **geographical model**. That is, a conceptual model of a reality and its behavior.
- Establishing a **representation model**. That is, a way of coding the conceptual model, reducing it to a finite set of elements.
- Establishing a **storage model**. That is, a storage strategy for storing the elements of the representation model.

Representation models are the most important ones, and we will focus on them here. The two main representation models are the **raster model** and the **vector model**. Layers using these models are commonly known as **raster layers** and **vector layers**.

Raster model

The raster model is based on a **systematic division of space**. The whole space is characterized by a set of elements that cover it, each of them with an associated value.

The most common raster model is based on a grid of **square cells**, or sometimes rectangular ones. Knowing the orientation of the grid, the size of the cells (which is the same for all of them), and at least the coordinates of one of them, it is possible to know the location of all cells, thanks to its **regular structure**. With that, the values of the variable we are working with are known in all points of the area covered by the layer. The **cell size** is a parameter related to the scale of the layer, since it defines its resolution and depends on the level of detail used when the corresponding measures were taken.

Figure 4.3 shows an example of a raster grid.

10	16	23	16	9	6
14	11	18	11	18	19
19	15	13	21	23	25
20	20	19	14	38	45
24	20	20	28	18	49
23	24	34	38	45	51

Figure 4.3: Cells in a raster grid with their associated values.

The number of values stored for each cell defines the number of **bands** in a raster layer. A band contains a single value for each cell. We can understand a raster layer with more than one band as a set of sublayers, all of them having the same spatial structure (extent and cell size), and wrapped as a single layer.

We can find a clear example of that in digital color images. A digital image is composed of a grid of values (called **pixels**), each of them with an associated color. In the most common case, that color is expressed with three values, corresponding to the intensity of colors red, green and blue, which, when combined, give the pixel color. That is, an image like that is a raster layer with three bands, each of them containing one of the red, green and blue components.

Another typical use of the raster model is for the so-called **Digital Elevation Models** (DEM), which contain the topography of a given area. These are always single band layers.

In most cases, the values of a raster layer are numerical, and GIS software is usually not adapted to handle other types of values in the thematic component of a raster layer. Due to this, raster layers can be seen as **matrices**, and the corresponding mathematical tools can be used for their analysis.

Vector model

The other main representation model is the vector model. In this model, there are no fundamental units that divide and cover the area that is modeled. Instead, the variability and characteristics of that area are modeled using **features**, which represent elements in which those characteristics do not change. The geographical part of a feature is made of

geometric primitives, and these can be of three different types: **points**, **lines** and **polygons** (Figure 4.4).

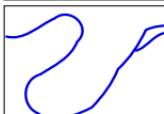
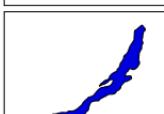
Primitive	Feature	Representation	Attributes																					
Points			<table border="1"> <thead> <tr> <th>ID</th><th>HEIGHT</th><th>DIAMETER</th></tr> </thead> <tbody> <tr><td>1</td><td>17.5</td><td>35</td></tr> <tr><td>2</td><td>22</td><td>45.6</td></tr> <tr><td>3</td><td>15</td><td>27.2</td></tr> <tr><td>4</td><td>19.7</td><td>36.1</td></tr> <tr><td>.</td><td>.</td><td>.</td></tr> <tr><td>.</td><td>.</td><td>.</td></tr> </tbody> </table>	ID	HEIGHT	DIAMETER	1	17.5	35	2	22	45.6	3	15	27.2	4	19.7	36.1
ID	HEIGHT	DIAMETER																						
1	17.5	35																						
2	22	45.6																						
3	15	27.2																						
4	19.7	36.1																						
.	.	.																						
.	.	.																						
Lines			<table border="1"> <thead> <tr> <th>WIDTH</th><th>DEPTH</th><th>LENGTH</th></tr> </thead> <tbody> <tr><td>15</td><td>4.3</td><td>35</td></tr> <tr><td>6.3</td><td>3.9</td><td>5.2</td></tr> </tbody> </table>	WIDTH	DEPTH	LENGTH	15	4.3	35	6.3	3.9	5.2												
WIDTH	DEPTH	LENGTH																						
15	4.3	35																						
6.3	3.9	5.2																						
Polygons			<table border="1"> <thead> <tr> <th>AREA</th><th>DEPTH</th></tr> </thead> <tbody> <tr><td>31494</td><td>1637</td></tr> </tbody> </table>	AREA	DEPTH	31494	1637																	
AREA	DEPTH																							
31494	1637																							

Figure 4.4: Geometric primitives in the vector representation model and some examples of each of them and their associated attributes

Using points, lines and polygons, geographical space can be modeled by associating values to these primitives. A feature can have **multiple primitives**. For instance, in a layer that contains countries, a country such as the United States will require several polygons (continental US, Alaska, Hawaii islands, etc.). All those polygons form a single feature, since all of them belong to the same country and will share the same associated values.

A layer can contain features with primitives of different types, but usually it is restricted to just one single type. It is common to speak of a “points layer” or a “polygons layer” to indicate that.

Elements can be represented using different types of primitives. For instance, a city can be represented as a single point or as a polygon with its perimeter. Using

one or another geometry should depend on the type of phenomenon that we want to model or the level of detail that is needed, among other factors.

The thematic component in the vector model is defined using **attributes**. A layer usually contains multiple attributes. Attributes are associated to features, can have information of all types and they are more versatile than the values associated to raster layers, which, as it was mentioned, normally contain just numerical values. Due to its particular structure (a set of attributes associated to a feature), the thematic component in the vector model can be represented as a table and stored in a **database** (we will talk more about this in the chapter devoted to databases). Also, it can be analyzed independently of the spatial component.

A particular element of the vector representation model is **topology**. A vector layer is said to contain topology if it contains the spatial relations between its features. Topology is required for certain analysis and changes the way some operations, such as geometry editing, work in GIS.

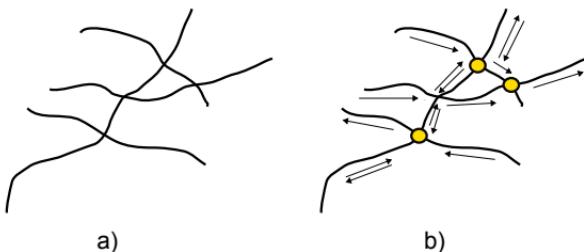


Figure 4.5: A roads layer without topology (a) and with topology (b). Circles in this last case indicate connections between roads.

Although most vector layer operations can be performed without topology, some of them such as **network analysis** are not possible without it. If we think about a roads layer, if it just contains lines representing roads but

no information about how they are connected, there is no way of constructing the network from them. The points where lines intersect might be crossings or roundabouts (so it is possible to move from one road to another), but they might also be points without connection between the roads (one passing above the other). Without knowing that, we are missing information and the network analysis cannot be performed. (Figure 4.5)

Line data without topology is popularly known as *spaghetti* data.

Raster vs vector

Both the raster and vector representation models can be used to store **any geographical information**. Figure 4.6 contains an example of that, and it shows a roads layer represented using both models.

We mentioned DEMs as a typical case of raster layers. Representing elevation as a raster layer has many advantages, especially for performing analysis, but it is not the only option. We can have a vector layer with points (that will be the case if the elevation data comes from a topographic survey), or a lines layer with contour lines (the most common way of representing elevation in a traditional map).

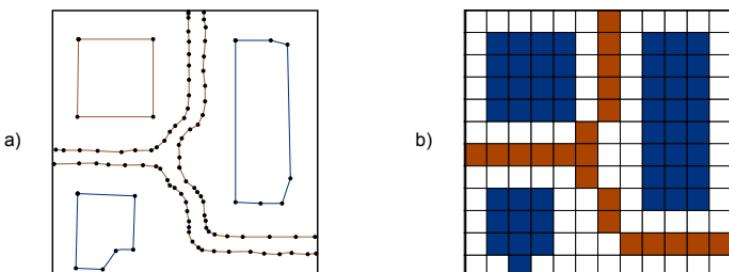


Figure 4.6: Comparison between vector (a) and raster (b) representation models.

It is clear that both models have many differences, and each of them has its pros and cons. The following are some factors to compare.

- **Approach.** Raster model focuses on the properties of the space that is represented (*what* and *how*), while the vector model focuses on the location of that property (*where*).
 - **Accuracy.** Raster model has its precision limited by the cell size. While this can be as small as we want, that would result in very large amounts of data. Features smaller than that size cannot be represented and it is assumed that there is no variability within a cell.
Also, shapes are limited to straight angles, since the base unit for the raster grid, as we have seen, it is a square or rectangle (Figure 4.7)

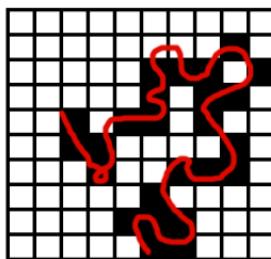


Figure 4.7: Limitations of the raster representation model. Since the space is divided in square units, elements such as curves cannot be faithfully represented.

- **Complexity.** Analysis algorithms, specially those in which several layers are used and combined, are usually simpler and easier to implement with raster layers, mainly due to their regularity and systematicity. Working with vector layers, which do not have any regularity, tends to be more complex from the algorithmic point of view.

Overall, there is no representation model that is better than the other. Depending on the case, one will be more suitable than the other. The following factors should be considered when evaluating the suitability of these models to a particular circumstance:

- **Type of variable or phenomenon to represent.** In general, it is better to use raster layers for **continuous** variables such as elevation, in order to make it easier to perform analysis based on them. **Discrete** variables, on the other hand, are better represented using a vector approach.
- **Layer purpose.** It is important to know how we plan to use a layer, to determine the representation model that might better suit our case. For instance, if we have elevation data and we plan to perform analysis, it will be better to have a raster DEM, since most algorithms require the elevation data to be a raster layer. However, if we want to use that elevation data just for visualization and combine it with other layers to create a map, it might be better to have a vector layer with contour lines, since those will be a better cartographic solution and will interfere less with the remaining variables.
- **Context.** The context might make it better (or even mandatory) to work with a given representation model. For instance, if we are working with images and plan to do some analysis with other layers as well, those should be raster layers, since images, as we have seen, are always raster layers.

There are algorithms that allow **converting between the raster and vector representation models**, so if we have our data in one of them, we can obtain a new layer that uses the other model and might be more suitable for our work.

DATA SOURCES

Not so long ago, all information used in a GIS had its origin in a paper map whose content was later transformed to adapt it to the particular nature of that GIS. Geographical data were obtained from the **digitalization** of printed cartography; that is, from the conversion of analogical maps into digital data that GIS can handle.

Apart from the fact that we can use them in a GIS, digital data have many advantages and represent an important qualitative improvement. Digital data are easier to update, easier to distribute (specially since the Internet was created), use less physical space and are easier to maintain (digital data do not degrade: their physical support does, but they are easy to replicate without losing quality).

Techniques for geographical data acquisition have advanced and it is possible now to create data that can be directly integrated into GIS. Data sources that produce data ready for use in GIS are called **primary** data sources. Those that generate data that has to be adapted or converted are called **secondary** data sources.

In this chapter, we will see the main data sources that provide data for GIS.

REMOTE SENSING

Remote sensing is the **acquisition of information about an object or phenomenon without making physical contact with it**. Instead of measuring the object itself, it measures the perturbations —mainly the electromagnetic ones— that it causes on its surroundings. In our case, it is applied to objects on the Earth's surface.

A remote sensing system contains the following elements (Figure 5.1):

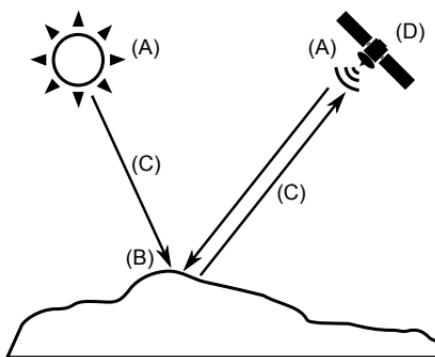


Figure 5.1: Elements in a remote sensing system.

- **A source of radiation (A).** It can be natural or artificial. Radiation emitted by the source reaches the Earth's surface and it is altered by the presence of the objects on that surface. Remote sensing studies that alteration. Objects themselves can emit radiation as well.
- **Objects (B) that interact with radiation** or can emit it, as mentioned above.
- **An atmosphere (C)** through which radiation moves from the source to the objects. The atmosphere also interacts with the radiation and alters it.

- **A receiver (D) which receives the radiation** once it has been emitted or altered by the objects. The receptor measures the intensity of the radiation coming from different points in the area being studied and, with them, generates its final product (in most cases, an image).

In this chapter, we will describe these elements in detail. To study the first two, we will also study some fundamentals about radiation and its interaction with matter. For describing the receivers that make part of a remote sensing system, we will separate them into two components: **sensors** and **platforms**.

Interaction with the atmosphere must be managed in order to eliminate its influence since, in most cases, we are interested in the object on the Earth's surface, not the atmosphere itself. Removing that influence is part of the post-processing of the data. Those processes are, however, rather complex, and they go beyond the scope of this book, so they will not be explained here.

Electromagnetic radiation

Electromagnetic radiation is caused by alterations in the electric and magnetic fields, which generate waves corresponding to each one of them. These waves move at the speed of light and can be described with the usual parameters such as wavelength and frequency. The range of frequencies (and corresponding wavelengths) of electromagnetic radiation is called the **electromagnetic spectrum**.

The spectrum is subdivided in regions depending on the wavelength, such as (from shorter to larger wavelength) gamma rays, X rays, ultraviolet region, visible region, infrared region or microwaves.

Radiation emitted by a radiation source is altered by the presence of objects that **absorb**, **transmit** or **reflect** it.

These three phenomena take place in a different proportion depending on the characteristics of the object and the radiation. For remote sensing, **the radiation that is reflected is the one of interest**, since it can be collected later and used to produce the data output.

Since each object reflects radiation at different wavelengths in a different way, this can be considered as a property of an object. The particular response of a given object and the way it alters a given radiation (which depends on its shape, material, etc.) is known as its **spectral signature** and can be used to identify the object.

Sensors and platforms

The two main technological elements in a remote sensing system, both of them related to the receiver, are the **sensor** and the **platform**.

The sensor is the element that can read the electromagnetic radiation and register its intensity for a given zone of the spectrum. It can be a simple photographic camera or a more specialized sensor.

Passive sensors use natural source of radiation (in most cases, sunlight), and just measure that radiation as it is reflected on the Earth's surface. **Active** sensors emit their own radiation, and then collect it back after it has been reflected. Here is a simple example to better understand this: a photographic camera is a passive sensor, while a photographic camera that uses a flash unit is an active one. The radiation emitted by an active sensor does not have to be visible light (as in the case of the flash); the sensor can emit in other regions of the spectrum.

Technologies such as **radar** or **LiDAR** (similar to radar but with light pulses instead of radio waves) are based on active sensors.

The sensor is **mounted on a platform**, and it performs its data acquisition from it. **Several sensors** can be mounted on a single platform.

The two main types of platforms are those located inside the Earth's atmosphere (mostly on airplanes) and those outside of it (on satellites).

The advantage of airplanes is their **availability**, since they can be piloted and used to cover any place on Earth at any moment. Satellites, on the other hand, cannot be guided, and its movement is fixed and defined by a set of parameters known as **orbital parameters**, which define the orbit described by the satellite around the Earth.

Orbits can be classified according to their **rotation axis** or their **movements**. Two particular cases are the **geosynchronous** orbits (the satellite is located at a fixed point and its movement follows the Earth's rotation) and the **heliosynchronous** (the satellite passes over any given point in its path, always at the same local solar time).

Resolutions

The most important parameters that define the characteristics of a remote sensing system are its **resolutions**. These define the level of detail of the products that it creates. Resolutions depend on both the sensor and the platform as a single operative unit, and on the individual characteristics of each of them. Four resolutions can be defined:

- **Spatial resolution.** It indicates the size of the smaller object that can be distinguished. If the output is an image, the spatial resolution is the real size of the area represented by a single pixel.

- **Spectral resolution** indicates the amplitude of each of the regions of the spectrum that is registered. It is defined by the total amplitude that is covered and the number of sections into which it is divided. The measurement corresponding to each of these sections will usually be stored in a separate band in the resulting image.
- **Radiometric resolution** indicates the level of detail of the intensity measurement taken for each of the spectral regions that are registered.
- **Temporal resolution** indicates the time that it takes the sensor to return to a given place. It makes sense only for orbital sensors and depends on the platform characteristics, such as altitude, and also on the sensor characteristics.

It is not possible (for technical and theoretical reasons) to have a sensor in which all the above resolutions are maximized simultaneously. Some sensors might favor certain resolutions, while others might favor different ones.

When using images coming from remote sensing in a GIS, we should consider which resolution is more important (for instance, to locate elements that have a small size, high spatial resolution is needed). Using data from several different sensors is a good strategy for overcoming these limitations.

Photogrammetry

Photogrammetry is the technique used to study and precisely define the shape, size and position in space of any object, using measurements from photographs. Of special interest to GIS is the branch of photogrammetry known as **aerial photogrammetry**, which uses aerial photographs and it is mainly used for generating elevation data through a process known as **restitution**.

Instead of single images, the branch of photogrammetry known as **stereophotogrammetry** uses pairs of images, each of them taken from a different point. These images form a **stereo pair** and with them a three-dimensional reconstruction of the original scene can be produced. This can be used by an operator to see the scene with **depth and volume** so that terrain forms can be identified and elevation information obtained.

If using satellite images, stereo pairs can be obtained from those platforms and sensors that allow **changing the angle of vision**, so in the same satellite pass, pictures of a given area can be taken from different points.

Photogrammetry can be **analogical** or **digital**, the latter being the one more related to the field of GIS.

Stereoplotters are used to combine and align the images that form the stereo pair. Current stereoplotters are called *analytical stereoplotters*, and contain elements from GIS, along with more specific elements. Among these, we find specific visualization software and peripherals such as 3D mouses or other mechanical elements found in analogical photogrammetric devices, making it easy for operators to adapt to this new type of tools.

PRINTED CARTOGRAPHY. DIGITIZATION

A large amount of cartography exists in printed form, such as maps or old analogical aerial photographs. To be used in a GIS, this cartography has to be **digitized**, which means creating raster or vector layers from them. In this latter case it also implies **separating the different types of information** that the map might contain, since the information in a single printed map would be stored in independent layers, in GIS

Digitizing a printed cartographic document involves three steps:

- **Georeferencing** the original document. That is, setting a geographical context (coordinate system, control points, etc.), so the digitized elements produced are correctly referenced.
- **Digitizing the spatial component.** That is, creating the corresponding geometries.
- **Digitizing the thematic component.** Creating cell values for raster layers or attributes in the case of vector layers.

Digitization can be **manual** or **automatic**. If manual, an operator introduces the value, while an automatic process is done through an algorithm.

To create raster layers, the most common method is **scanning** the original document using a **scanner** which creates a digital image from an analogical one.

High-end scanners specifically designed for working with cartographic documents are available. Generic scanners, however, can be used for this task with acceptable result in terms of accuracy and distortion.

Two parameters define the characteristics of a scanner: its **spatial resolution** and its **radiometric resolution**. The first one is usually measured in **dots per inch** (DPI) and indicates the number of points (cells) that the sensor will create in the resulting image for each length unit in the original document. Radiometric resolution defines the ability of the sensor to separate between two different colors.

The ideas discussed in chapter 2 about scale should be taken into account here as well. Working with a higher resolution (if the scanner allows it) will not always mean adding more information to the resulting image, since it

might not exist in the original document. We would just have a volume of data larger than the one needed to capture all the information in the printed document but not more information.

In the case of vector layers, **manual digitization** is the most common method. An operator defines the features, tracing its geometries and entering the associated attribute data.

To digitize geometries, the operator can use the **editing functionality of a GIS** and work on the screen of the computer using its mouse as tracing device or use specialized peripherals such as a **digitizing tablet**. In the first case, digitizing takes place on the screen, so a digital version of the printed document is needed (although not a vector one), which can be obtained by scanning it. The full digitization process involves two steps which include two different types of digitization: from printed map to raster image (automatic), and from image to vector features (manual). If using a tablet, the printed map can be used directly to trace geometries on it.

Automatic digitization of geometries in a vector layer is known as **vectorization**. A digital image is needed, so the original document has to be scanned first. The vectorization algorithm analyzes the map and finds the elements that it contains, creating the corresponding vector layer elements from that. Manual work is usually needed to complete and correct the resulting data, since this tends to be a complex and error-prone process in which data preparation has a great importance, and a fully automatic alternative is not possible most of the time.

A particular case of digitization is the **creation of layers from a set of values representing some spatial process**. That is, when the original analogical document is not a map, but just a set of values. This process is known as

geocoding, and it involves assigning coordinates to those values, and then creating the corresponding layers with the combination of the original thematic data and the geographical information (the spatial component) that resulted from the geocoding process.

The original alphanumeric data can be introduced manually or by using an automated approach, such as scanning the document and then using some character recognition (OCR) software.

A particular and very popular case of geocoding (although in this case the document is not analogical) is **geotagging**, in which coordinates are assigned to digital images.

Digitization Quality

One of the most important aspects of digitization is the **quality of its result**, which should be as close as possible to the quality of the document being digitized. Digitization is never perfect, regardless of the accuracy of the equipment that is used or the skills of the person that performs it. There will always be errors and deficiencies.

Along with the errors introduced in the different phases of the digitization process, the original source documents might contain their own errors as well. For instance, scanning a map might introduce errors due to geometric distortions, but that map might itself contain distortion due to its previous use.

Information contained in a cartographic document might include elements that are problematic and will decrease the quality of the resulting data. A map that contains stains or has some lines that are not correctly visible will result in errors when digitizing its vector features (specially in the

case of automatic digitization) regardless of the quality of the scanning process that is required before.

Among the errors due to the digitization process itself, and not related to the document characteristics, **mismatching nodes in vector geometries** are one the most common (Figure 5.2)

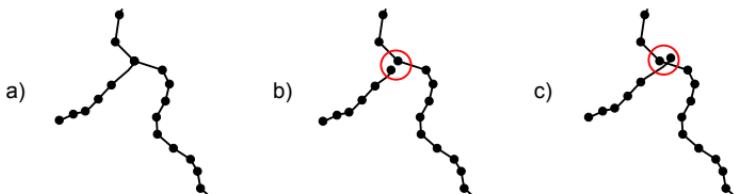


Figure 5.2: Digitization errors. a) Correct version with matching nodes. b) and c) Incorrect versions, with mismatching nodes, causing wrong disconnected lines.

For this reason, the editing capabilities of GIS include additional functionalities to avoid these errors while digitizing, helping the user and allowing an accuracy and quality that would not be possible without such aids. Among them, the automatic adjustment of geometry nodes based on predefined tolerances (known as **snapping**) is specially relevant, as it can guarantee a correct relation between the nodes of different geometries, whether or not they belong to the same feature.

This is particularly important in the case of digitizing not only the geometries, but also the **topological information** that is contained in the original document (such as the connection between the roads in a map sheet). Also, when digitizing geometries and their topology, certain additional rules must be followed, such as only digitizing the shared sides of adjacent polygons once. Additional information must also be digitized, such as the definition of nodes when lines cross each other and there is a relation

between the objects they represent (for instance, a crossing between two road lines that represents a point where vehicles can pass from one road to the other).

GPS

One of the most relevant advancements in geographical data sources have been **global navigation satellite systems** (GNSS). **For any given point and at any time**, these systems allow us **to know the exact location of that point** with an accuracy of a few meters or less. To do that, they use a constellation of satellites to which information is transmitted from the study point, and use that transmission to compute the coordinates of the point.

The first and most popular of these systems is the **Global Positioning System** (GPS). It has 24 active satellites (the satellite segment) along with terrestrial stations to control them (the control segment) and it is based on **trilateration**. Distances are measured from a GPS unit (the user segment) to a certain number of satellites. Knowing those distances and the exact position of the satellites, the position of the unit can be computed. Position is computed with its x , y and z coordinates. The GPS system uses WGS-84 as its reference ellipsoid.

The satellite network is designed to guarantee that, from any point of the Earth's surface, and at any time, a GPS unit can locate the required number of satellites to compute its position.

There are **several error sources** that might affect the accuracy of the position computed by a GPS unit. Among them, we find errors in the position of satellites, errors due to the effect of the Earth's atmosphere on the GPS signal, and also errors caused by accuracy problems with the clocks used to measure signal travel time (which is then

converted into travel distance). **Selective availability** was a random error introduced in the GPS signal with military purposes. It was, however, removed on May 2, 2000.

Among the techniques used to correct or minimize these errors, **differential GPS** is the most important one. It was originally conceived to remove the effect of selective availability but can be used to correct a large part of the other errors that affect the GPS system.

To apply differential GPS, along with the receiver unit for which its position is to be computed, a **second receiver** is needed. It has to be **fixed**, not mobile, and its coordinates have to be known with great precision. This receiver is itself a high-precision unit, and broadcasts information that other units can use to correct their position.

The idea is that errors that affect the mobile GPS unit **also affect the fixed reference one**. The error for the reference unit can be computed, since its position is known, and using the discrepancy between that real position and the one computed using the GPS system, the position of other GPS units can be corrected.

Using this differential correction, a regular GPS can obtain coordinates with an accuracy of around 2 meters in their x and y components, and around 3 meters in the z component. Without differential GPS, an accuracy of just 10–20 meters is to be expected.

Precision of the GPS system depends on the GPS unit. There are many classes of GPS receivers, but the main ones, from the point of view of GIS, are two:

- **GPS for general use.** These units are low-cost **small and portable** for outdoors activities, where a high accuracy is not required. The GPS receivers found in smartphones fall into this category.
- **GPS for surveying.** Larger units, usually with independent antennas that are connected to the receiver to

increase accuracy. It also ensures better location of satellites in difficult conditions, such as under forest canopy. They are designed for professional use.

Regarding its connection with GIS, GPS units can collect coordinates and then create GIS layers with them. They can store points (called **waypoints** in the usual GPS terminology), or capture the path followed by the user (called a **track**). In this last case, the receiver stores points at regular intervals so the user can just move and does not have to manually store the coordinates along the route. This information can be later introduced in a GIS for further analysis or visualization.

VOLUNTARY GEOGRAPHICAL INFORMATION

The participative ideas of the so-called **Web 2.0**, when combined with tools such as recreational GPS units, or with simplified software for editing and digitizing, result in interesting initiatives in which people, with no specific training in cartography or surveying, can acquire and share geographical information.

Although this cannot be considered a different data source (the techniques and devices used have already been described in previous sections), there is an important change in the philosophy behind data collection and usage, which makes it worthwhile to treat this type of data separately in the context of this book.

The term **Volunteered Geographical Information** (VGI) refers to the use of the Internet to create, manage, and share geographical information which has been voluntarily contributed by a community of users, also using the Internet. The set of techniques and tools used by those users is termed **neogeography**.

Neogeography has changed some fundamental ideas in cartography, since it has modified the traditional concept of geographical information (which was created by a very skilled few), its characteristics, or the role it played in certain fields. The following are some ideas about these changes and neogeography itself.

- **Popularization and democratization.** Cartographic production has always been in the hands of governments or agencies and, in many cases, has been strongly censored due to its high strategical value. With the advent of VGI, geographical information becomes more democratic, and its creation is a free, participative and unrestricted process. The top-down approach that had dominated the production and use of cartography is thus inverted.
- Citizens become **sensors** and are more conscious of their geospatial reality.
- Cartographic production loses its mysticism.

The most relevant VGI project at this time is **OpenStreetMap** (OSM), a “collaborative project to create a free editable map of the world.”.

METADATA

Regardless of their origin, data might need additional data to be interpreted. For instance, if we have the coordinates of a point, to correctly interpret it we need, among other things, the coordinate system in which those coordinates are expressed. The data we work with (the coordinates) should be accompanied by some ancillary data (such as the EPSG code of the coordinate system).

This ancillary data are known as **metadata**. Metadata are **data about the data**, and their purpose is to **explain the meaning of the data**. That is, they help the user

to better understand the meaning of the data and the information that they contain. Metadata are an additional document that accompanies the data and that allows for better management and use.

Within GIS, metadata are usually **associated to a layer and its content** and can be **referred to both components** (spatial and thematic).

The concept of metadata is not exclusive of digital geographical data. A printed map has metadata in a certain way. A legend or a text in its margin with information about the date in which it was created, are also metadata. In the case of digital metadata, the metadata are **independent from the data itself**. That allows us to perform operations **separately on the metadata**, which opens many possibilities and gives them a greater value.

Two of the main functions of metadata are **ensuring the correct use of data** and **facilitating its management, discovery and exploitation**.

Geographical data, as with many other types of data, are usually created for a given purpose, and that purpose does not have to be contained in the data themselves or be easy to infer from them. When data are then used for a different purpose, problems might arise, since the data might have some deficiencies when used in this new context. With the help of metadata, this can be solved. Metadata may help, for instance, **avoid the use of outdated or inaccurate datasets**. Knowing the parameters that define a dataset (original scale, date, etc.), we can better judge whether or not to use them for a given task.

Data creators must **add enough metadata to them** and users must **consult associated metadata** before using any data.

Regarding data management, metadata facilitate operations such as searching data in larger datasets or data

collections. Metadata can contain a summary of the full data they accompany, and that can be used to **make search operations more efficient** if they use geographical criteria. For instance, if a user wants to find whether a data collection has some data for a given area, the full content of each vector layer in the collection should be checked, to determine if any of its geometries falls within the requested region. This can be a lengthy operation. If each layer has an associated metadata with its extent, the query can be executed with the metadata, which is much faster, as it is independent of the number of features in the layer. This is fundamental when creating **data catalogs**, which respond to user queries based on the information contained in the metadata.

Content of metadata. Metadata creation

The information contained in the metadata associated to geographical data depends on parameters such as the **representation model** used, the **format** in which data are stored (file format, database, etc.), the **organization, entity or individual responsible** for the data or the **element(s) with which they are associated** (set of layers, layer, feature, etc.).

Some of the common elements that are added to metadata in the case of geographical data are **identification** information (to identify them in a unique way and distinguish them from other data), information about their quality (including their origin, and the origin of the data they might derive from –the **data lineage**–) or information about its distribution (access, license, etc.).

Metadata can be created **at the data origin**, at the same time as the data themselves are created. It can also be **extended later** by data distributors, managers or users.

Most of the metadata content is created manually, using specific applications, sometimes connected to desktop GIS tools. Metadata elements that can be extracted from the data itself (such as the data extent), can be created automatically.

Metadata are generally stored as **additional files** that accompany the data files or they are stored **as part of a database**.

SOFTWARE AND TECHNOLOGY

The classic concept of GIS is that of a complete software application which implements all the tools needed for working with geographical data: creating or editing, managing, analyzing and visualizing. Along with that, other types of applications have appeared which, although they do not match exactly that definition, have to be considered as part of the GIS world.

We will divide GIS applications into three main blocks: **desktop GIS**, **web-based GIS**, and **mobile GIS**. They will be described in detail in this chapter. We will also provide additional information about some technologies that they are based on.

DESKTOP GIS

There are five fundamental functionalities of a desktop GIS: **data input and output**, **visualization**, **editing**, **analysis**, and **map design**. Most desktop GIS tools have these five capabilities, although the level of functionality for each of them might differ. Some tools might be more prepared for data editing, while others might focus on analysis.

Data input and output

A desktop GIS must be able to **read data** and, optionally, to **save** it. This last functionality is needed in case the GIS can produce new layers, but not in those that do not contain analysis or editing capabilities.

There are a **very large number of data formats for geographical data**, and most GIS use common libraries to be able to read and write them, allowing them to share data among them and improve their connectivity.

Apart from being able to read data files, it is now also important to be able to connect to *databases* and **remote services**. We will talk about those later in this chapter.

Visualization

Visualization is a fundamental capability of GIS. It is, of course, important when the main purpose of using GIS is to create cartography, but also when our work is focused on data editing or analysis, since visual exploration of the data is a previous step.

The visualization part of GIS is mainly comprised of a **canvas** on which layers are rendered. The user can add or remove layers, and also change their **symbology**, that is, the way in which the layer data is converted into graphical elements. Layers are rendered in a given order which allows to create a *rendering hierarchy*.

Along with the canvas, there are **navigation tools** that allow the user to modify the area that is being displayed by zooming in, zooming out or panning (Figure 6.1).

The most remarkable feature of geographical data visualization in GIS is that, unlike what happens with a classic printed map where its characteristics cannot be changed, the user can select *what* he/she sees and *how* he/she sees

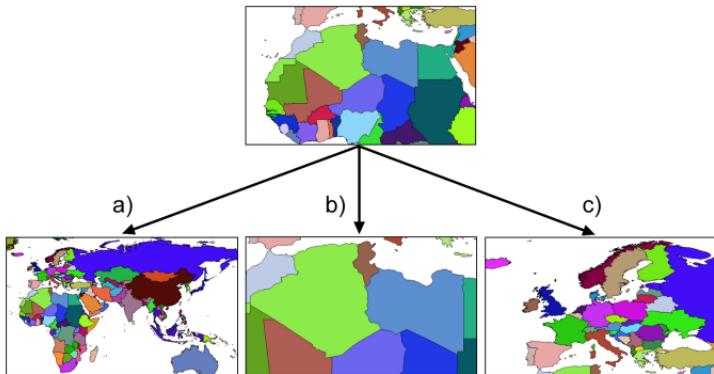


Figure 6.1: Navigation tools that are common in a desktop GIS.
a) zoom out, b) zoom in, and c) pan

it. Geographical data is **independent of the information needed to visualize it** and, therefore, it can be represented in different ways. This is true even for data that has an inherent visual nature, such as images, since even in that case the rendering can be adjusted and modified by the user.

Although in the most common case the canvas is bidimensional, certain GIS are also capable of **three-dimensional rendering**. In this case, navigation tools are more complex and they allow for adjusting perspective, vision angles or vertical exaggeration, among other parameters.

Analysis

Analysis is a fundamental functionality of GIS since its origins. Others, such as visualization, although we cannot imagine GIS without them nowadays, were very limited in the early days. Analysis, however, has always been at the core of GIS.

The current trend in GIS is to consider analysis capabilities as **modular tools** that are run on a base platform which includes the data input and output capabilities, along with the visualization component. Analysis tools are independent, but they can be used together to create more complex analyses.

Analysis tools might be completely independent of the visualization component or be linked to it. In the first case, the analysis is performed on a set of layers and parameters without any interaction with the map while in the second case, the user might interact with the view to define how the analysis is performed (for instance, selecting a coordinate or a region in the canvas which will then be used as a parameter for the analysis tool).

The result of an analysis tool in GIS can be geographical (a new layer) or not (a simple value, such as the one resulting from some statistical analysis of the input data).

Analysis tools can be organized into **workflows** which help **automate** analysis routines. Also, the analysis functionality of desktop GIS can usually be used from *scripting languages*, which allow definition of more complex models and data flows. This is one of the main strengths of current GIS tools, since it provides the user more power and flexibility.

Editing

The geographical data with which we work in GIS are not static. Information contained in a layer **might have to be changed or corrected**, and the functionality that allows the user to do that is important if we want the GIS tool to be versatile. Without them, geographical data lose part of their potential, and that is the reason why most desktop GIS tools implement editing capabilities to some extent.

This capabilities can be used to **create new layers** or to **update existing ones**. The following are some editing tasks that can be performed with GIS:

- Editing the geometries of a vector layer feature.
- Editing the attribute values of a vector layer feature, including editing the list of attributes of the layer, adding or removing them.
- Adding new features to a vector layer or removing existing ones.
- Editing cell values in a raster layer.

Tools used to edit geometries inherit a large part of their design from CAD software. In certain cases, they are extended with new functionalities, as happens in the case of editing geographical data with topology (CAD software does not consider topology).

Map design

Most desktop GIS are capable of producing cartographic documents which can later be printed and used as a classic paper map. These documents are composed in the GIS from the data, and use the same functionality that it is used for the on-screen rendering (symbology, etc.).

Along with that, other tools allow the user to **design and compose the map**, and to *adjust its elements* (rendered layers, legend, title, etc.), and are inspired by those found in design software.

Some desktop GIS include elements to *automate cartographic production*, such as templates or tools to generate map series (Figure 6.2).

This is possible thanks to the **separation between geographical data and the design of the cartographic document**, similar to what we noted for the case of visualization.

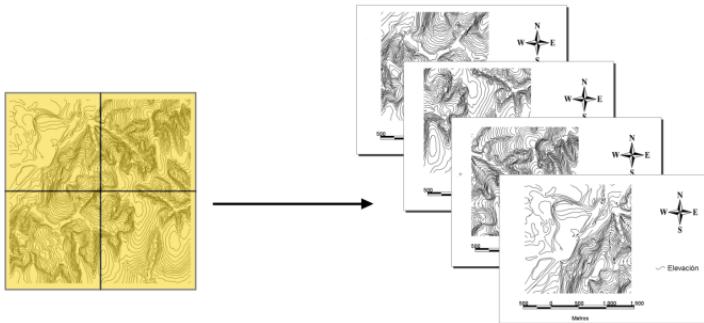


Figure 6.2: Automated creation of map series in a GIS.

WEB MAPPING. CLIENTS AND SERVERS

One of the most relevant advances in the history of GIS is the advent of **Web mapping**. Web mapping technologies are used to incorporate GIS elements as part of websites, with internet browsers being then the base platform on which GIS functionality is executed. These technologies include not just the elements run on the browser, and have been key in shaping and developing others such as **remote data services**, which are used not only by Web Mapping applications but also by desktop GIS.

The concepts of *server* and *client* are fundamental in this context. Let's discuss them in a bit more detail.

A **server** is the element that provides (serves) a given content through the network. In our GIS context, this content means basically geographical data. The **client** is the element that requests the data, receives it and works with it.

A web browser is a client, since it makes a request to get the content of a website and shows it to the user. When we enter a web address in the address bar of a web browser, we provide the information needed to establish

the connection between the server and the client and to transfer the data from one to the other.

Let's see how that works. Suppose that we want to visit the following website:

`http://victorolaya.com/writing`

The requests is done based on the web address —more technically, a Uniform Resource Locator (URL)—, which is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. We can divide it in the following parts:

- `http`: The protocol to use, which defines the way client and server will communicate with each other.
- `victorolaya.com`: The host name. This part identifies the server machine connected to the network where the page that we want to visit is stored. It is a human-readable version of a numeric code that indicates the address.
- `writing`: The page we want among all the ones that the server can provide.

The process that allows us to have that page in our web browser comprises the following steps:

1. The client makes the request.
2. The server machine is identified and the request is driven to it.
3. The server prepares the page that has been requested and sends it back to the client (or it sends an error message in case it could not find or prepare the page).
4. The client receives the page and renders it so the client can see it.

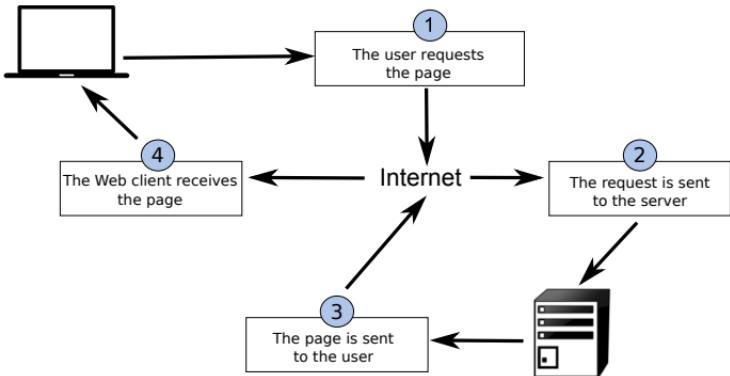


Figure 6.3: Client-server interaction mechanism that takes place when visiting a website.

Figure 6.3 shows a summary of this.

A relation is established between clients and servers, in which an arbitrary number of clients connect themselves to a server, from which they obtain data whenever each of them makes a request. In this client-server architecture, the server has the data to be shared through a service, while clients only provide information about themselves needed to validate and perform the data sharing.

Now let's see how these ideas apply to the field of GIS. Regarding servers, they can have four main capabilities:

- **Serve rendered geographical data.** Generally known as **map servers**, they provide maps. That is, images created from geographical data. If data is already an image (such as an aerial or satellite photograph), the server will just send it as it is. If data is not an image (vector layer or raster layer other than image), the server will create an image based on the geographical data. The symbology used to do this can be a default one that the

server uses for all requests or it can be provided by the client in the request.

In both cases, the client also specifies the dimensions of the requested map image that is served, and then the server prepares it.

- **Serve the data directly.** A more flexible option is to serve the geographical data itself. The client requests the data and once it has been transmitted across the network, can use it however he/she needs. In case the data is to be visualized, the symbology has to be set in the client side, since the server is not taking care of that and provides the raw data.
- **Serve the result of queries.** Another functionality that the server can have is to return not just the full set of geographical data, but a subset of it. The client can specify a **filter** and the server will use it to create a subset that will later be sent in the response. Also, the server can provide **descriptive values** about the data it has. The client, which might be connected to several services and obtain the values, can use those values to filter which services to use (for instance, asking them the extent of their data and then selecting only those that have data about a given study area). As we have already seen, **metadata** have a great relevance in this context, since they allow this kind of queries to be executed (and the corresponding requests to be responded to) efficiently.
- **Serve processes.** Finally, a server can provide new data, whether geographical or not, computed from geographical data. In this case, the server provides a processing service, and it processes the data that is passed to it as part of the request. The request can contain the data itself, or a reference to it. If a reference is passed, the data might already be in the server, or it can be in another one. In this last case, the first server will become a client

of the second one, will retrieve its data, process it, and send back the result to the original client (Figure 6.4).

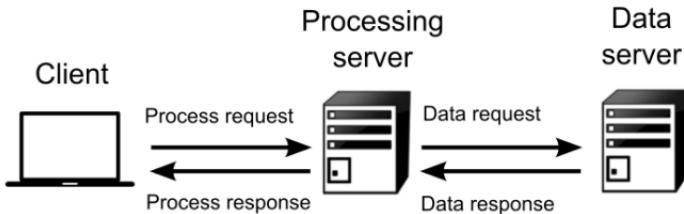


Figure 6.4: Remote processing service using data from a second server.

About clients, they can be divided in two classes:

- **Heavy clients.** Heavy clients are independent applications that do not run on top of another one such as a web browser. They usually have a larger size since the application has to care of all the program logic.
Heavy clients handle and use data not coming from web services, such as local data files. They are not just clients, but full-fledged applications that work even without its client part.
Nowadays, most desktop GIS are heavy clients themselves as they have the functionality of classic GIS but can also consume web services.
- **Light clients.** They normally have a smaller size and their capabilities are more limited. They run on web browsers and most of the time rely on remote data from servers.

Although originally, they focused on data visualization (adding map views to websites, with a certain degree of interactivity), they have begun to implement more advanced functionalities such as analysis functionality

(whether on the client side or using a processing service) or data editing.

The term *Web mapping* is used to refer to the lighter clients which focus only on rendering maps, while the term *Web GIS* is used for those with more functionality, incorporating some of the tools traditionally found in desktop GIS.

SOME TECHNIQUES RELATED TO GIS SERVICES

Two important techniques used in the context of the client-server architecture for geographical data are **tiling** and **caching**. These techniques, whether implemented on a light client or a heavy one, allow for more responsive interfaces and reducing the amount of data sent over the network, overcoming to a certain extent the problems that a slow network might cause. Both are used mainly with map servers (servers that provide rendered images).

Tiling divides the images that the client is working with into smaller ones, forming a mosaic. By correctly managing the tiles in that mosaic, the amount of data transmitted can be reduced. When the request is sent to the server, instead of a single image, a set of them is requested. Although this does not reduce the amount of data, the tiled structure will allow a more flexible and optimized handling of data once a new image is needed, as will soon be explained.

Caching is a technique frequently used not just for web SIG, but as a general tool in the context of the internet. Web browsers store previous responses from web servers, such as web pages and images, in a so-called *cache*. When data that was previously requested is requested again, it can be taken from the cache instead of from the corresponding server, which is usually faster and more efficient.

Combining tiling and caching increasing responsiveness and results in an optimized data management. Let's see how that works, using the example shown in figure 6.5.

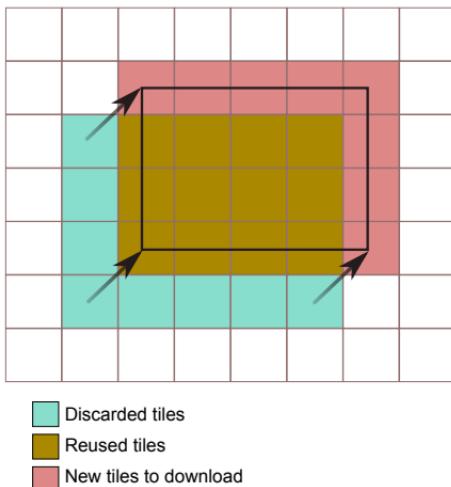


Figure 6.5: Combination of tiling and caching techniques to optimize data handling in a web GIS application.

Initially, the application displays an area that cover 20 elements or tiles. All those tiles have already been downloaded from the server, and stored in the application cache, which means that using them again does not require making a new request to the server. Without caching and tiling, when the application user changes the area to be displayed as shown in the figure, a whole image has to be requested, with exactly the same size as the image to be rendered in the screen. However, if tiling and caching are applied, we just have the whole new image to be painted divided in 20 parts, and, since some of them are stored from a previous request, we just have to request a small number of them (8 in this case, corresponding to the areas not covered by

the previous image). The amount of data that is requested to the server and transmitted over the network is much smaller.

Caching can also be implemented on the server side. We have seen that map servers provide already rendered images based on some data. Rendering that data can be time consuming and, if it has to be done for each request, that would mean a lot of computing cost for the server. Instead, images are pre-rendered at different scales so when a client request is received by the server, it just has to crop the pre-rendered image instead of producing the response image from the base data.

A recent technique that is gaining popularity is *vector tiling*. Using the same approach as in the case of the tiling we have just seen (that is, cutting the data in pieces), vector layers are divided and only the required data are sent to the client.

This allows the client to request and use vector layers and be responsive at the same time. Without vector tiling, this would be impossible for large layers. The advantage in this case is that the symbology can be defined by the client. Also, the user experience is improved, since for instance, transitions become more fluent when changing the map scale, due to the scalability of vector data.

STANDARDS

To ensure the the client-server system works correctly, it is important to define how the communication between servers and clients takes place. Some **normalization** is needed, and there must be common and well-defined elements implemented by both the client and the server. This *lingua franca* that allows clients and servers to communicate is what we call a **standard**.

In an ideal situation, a complete **interoperability** would exist independent of formats and applications used. Clients and servers would be able to connect with each other, regardless of their own characteristics. Standards are the element that allows that to happen, because they define a common framework in which clients and servers communicate. As long as a client or a server follows the standard, it will be able to communicate with all others that do it as well. Standards provide **technological homogeneity**.

Interoperability means that any element of the client-server system can be replaced with another one, and the interaction between all parts of the system will not be affected. A client or server might have different functionalities, but regardless of its origin (its manufacturer), it will be able to interact with the other elements, if all implement the same standard.

A standard is considered as such when it is used by a group or community, which accepts it to define the characteristics of a product or service within it. Standards can be established by public acceptance and custom (*de facto* standards) or they might have legal recognition and be proposed by some official organization (*de iure* standards).

A standard is **open** if **its definition is available** to everyone who wants to know more about it and use it for any activity related to it.

The following are some of the fundamental principles that open standards are based on:

- **Availability.** Open standards are available to anyone, to read and to use.
- **Maximize end-user choice.** Open standards create a fair, competitive market, and do not lock users in the closed environment of a given vendor.
- **No royalty.** Implementing a standard is free and has no cost, unlike the case of a patent.

- **No discrimination.** Open standards and the organizations behind them do not favor any implementer of the standard over the rest of them.
- **Extension or creation of subsets.** Standards can be extended with additional elements or reduced to less-detailed subsets.

To know the impact that a standard has in the context of GIS, let's take a look at figure 6.6, which represents a non-interoperable architecture that does not use standards.

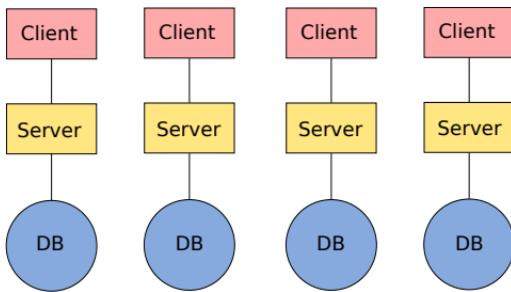


Figure 6.6: Non-interoperable architecture.

Data stored in each database are available only by using one client, the one corresponding to the server that serves those data. The remaining data are not available for that client. Each client-server-database group is an independent island, technologically isolated from the rest of them.

Disadvantages of a non-interoperable architecture like that include the following:

- **Waste of resources.** Each service must manage its own data. That is complex and has higher costs than sharing data with other compatible services.
- **Need to know multiple clients.** Since we need a different client for each service, the user must be familiar with all of them. Being capable of using just one client is

not enough to use all the available data, since that client can only access a small part of all that data..

- **Combining data is not possible.** Two datasets that are available through two different services cannot be used in the same client, as it cannot communicate with the corresponding servers.
- **Combining functionalities is not possible.** If data is only available to a given client, the functionality in another one (which might not be implemented in the first) cannot be used on that data. When working with that data, the user's possibilities are limited to what the corresponding client can do.

Now let's take a look at a fully interoperable architecture based on open standards, as seen in figure 6.7.

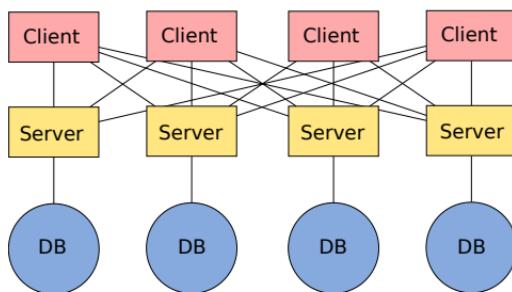


Figure 6.7: Interoperable architecture.

In this case, there is a server that manages and offers the services for each database, but all clients can access all servers, since they are based on open standards and communication is possible between any two of them.

Relevant standards in GIS

The most common standards for geographical information are created and promoted by the **Open Geospatial**

Consortium (OGC). The OGC is “an international not for profit organization committed to making quality open standards for the global geospatial community”.

Some of the most relevant OGC standards are the following ones:

- **WMS.** To serve maps (images)
- **WCS.** To serve coverages (currently only raster layers).
- **WFS.** To serve geographical features and attributes (vector layers). It can also allow editing features from the client.
- **WPS.** To serve remote processing services.
- **GML.** To store geographical information.
- **CSW.** To make queries to a catalog that contains geographical data.

Each one of these standards is described in the corresponding specification, which is subject to change and improvement. Several versions exist for each of them.

Along with these standards are those made by organizations such as **ISO** o **W3C**, with a more general scope, but also important in the context of GIS. Among them, the most relevant standards are the ISO ones that define **how to store metadata** and the W3C standards related to **communication over the Internet**.

MOBILE GIS

GIS on mobile platforms such as mobile phones or tablets has a clear relation with both desktop GIS and web GIS. It takes the elements from them and adds others derived from running on a mobile platform which expand their possibilities.

Mobile devices nowadays offer two main capabilities from the point of view of GIS: **wireless access to Internet** and **ability to know the position** of the device.

Internet access can be used to obtain maps and geographical information from a given server, or to send field data acquired with the aid of the mobile device.

The position of the mobile device is usually known based on the GPS receiver that is part of most mobile phones. However, other approaches are also possible, from computing the position **based on the phone network** to using some **indoor positioning system** if it is available.

Knowing the position of the device allows the mobile GIS application to provide additional functionality. For instance, it might be used to make **field data acquisition** easier and more efficient (the coordinates of measured points do not have to be entered manually), or to provide **location-based services** (LBS).

Some of the main groups in which these services can be grouped are listed next.

- **Navigation.** Shortest path computation, route guidance, etc.
- **Data acquisition.** Any type of data can be registered in the field, and the device associates to them its own position automatically.
- **Information.** Business directories, travel guides, etc.
- **Advertising.** Location-based advertisements, promotions for nearby shops, etc.
- **Tracking.** Of both people and products, along predefined routes or arbitrary ones.
- **Management.** Of infrastructures, installations or fleets.

When running in a mobile platform, a GIS has additional information about the context it is running on (position, direction of movement, speed, illumination, etc.), that

it can use to provide more functionality than a desktop or web GIS running on a non-mobile platform.

DATABASES

Databases are used in all disciplines in which an efficient data handling is needed, especially if those data are large. The data that are used in GIS are usually rather large, and the improvements in data acquisition have caused geographical data to be now more precise and consequently, larger.

Databases not only have the advantage of being able to work with large datasets, but also other advantages such as managing multiple users or providing efficient access and indexing. For this reasons, they are a fundamental element in any software context, including GIS.

A **database** is a **systematically stored and organized collection of data**. Databases provide a better way of handling and using data, thanks to their structure.

Some of the advantages of using a database instead of a traditional file-based approach for storing data are:

- **More independence.** Data are independent from both the users and the software.
- **More availability.** Databases facilitate the access to data from different contexts and applications, making them more useful for a larger number of users.

- **More security (data protection).** Replication and synchronization of data becomes easier.
- **Less redundancy.** There is a smaller volume of data and faster access.
- **More efficiency in data capture, encoding and input.**

This has a direct influence on the results that are obtained from the exploitation of database data and we find the following advantages:

- **More coherence.** Better management leads to better data which produces results with a higher quality.
- **More efficiency.** Accessing the data is easier and more effective.
- **More informative value.** It is easier to extract the information that is contained in the data, since one of the goals of a database is to increase the value of data as source of information.

Users also enjoy advantages when using a database, such as the following:

- **Easier access.** The user of the database just has to worry about *using* the data. A solid infrastructure to do it is available and so are the tools needed for it.
- **Easier data reutilization.** Data is easier to share when using a database.

In short, we can say that the main characteristic of a database is the **centralization** of data that it implies which results in a **better data access, management and organization**.

RELATIONAL DATABASES

From the many different models that have been defined for creating a database, the most popular one, both in the GIS context and outside of it, is the one used in databases known as **relational databases**. This model uses a scheme based on **tables**, which is both easy to understand and to use for analysis and data queries. Tables have a certain number of **records** (rows) and **fields** (columns).

The table itself is known as a **relation**, since it contains the relation that exists among its elements. Columns represent the **attributes** associated to a feature, while rows contain the **records**. A row is formed with a set of n attributes which form a **tuple**.

A database usually contains more than a table, since the information to store is of many different types and it is convenient to separate it into several tables. Apart from the relations that the table itself implies, relations between tables can also be defined. This is commonly known as a **table join**. To perform a table join, we need to have some attribute that can be used to unequivocally represent a tuple, known as a **key attribute**, and it must be **unique and invariable** for each tuple. For instance, if we have a table where each row represents a person, an attribute containing the Social Security number can be used as a key attribute.

When working with geographical data, it is common to use **the spatial component as key**, since it is usually unique.

Relations between tables can be of several types, depending on the records of one table that are related with those of the other table. We have **one to one**, **one to many** and **many to many** relations. For instance, if we have a table with cities and another one with persons and

we define a relation *lives in*, it will be a one to many relation, since many people can live in a single city and each person lives in only one of them.

DATABASE MANAGEMENT SYSTEMS

Along with databases, the fundamental element to exploit them are **database management systems** (DBMS). These systems are an **intermediary element between the data and the software that uses them**. Software such as a desktop GIS does not access the database directly, but **through a DBMS**.

The following are some of the characteristics that a DBMS must have:

- **Transparent access to data.** The DBMS creates an abstraction of the data that makes them easier to work with, hiding the internal elements that are not relevant for exploiting the data.

Procedures such as **queries** are done through a DBMS which takes care of interpreting them, applying them on the database and returning the corresponding result. The GIS does not query the database, but instead communicates with the DBMS.

- **Data protection.** If the database contains sensitive information, a DBMS must **control the access to it**, restricting it to certain users and implementing the protection mechanisms that are needed.
- **Efficiency.** A DBMS must be capable of efficiently handling **a large volume of data and a large number of operations** (for instance, many users accessing simultaneously), and provide a quick response to user requests.
- **Transaction management.** Operations on a database such as adding or deleting a record are performed using a

transaction. A transaction is a unit of work performed within a database management system against a database. A DBMS is said to be **transactional** if it can guarantee the integrity of the data and does not allow transactions to remain uncompleted.

Since software such as a GIS communicates with the DBMS and does not access the database directly, a language to establish this communication is needed. Languages used to make queries to a DBMS are known as **query languages**. The most popular of these languages is the **Structured Query Language (SQL)**.

Spatial databases

We have reviewed the fundamental ideas about general databases which can contain any type of data. Adding spatial data to this is not trivial, as it adds more complexity and makes it necessary to use a different approach. For a database to be considered spatial, it should be adapted to the particular nature of spatial data, and include additional elements.

First, the database has to be able to **store spatial data natively**. That means that a geometry can be stored in the table, just as happens with other data types that can be used for table attributes, such as numerical values or text strings. The database must be able not only to store spatial data, but also to **understand it** and be aware of its properties, so it can support queries related to that data.

This is what makes the database fully spatially enabled, unlike a storage mechanism in which the geometry is stored using some of the basic data types (for instance, using a string containing the geometry coordinates), and the database does not know about its spatial nature (it

does not differentiate between that string and any other containing other type of information).

Although raster data can be stored as well, **spatial databases work mostly with vector data** and are better adapted to them. The geometries are stored as part of the attributes of a table record which corresponds to a feature in the vector representation model. The thematic component can be stored in the database without requiring further adaptation.

Assuming that the database is prepared to store spatial data and correctly work with it, we now need to **adapt the query language**. Along with the usual operations that a DBMS can perform, new ones are added that use the spatial properties of spatial data. A query language that supports queries related to the spatial component of the data is known as a **spatial query language**.

QUERIES

A query is an operation in which we *ask* the geographical data about the information they content. This type of analysis is one of the key elements of GIS, since it represents a large part of the work that is done with a GIS software.

Although queries are not exclusive of databases, they become more powerful and efficient with the help of a DBMS and a query language.

In the context of GIS, a query represents something similar to what we do when we use a classic paper map and we respond to questions such as *which is the closest river to X city?* or *which rivers cross the Y province?*. We must not forget, however, that geographical data has two components: a thematic one and a spatial one. Questions such as the ones above refer only to the spatial component

but we can make queries that refer to the thematic one, or to both of them simultaneously.

A very simple example of a query is **selection**. This is an operation that is commonly performed in a GIS, to just work with a subset of all the features in a layer. In figure 7.1 you see how the GIS user defines a rectangular area and features that fall within its limits are selected. Selection criteria can be as simple as this one or more complex and they might also include the thematic component (here we are just using the spatial one).

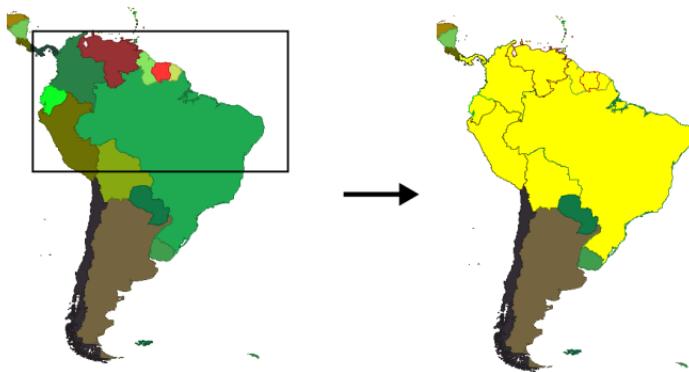


Figure 7.1: Manual selection of features by defining a rectangular region.

A query can also be used to extract certain information from a database according to our needs and to later create a new layer with it. This operation is very useful when the database contains a large amount of data but we only need a part of those data. We might create a subset based on spatial criteria (for instance, if the database contains information for the whole world and we just want data corresponding to a given country), a thematic one (the database contains many attributes associated to each feature, but only a few

of them are of interest to us), or a combination of both. To extract that information and create a subset of the original data, we will use a query.

Let's consider more examples of queries. Let's assume that we have a layer with the world's countries, and a set of economic and social parameters associated with each of them. For each country, we also have a polygon representing its boundaries.

We can make queries like the following:

- Which countries have a GDP larger than Spain's?
- Which countries have grown economically during the last year?
- Which countries have a population of more than 200 million people?

In these queries, we are not using the spatial component (we do not need the polygon associated to each country). We could make those queries if we had country data without any spatial component and were using a regular database with no spatial capabilities.

Queries might include several criteria. For instance:

- Which countries that have grown economically during the last year have a population of more than 40 million people?
- In which countries where English is spoken, has the population increased during the last year?

To express queries in a way that can be later adapted to a query language, we need to use **logical operators**. The above queries would be rewritten as follows.

- Which countries have grown economically during the last year *and* have a population of more than 40 million people?

- In which countries is English spoken *and* the population increased during the last year?

Query languages that can be used to communicate with a DMBS support these operators for queries.

If the DBMS is spatial and *understands* that certain columns of a table contain spatial information, it will support queries that use that information, such as the following ones:

- Which country spans the most degrees of latitude?
- How many countries are completely contained in the southern hemisphere?
- Which countries are located less than 2000 km from Spain?

To respond to these queries, we just need to analyze the spatial component and do not need the rest of the attributes data. These queries are purely spatial. Although they extend what we had done before, we are not adding here any new way of studying geographical data that was not possible without a GIS. We could respond to those queries using just a classic paper map.

The true power of spatial queries is to allow querying both the thematic and the spatial component. For instance, with queries such as these:

- Which countries in the southern hemisphere have a population density higher than that of Peru?
- How many countries with a population of more than 10 million people share their borders with Russia?

These queries require analyzing the thematic component and at the same time, include criteria that are based on the spatial and topological relations of the associated geometries.

Queries can include **several layers**. For instance, if along with the countries layer that we have been using, we have a layer with rivers, we could respond to a query such as *Which countries does the Nile river cross?*. This is a purely spatial query that uses two layers.

Table joins, which were discussed for regular databases with no spatial data, can also be performed with a spatial criteria. These are known as **spatial joins**.

Here is an example of a spatial join. Suppose that we have a layer with world cities and the layer with countries that we have been using in previous examples. We can define a relation between the two corresponding tables, which will associate to each city all the attributes of the country it belongs to. A field with the country name in both tables is needed to use it as the linking point.

However, even if we do not have such a field, we can join the tables if we have spatial data for both cities and countries. All cities that belong to a given country must be located within its boundaries. This can be used to define the relation between the tables and we can know which country a city belongs to just by finding the polygon from the countries list in which the point representing the city is located.

Spatial indexes

If we make a query to a spatial database, responding to it might involve a large number of operations. If, using our countries layer, we want to know which countries have a population of more than 10 million people, we need to read the population of every single country in our table and compare it to that value. If the table has a large number of records, the query might take long to be processed. We can

clearly see that this is not the optimal way of processing a query.

By using what is known as **indexes**, we can reach the data that will form the response of our query in a shorter time, without having to pass through all the data contained in the database.

This is easy to understand using an example. Imagine a telephone book. It contains a large number of entries, but you can easily find a name without having to read them all. This is because a) the data is ordered (**indexed**) in a particular way (alphabetically) and b) you know how to use that indexing (you know the order of letters in the alphabet). With that, you know that it does not make sense to search for a certain Mr. Johnson in the pages that correspond to letter A or B, and you can skip them.

Apart from indexes for numerical or alphanumerical values which are easy to create, another type of indexes, known as **spatial indexes**, are of great importance in the context of GIS. The concept is similar to non-spatial indexes and serves the same purpose: to **optimize searches using a correct data structure**, in this case based on its spatial component.

We will use another example to help understand how an spatial index works. Suppose that we are using our layer with countries and want to find those that are located at less than 2000 km from Spain. How would we respond to this query?

A naive approach would be to measure the distance between Spain and all the remaining countries, then select those at a distance of less than 2000 km. We would get the correct result, but this approach is far from optimal.

Finding a better approach is easy. For instance, with a little knowledge of world geography, we can immediately exclude all countries in the Americas. We can be sure that

they will not be part of the response, since the distance between Spain and the Americas is already larger than 2000 km. We do not know the distance between those countries and Spain but we are sure that it will be more than 2000 km. Therefore, it makes no sense to measure the distances to all of them.

That knowledge of world geography that allows us to reduce the number of countries to work with is actually like a spatial index. It cannot be used to respond to the query, but it **provides an approximation that makes it easier to respond to it**. We can discard a large number of countries, and then perform the more complex operation (the measurement) with just a subset.

Thanks to spatial indexes, queries are more efficient and we can work with larger datasets.

Indexes (both spatial and non-spatial ones) are stored along with the data they refer to, whether in separate files or inside the database itself. Spatial DBMS have **built-in capabilities to compute those indexes** and store them, and once they have been computed, they are used whenever the DBMS has to respond to a spatial query.

SPATIAL ANALYSIS

Analysis is one of the key capabilities of GIS. Spatial analysis is the **quantitative analysis of phenomena, considering the geometric, geographical or topological properties of their elements**. Properties such as position, distance and area are relevant when performing spatial analysis.

We perform spatial analysis when we use a classic printed map to search for the highest peak in a given map sheet, read the elevation of a given element such as a city, or plan a touristic activity checking the places to visit and how to move between them using the best roads or following the fastest route. Of course, we can also perform this kind of operations within GIS.

Analysis generates new data, and that data can be in the form of **new layers, tables, or simple values**.

The result of an analysis might express **the same variable** as the original data (for instance, computing the average value), or a **different one** (for instance, if we compute a slope layer from an elevation layer).

Spatial analysis requires spatial data, which can be of a **single type**, or, instead, of **multiple types that are combined**. For instance, in the case of finding the highest

point in a map, the result is just a coordinate and the only variable used is the elevation. In the case of computing the average elevation of a city, two variables are used: the elevation and the space occupied by the city (defined, for instance, by a polygon with its boundaries). Although all that information is traditionally contained in a single map sheet, in a GIS it will be in two separate layers, both of which will be inputs for this particular analysis.

Analysis in a GIS can help answer questions related to:

- Position or extension.
- Shape or distribution.
- Spatial associations.
- Spatial interactions.
- Spatial variation.

SOME EXAMPLES OF SPATIAL ANALYSIS

The following sections describe some common types of spatial analysis.

Spatial queries

We already discussed queries in the chapter devoted to databases.

Queries can be combined with other analysis tools for instance, to **select** a subset of features with which we will later perform some other analysis.

Topological analysis

Queries can be referred not just to the position of geographical elements but also to their *relation with other elements*. If we have topological information, we can perform analysis that responds to questions such as:

- How can I reach a give coordinate from my current position using the existing road network?
- Which countries share a border with France?

Measurement

Spatial properties can be quantified and measured. Among the most basic ones, we find length, area, perimeter or shape factors. More elaborated ones such as slope or multiple indices derived from basic measurements can also be computed with the help of GIS.

Combination

One of the most typical procedures within GIS is the **combination and overlay** of layers. The separation of geographical data into layers facilitates this kind of operations and turns GIS into the optimal platform to perform any analysis that requires combining information from different variables.

In the case of vector layers, overlay operations such as **union, intersection, difference or clipping** are frequently used. Figure 8.1 shows an example of an overlay operation between polygon layers.

Transformations

We include in this group a large set of operations that alter the input data in different ways. Among them we find **coordinate transformations, simplification of geometries** or the **creation of influence areas (buffers)**. These transformations may affect both the spatial component and the thematic component of the data.

A particular case, already mentioned in a previous chapter, is the **conversion between representation models**.

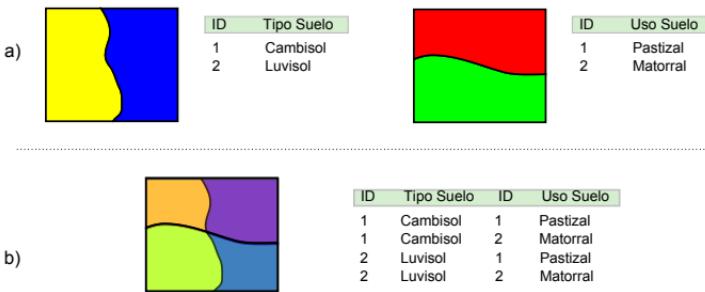


Figure 8.1: Intersection between two polygon layers.

Figure 8.2 shows an example. Starting with a scanned map (a raster layer) with contour lines, these can be traced and a vector layer created based on them. The lines in that vector layer can be later converted into a raster DEM using interpolation techniques. From the raster DEM it is possible to obtain contour lines at an arbitrary contour distance (of course, within the level of detail of the original data).

Terrain analysis

Terrain analysis is one of the most powerful capabilities we find in GIS. From basic parameters such as **slope** or **aspect**, to highly specific morphometric ones, and passing through a large collection of tools for **hydrological analysis**, a vast array of analysis capabilities is available in this field.

Descriptive statistics

The common elements of classic statistics have their equivalents when working with geographical data, and they allow us to **quantitatively describe** the data we work with. Here we include centrality and dispersion measures,

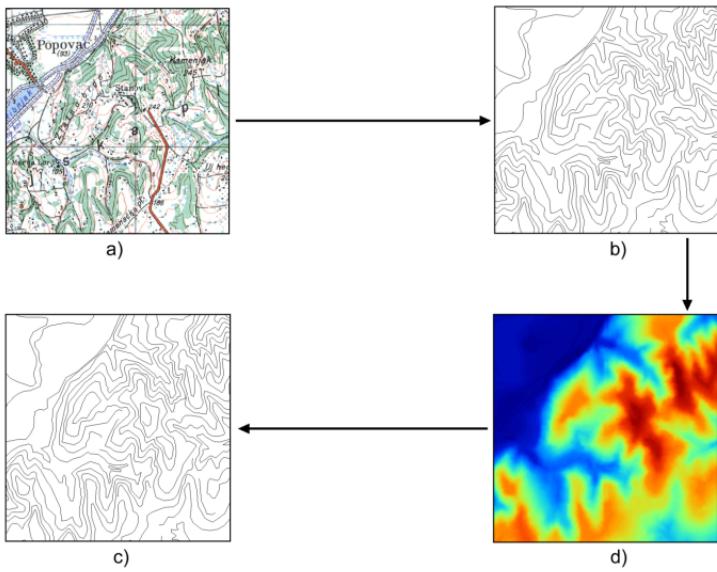


Figure 8.2: Conversion between representation models for an elevation layer.

pattern analysis, and many others. These can be themselves used in hypothesis testing, in case there is a spatial component involved.

These statistical values allow us to respond to questions such as:

- Is average height a constant value across a given country?
- Is there a predominant movement direction for individuals of a given species, or do they move erratically?

Inference

Another important statistical analysis in GIS is the one that helps to infer the behavior of variables and their evolution.

Change modeling is one of the many fields that are rapidly evolving thanks to the help of GIS.

Optimization and decision-making

The layered structure of geographical information in a GIS, which, as we have seen, was ideal for overlay operations, provides also an optimal framework for studying the combined effect of multiple phenomena. GIS is the perfect framework for **multiple-criteria analysis**.

Questions such as the following ones can be responded to using GIS:

- Which one is the best place to build a new power station considering its effect on the environment and the people living close to it?
- Where should a hospital be located to provide the best possible service to the inhabitants of a given region?

PARTICULARITIES OF SPATIAL DATA FOR ITS ANALYSIS

Spatial data have some great potential thanks to their particular properties but at the same time, these properties might **limit or condition** working with them. In some cases, they might represent problems that have to be considered when analyzing the data; in others, they are just something that anyone working with spatial data should know but that are not problematic *per se*.

Scale

We can study geographical information at *different scales* and depending on which one we use, the results obtained will be different. For this reason, apart from considering scale when rendering and visualizing geographical data,

the **analysis scale** should be considered as well when performing any analysis.

The analysis scale should depend on the *data properties* (accuracy, data type, etc.) and the **analysis to be performed** with them.

This can be easily understood with the help of figure 8.3. If we want to categorize the form of terrain at a given point, we need to analyze the elevation of the point and also the elevation in its surroundings. Depending on the size of that analysis window around the center point (which is what defines the analysis scale), the results can be very different. In the image, for a small value of the analysis radius, the terrain will be categorized as being a peak. For a larger region, however, it will be considered the bottom of a valley.

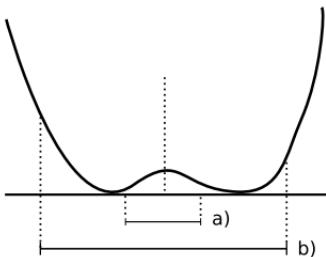


Figure 8.3: Depending on the analysis scale, the form of the terrain can be classified as a peak (a) or a valley bottom (b).

Therefore, we must look at the terrain at the correct distance for which the information that it gives us is the most interesting and correct for the kind of analysis that we are performing. Apart from the fact that there is an optimal analysis scale for each type of analysis, it is also interesting to work at **multiple scales**, as that will provide us more information than what we can obtain working only at a single scale.

Another example of how the analysis scale affects the analysis result is found in the case of **taking measurements**. As it can be seen in figure 8.4, the measurement unit (which is implicitly defined by the level of detail of the data) that is used causes the results to be different.

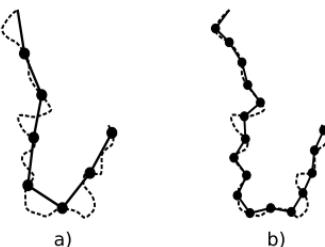


Figure 8.4: Measurement unit affects the value of the measurement.

A value by itself might not have meaning if it is not accompanied by the scale that was used to obtain it.

The concept of **fractal** has a direct link with this.

The Modifiable Areal Unit Problem

Many of the variables with which we work in GIS **cannot be measured at a single point**, and they must be **aggregated for a given area around that point**. Examples of that are the percentage of the population within a given age range or the population density.

Areas defined to work with these variables are **essentially arbitrary** such as countries, counties, districts, etc., and they are defined without taking into account any criteria related to spatial analysis. Using different areas (different units for computing the values of the variable) will yield different results.

This problem is known as the **Modifiable Areal Unit Problem** (MAUP). Solving or reducing its effect is complex

and no solution exists that can be applied in all cases, but whenever we work with this type of geographical data, it is important to keep in mind that there will be a source of statistical bias that cannot be neglected.

Another problem related with the MAUP is the so-called **ecological fallacy**, which result from (wrongly) assuming that the values computed for a given area can be assigned to the individuals of the population within that area. This would only be true in the case of complete homogeneity.

Spatial autocorrelation

Spatial autocorrelation is the **correlation of a variable with itself**, in such a way that values of the variable at any point are correlated with values of that same variable in nearby points. For instance, in the case of temperature, points close to a heat source will have a higher temperature than those far from it or closer to a cold spot. If we study the distribution of an infectious disease, reported cases are likely to appear grouped and a large number of them normally cause the nearby populations to also be significantly affected by the disease.

Another way of expressing this is using the well-known **Tobler's First Law of Geography**, which states that “everything is related to everything else, but near things are more related than distant things”.

In the above cases, spatial autocorrelation is said to be **positive**. However, it can be also **negative**, if higher values are surrounded by lower ones, or there can be no correlation at all, when values in separate points are **independent** and do not affect each other regardless of distance.

Figure 8.5 shows three raster layers which demonstrate the above types of spatial autocorrelation.

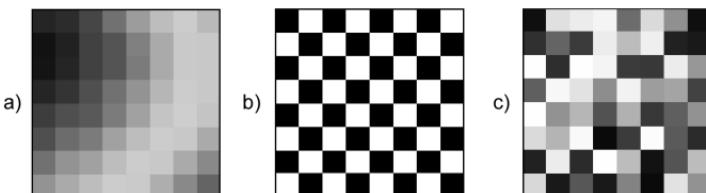


Figure 8.5: a) Positive spatial autocorrelation. b) Negative spatial autocorrelation. c) No spatial autocorrelation (independence).

The existence of spatial autocorrelation has several important consequences.

First, many of the most common statistical analyses assume the independence of the variable that is being studied. Since there is a dependency on the spatial component, this component has to be introduced as another variable to consider, in order to ensure that results are sound.

Something similar happens when data have a **spatial trend** (values of a variable depend on the position; for instance, temperature values, which show a clear trend as latitude changes), since that also invalidates the assumption that data are independent.

If positive correlation exists, **statistical inference is less effective**. The same number of observations contain less information about the phenomena represented by the variable.

The consequences of spatial autocorrelation, however, are not only negative. If points located in the vicinity of a given one are related to it, and the value of a variable is affected by that proximity, that can be used to **estimate values** at any point, knowing the values in a set of nearby

points. That is the fundamental idea behind **interpolation methods**.

Structure

Both the data itself and the properties of the phenomenon they represent (such as the aforementioned spatial correlation) have some sort of structure. This structure can have a relevant effect on the analysis results and should be taken into account.

The two basic statistical concepts related to the spatial structure of a process are **stationarity** and **isotropy**. Stationarity indicates that the process is **translation-invariant**. That is, its properties are constant across the whole space and there is no spatial trend. Isotropy means that the process is **rotation-invariant**, and happens in the same way in all directions.

Border effects

The areas in which we perform spatial analysis **have boundaries**. These might be artificial, for instance, the limit of the aerial photograph we are working with, or natural. If we study a forest that is close to a lake, the shore will be the limit of the forest. Boundaries **distort the result of analysis**, specially for those variables that have to be aggregated (density, etc., as we saw for the case of the MAUP)

In some cases, the border effect might manifest itself only for those points close to the border. In others, however, **all the points related or somehow connected to the border** might be affected, regardless of the distance to it.

VISUALIZATION OF GEOGRAPHICAL DATA.

When working with a GIS, most of the time we will visualize the data we work with. Although certain data, such as satellite images or maps from a map server, include their own rendering and can be visualized *as they are*, in most cases it is the user of a GIS who defines the way geographical data is rendered. In other words, the GIS user **takes the role of the cartographer** and for this reason, must be familiar with the ideas and techniques used by cartographers.

Along with the concepts and tools from classic cartography, GIS include elements from what is known as **scientific visualization**, such as **interactivity** or **multi-dimensional data rendering**. This approach, richer than the classic one from cartography, is known as **geovisualization**.

In this chapter, we will see some fundamental ideas about data visualization, and how they are applied to both the traditional field of cartography and the GIS and geovisualization context.

BASIC IDEAS ABOUT DATA VISUALIZATION

When we visualize any kind of geographical information, whether on a computer screen or on a printed map, we are using a **visual language** to convey that information.

The study of signs of a language is called **semiology**. In the case of a visual language, we have a **graphic semiology**. This semiology works with the signs of the language that we use to visualize geographical data and helps us understand why and how visual elements serve their purpose of correctly conveying the information from which they are created.

Visual variables

Visual elements have several properties that can be used to transmit information. Depending on the case, some of them might more suitable than others.

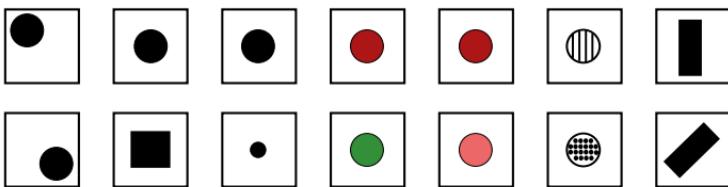


Figure 9.1: Visual variables. From left to right: position, shape, size, hue, value, texture and orientation.

These properties are known as **visual variables** and are applied to the geometric elements used to visualize geographical information. Those elements can be differentiated using the following visual variables, which are shown in figure 9.1: position, shape, size, hue, value, texture and orientation.

The use of **position** is rather restricted in the case of a map, since the real position of the element to be rendered should be respected. It is seldom used.

The **shape** is defined by the perimeter of the object. This variable is mostly used in the case of point data, using a symbol of a given shape located at the exact coordinates of the point to be rendered. It is difficult to apply to linear symbols and in the case of areal symbols it requires altering the shape of the symbol itself.

Size indicates the dimensions of the symbol. In the case of points, it can be applied by changing the size of the symbol itself. In the case of lines, changing their thicknesses is the most usual way of applying this visual variable on them. It is not used in areal symbols, except in the case of using a texture fill, in which the size variable is applied to the texture and not to the symbol itself.

Size alters how other visual variables are perceived, especially in the case of small sizes.

Texture refers to the pattern used to fill the body of the symbol. It can be applied to lines, using dash patterns, but it is mostly applied to areal symbols.

Color is the most important of all visual variables. Two of its components can be used as individual visual variables themselves: hue and value.

Hue is what we usually call color. That is, the name of the color (blue, red, green, etc.)

Hue can be altered by the hue of surrounding elements, especially in small symbols. Although human perception has a great sensitivity, it might be difficult to identify in small symbols, and it can be wrongly identified if the symbol has other larger ones with different hues in its surroundings.

Value defines the darkness of the color. For instance, light blue and dark blue have the same hue, but they have different value.

Differentiating two symbols by their value can be difficult depending on the type of symbol. It is easier in the case of areal symbols, while in the case of linear and point symbols it depends on their size. Smaller sizes make it more difficult to compare values and to extract the information that the visual variable is trying to convey.

Orientation is applied to point symbols, unless they have some sort of symmetry that makes it difficult to identify the orientation of the symbol. For areal symbols, it is applied to their texture. It's not applied in the case of linear symbols.

Properties of visual variables

Visual variables can have four basic properties.

- **Associative.** A visual variable is said to be associative if, when applied, doesn't change the visibility of an element. That is, it's not possible to give more importance to an element using that visual variable.
- **Selective.** A visual variable is said to be selective if, when applied, generates different categories of symbols.
- **Ordered.** A visual variable is said to be selective if it can be used to represent a given ordering.
- **quantitative.** When, apart from being ordered, it can be used to express ratios.

In the above list, variables are ordered according to the so-called **levels of organization**. The associative property is at the lower level, while the quantitative one is at the highest. The level of organization of visual variables is relevant when combining them, as we will see later. Also,

the level of organization of a variable defines the type of information that the variable can transmit.

Figure 9.2 shows different renderings of a set of point symbols, explaining in each case, one single visual variable.

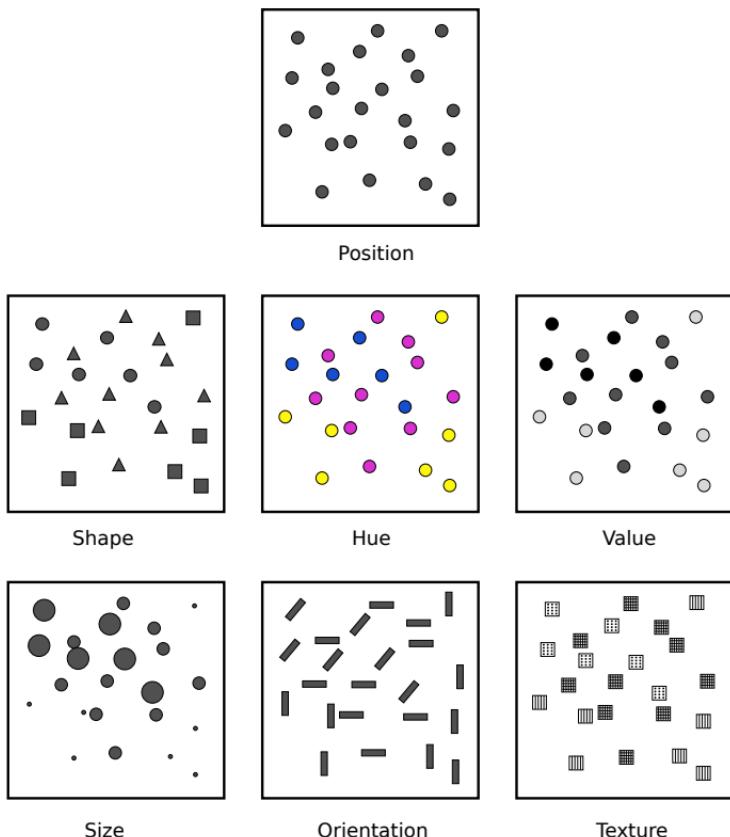


Figure 9.2: Visualization of a set of point symbols using a single visual variable in each case.

Starting with the associative property, we see that, except for size and value, all other visual variables do not do not emphasize one element over the others. In other words, one element is not seen as more important than the rest

of them when the visual variable is texture, color, shape or position.

With size, however, it is clear that a larger one gives symbols a more prominent role. In the same way, a darker value attracts the attention of the observer much more than a color with a lighter value.

Regarding the selective property, we can say that a variable has a selective quality if, at a quick glance, we can easily identify the elements that belong to a given group which is defined by a visual variable. The clearest example of this is hue. We can quickly separate from a set of symbols those that are red or yellow. All visual variables, excepting shape, have this property, although it might not be so as in the case of hue. Shape does not make elements form groups spontaneously.

The ordered property is found in those visual variables that we can use to define an ordering. Only position, texture, size and value are ordering properties. For instance, in the image corresponding to the visual variable hue, we cannot say which element we would place at the beginning or end of a scale defined by hue itself. With value, however, we can, since that scale would range between the lighter tones to the darker ones, and we can visually differentiate and sort them.

Finally, the quantitative property is found in those visual variables that can be used to visually estimate quantities and ratios. Only position and size have it. For instance, we can see that the big circles in the image corresponding to the size visual variable are more or less twice the size of the smaller ones.

Table 9 contains a summary of all these ideas.

Visual variables can be combined (for instance, representing objects with different size and hue). The properties of all the visual variables that are used must be considered,

	Position	Size	Shape	Value	Hue	Texture	Orientation
Associative	◊	-	◊	-	◊	◊	◊
Selective	◊	◊	-	◊	◊	◊	◊
Ordered	◊	◊	-	◊	-	-	-
Quantitative	◊	◊	-	-	-	-	-

Table 9.1: Properties of visual variables.

and if a given property is needed for the information that we want to convey, all those visual variables should have it.

The perception of visual variables

The perception of visual variables **might be altered by the environment**. It is important to study this from two points of view: **perceptual constancy** (how much we can modify visual elements and their surroundings before they fail to convey the same information and can be misidentified) and **perceptive aids** (how we can help visual elements to be perceived exactly in the way that we want).

Perceptual constancy defines how objects are **perceived in the same way regardless of the changes in the environment**. For instance, if an object is round, such as a wheel, it will have a round shape when we look at it from a perpendicular direction. If we now look at it from a different angle, we will see an ellipse instead of a circle. However, we will read it as round and will still identify its shape correctly. That is an example of the perceptual constancy of the shape.

Not all visual variables have such a perceptual constancy. When the perception of an element changes even if the object itself does not, a **perceptual contrast** is said to exist. Perceptual contrast might cause a visual element to be wrongly perceived and the information that it transmits to be misinterpreted.

The following are some of the main ideas about perceptual contrasts to take into account when creating a map:

- Size is the visual variable that is more affected by perceptual contrasts. The apparent size of an object might change if it is surrounded by other elements of a different size. This is particularly relevant when using point symbols in a map.
- Values is also altered when other elements with a different value appear nearby, specially if there are a large number of them.
- Hue is altered by the presence of other hues. In a map, we should consider how the background color might affect the foreground symbols.
- Complementary hues, when put together, might cause a vibration sensation in the border between them.

Regarding perception aids, the most important factor when creating a map is the **correct separation between the foreground objects and the background**. The properties of the visual variables must be used to create different levels in the visualization, assigning more relevance to some elements in order to focus the attention on the information that they transmit.

To make certain layers (the most relevant ones for the purpose of the map) more visible, a **correct hierarchy** must be established with the help of visual variables. This hierarchy will add depth to the information displayed in the map, and some elements will be perceived as being more

important than others. Layer ordering already defines a structure and a hierarchy, but that is not enough in most cases and visual variables should be used to reinforce it.

Figure 9.3 shows why a correct hierarchy is needed to create a good map.

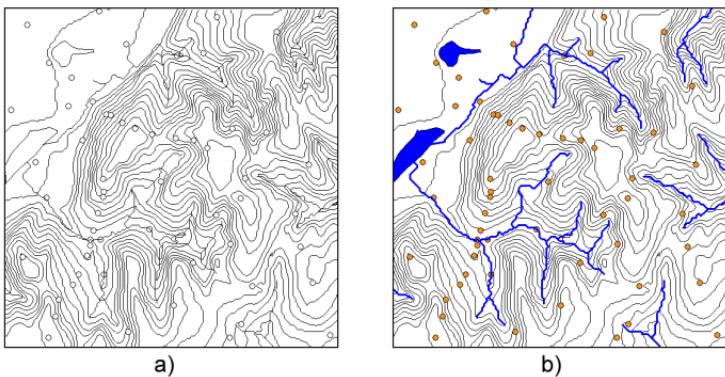


Figure 9.3: Comparison between a map without hierarchy (a) and a map with a correct hierarchy (b).

MAPS AND CARTOGRAPHIC COMMUNICATION

Maps are a method of communication that uses a language with a particular purpose: **describing spatial relations**. A map is, therefore, a symbolic abstraction of a real-word phenomenon, which implies that it has some degree of simplification and generalization.

The visual language that we have just seen becomes a cartographic language when it is adapted to the particular case of creating maps and knowing its rules is needed to create cartography that is later useful for the map user. All these ideas related to map production form what is known as **cartographic design**.

Cartographic design involves making decisions (in this case, by the GIS user who takes the role of the cartographer). These decisions must be guided by the **purpose of the map** and the **target audience** and depending on these factors, the cartographer must decide **the projection** (which does not always have to be the original one of the data), the **scale** (depending on the level of detail and taking into account the limitations of the data), the **type of map** (we will see more about this later in this chapter), or the **symbols** to use, among other things.

There are two main types of cartography: **base cartography** (also called **fundamental** or **topographic**) and **thematic cartography**.

Historically, base cartography represents the classic maps that have been created by cartographers. This type of map serves the purpose of precisely describing *what is* on the surface of the Earth.

Thematic cartography focuses on **displaying information about a given phenomenon** (a given geographical variable), which can be of any type: physical, social, political, cultural, etc. We exclude from this list those phenomena that are purely topographic, which are the subject matter of base cartography.

We can also say that base cartography represents **physical elements** (a stream, a coast line, a road, a valley, etc.), while thematic cartography focuses on **representing values and attributes**.

Thematic cartography uses base cartography (usually included in thematic maps) to help the map user to understand the spatial behavior of the variable being represented, and also to provide a geographical context for it.

Types of information and their visualization

We already know that the thematic component of geographical information can be numeric or alphanumeric and that numeric variables can be nominal, ordinal, intervals, or ratios. Selecting a correct symbology according to the type of information that we are working with is key to producing an effective map. In particular, we must use a visual variable that has the correct properties (levels of organization) for the variable that we want to visualize.

For instance, the associative property and the selective property are of interest just for qualitative information, while size is the only visual variable that we can use that has the quantitative property and therefore, the only one that should be used to represent ratios.

The following are some of the more important ideas about this, referred to the aforementioned types of information.

- **Nominal.** Nominal information is correctly represented using the visual variable shape. This information shows *what* is found in the different locations of a map, and not *how much* is found, and it is more related to base cartography than to thematic cartography. Using different symbols for point elements and line elements is a common and very effective solution. For the case of areal symbols, hue and texture are the most common solutions.

Alphanumeric information has similar properties, and the same ideas apply to it.

- **Ordinal.** Since values of the variable define an order, a visual variable with the ordered property is needed to correctly visualize this type of information
- **Interval and ratio.** Visual variables with the ordered property can be used in this case. However, size is a better

choice, as it is the only one which has the quantitative property.

Values are normally grouped into classes so the same value of the visual variable (same size of the symbols or same color value, for instance) is used for different values of the variable that we are visualizing. There are different strategies for this, which try to maximize the information that the map transmits. The most commons ones are **equal intervals**, **intervals using percentiles** or **natural intervals** (intervals that try to minimize the variance within each class).

Using one or another of these methods can have a noticeable effect in the visualization, as is shown in figure 9.4.

It is important to remark that, although levels of organization indicate increasing potential (that is, with a variable such as size or value we can convey all the information that can be conveyed with hue, since they have properties with a higher level), **it is not always better to use visual variables with a higher level of organization**, and it is not true that they will always be better than those with a lower one. For instance, using the visual variable value for a map with qualitative information (like using a ramp of different tones of blue for a map with soil type information) might not be a good idea, because it has the ordered property, and that might cause the map user to think that there is some hierarchy (that some soil types are “better” than others), which is false.

Map elements. Map composition

A map is not just the part that represents the geographical information, but a set of multiple elements, for example, the one that contains the geographical information itself.

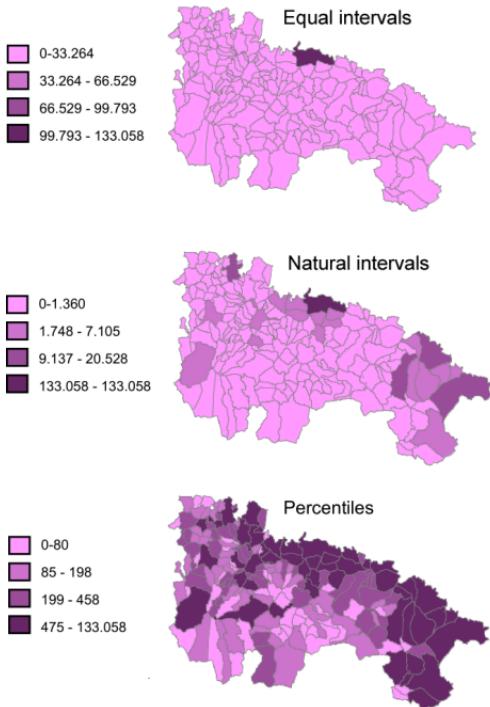


Figure 9.4: Comparison between different methods of defining intervals.

A **correct layout of the map elements** is as important as a correct symbology, since these, like symbology itself, are designed to help the map user to better interpret the information that it contains.

The following are the main elements that can be used to compose a map (Figure 9.5):

- **Name or title.** Needed to know what information is contained in the map.
- **Author.** Creator the map.

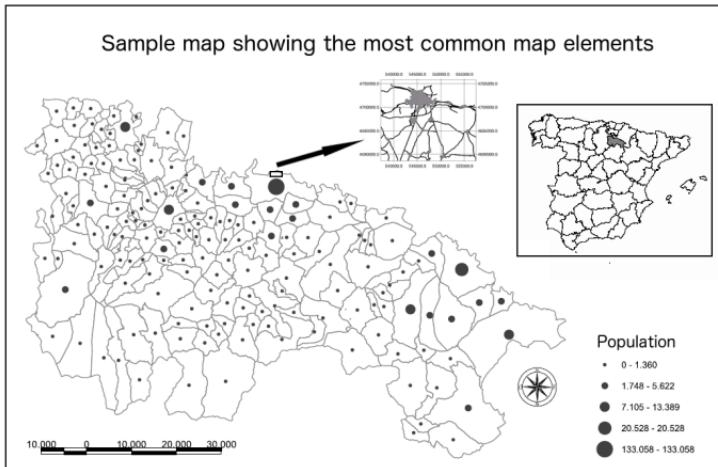


Figure 9.5: Map with its most common elements.

- **Additional information about the map.** For instance, the coordinate reference system used or its creation data, among others.
- **Data frame.** The frame which contains the rendered geographical information. It is the central element and will use most of the space of the map.
- **Graticule.** On top of the data frame, it locates the content of the map on the Earth and provides a geographical reference. It serves the same purpose as the scale, helping to estimate distances. It is usually added at all scales, but it is more relevant in the case of small scales.
- **Legend.** When designing a map, we should try to use a symbology that is as expressive as possible. However, sometimes it is not possible to include all the information with just the symbology itself and a legend is required. The legend has to be clear and easy to interpret as well. A legend that is too large or difficult to understand is

probably telling us that the symbology that we have selected can be improved.

The legend and the data frame form a single unit and should be together (legend inside the data frame), not separated in different frames or with boundaries between them, unless the data frame uses all the space of the map and it is not possible to visually separate both elements clearly.

- **North arrow.** Although by convention maps have a south-north orientation (north is at the top of the map), it does not always have to be that way. An arrow pointing north or a compass rose will help to clarify the map orientation.
- **Scale.** Map scale should be displayed both numerically and graphically (scale bar).
- **Locator map.** Allows the user to locate the map in a larger geographical context. It is especially relevant in the case of map series, to show the relation between the current map and the rest of them, acting as an index map.
- **Detail maps.** Used when there is an area that we can show with a greater level of detail. The area that it corresponds to should be indicated as well in the main map.

It is also important that the map **emphasizes its purpose**, giving more importance to those element that serve it better.

TYPES OF THEMATIC MAPS

There are many different ways of visualizing a given variable in a map. Several of them can be combined in a single map, especially if it includes more than one variable. In this case, the combination should strive to obtain the maximum possible clarity for all of them so the rendering of a variable does not overshadow the remaining ones.

In this section, we will describe the following types of thematic maps: proportional symbol maps, point density maps, isoline maps and choropleth maps.

Proportional symbol maps

A proportional symbol map represents **quantitative variables** using symbols whose sizes **are proportional to the value** of the variable. That is, the map uses the visual variable size (the only one with the quantitative property) to transmit the value of the variable being represented. If the symbol used is linear (such as a bar), its length is used to scale the values to render. If it is areal, area is used. That means that, in case of using circles, a value three times larger than a reference one will not be rendered with a circle with a radius three times longer, but with a circle with an area three times the area of the reference circle.

Symbol scaling can be done in a continuous way, but it is usually more convenient to use a discrete approach, grouping values in classes and assigning a single size for all values in each class, usually the size that corresponds to the center value of the class.

To avoid problems when perceiving the size of each symbol, it is important to show in the legend the relation between the different sizes and their corresponding values, as can be seen in figure 9.6

Point density maps

Point density maps are particularly suitable for countable variables such as population or crop yield. These quantities are represented using **repeated points** whose number is proportional to the quantity itself. Each point represents a unitary value and all the points within an area add up to the total value of the variable in it. All points have the

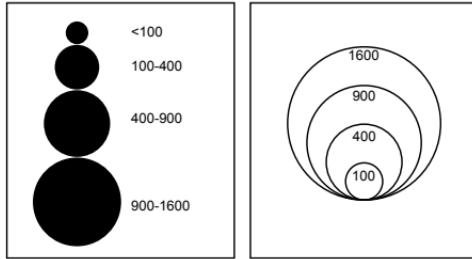


Figure 9.6: Two types of legend for proportional symbols maps.

same size and shape, unlike what we saw in the case of proportional symbols.

When creating a point density map, three parameters have to be defined: the **value of each point** (that is, how many units of the variable the point represents), its **size** and its **position**.

The value of each point should be defined **based on the range of values** covered by the variable, so points in the resulting map are not too scarce or too numerous. This value will be included in the legend, usually in text form, writing, for instance, that “a point represents 1000 inhabitants”.

Size must guarantee that points are visible and at the same time they do not take too much space in the map. The **optimal size is linked to the selected value of each point** and both parameters should be considered together, so as to find the best combination of them.

The position of the points is of great importance, as it should not convey wrong information or cause the user to misinterpret its meaning. If we do not have any additional information, points should be regularly distributed, covering the whole area that correspond to the variable value. If, on the other hand, we have more information about the distribution of the variable, we should use it to give the

points a more realistic position. For instance, if we are creating a density map with population values for regions, there should be more points in the surroundings of the cities within the region, since there are more inhabitants in those areas.

Another thing to consider is the meaning of the variable and whether or not the phenomenon that it represent can appear at a given point. For instance, if the variable that we are representing is the number of water birds know to nest in each region, it will be wrong to place the density points in forest areas or city ones, since it might be inferred that birds are found there, which is likely not true.

Image 9.7 shows an example of a point density map.



Figure 9.7: Point density map.

Isoline maps

Isoline maps are commonly used to represent **continuous variables**. Containing only lines, they mix well with other types of maps without being obtrusive.

An isoline map is formed by a set of lines, each of them connecting points that have the same value of the variable being represented. These lines cannot cross with

each other, since a point cannot have two values at the same time. The most common use of isolines are contour lines in topographic maps which represent points with the same elevation.

Isolines are defined by their **equidistance** which indicate the difference between the values represented by any two contiguous isolines. A lower equidistance means more isolines and a denser map.

Size is the only visual variable used with isolines. It is used to highlight those that represent a value that is a multiple of a given number, to make the map easier to read. These lines are known as **index lines**.

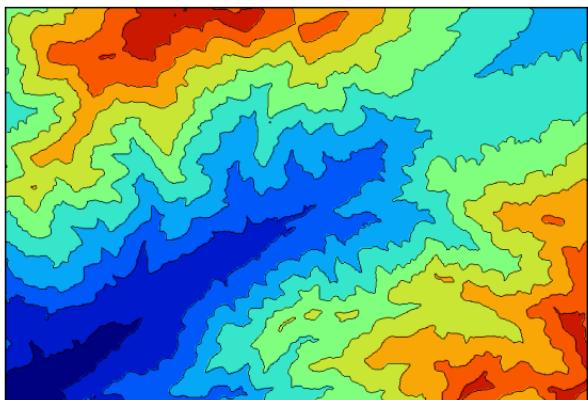


Figure 9.8: Map with isolines and hypsometric tints.

Lines are labeled with their value over the line itself, usually only on the index lines (the value of other lines between the index ones can be easily figured out knowing the equidistance).

A particular case of isolines are the so called **hypsometric tints**. Apart from drawing the lines themselves, the areas between them are colored, each with a different color (usually using a graduated scheme).

Figure 9.8 shows an example of this.

Choropleth maps

Choropleth maps are a very common type of map in GIS. For instance, maps in figure 9.4 were all choropleth maps.

In a choropleth map, there is a set of areas, each of them representing a single value of a variable. This value applies to the whole area, and is normally represented using hue applied to the areal symbol.

Choropleth maps have some important limitations. One of them is the **sharp change in the boundaries between areas**, which might be interpreted as an abrupt change in the variable value in that boundary. This could hide the continuity of the variable distribution, in case it exists. Another problem is the **homogeneity within each area**, which might lead to thinking that the variable has a uniform distribution, even if that is false.

In many cases, and in order to correctly transmit the information contained in the variable, its values have to be **normalized using the area of each region**.

VISUALIZATION IN A GIS

Now that we know the basic ideas about visualization and how they are applied to maps, it is time to see how these are used in the context of a GIS. Two ideas are particularly relevant in this contexts: the fact that we work with **multiple layers** to be represented together, and the particularities of **on-screen rendering and the interactivity it offers**.

Combining multiple layers

In most cases, visualizing a layer alone is not the best way of visualizing the information it contains. In a map, we

normally find several types of information, and that is not just for the sake of space but because it helps the map user to understand and interpret the main information. For instance, contour lines help to understand the shapes of rivers and lakes, providing a valuable context.

When combining layers, we should try to create a synergy between them so they complement each other. This is mostly done by **correctly ordering the layers** and using **a symbology for each of them that does not interfere with the others**.

When two layers have information for the same location, only the information of the layer on top will be seen. Layer ordering should maximize the information seen in the map and prioritize the most important layers over those that contains secondary information.

We know that raster layers fill the space and contain values in all of their cells (pixels in the case of an image). For this reason, they will cover whatever is underneath, and is not a good idea to place them at the top of the rendering order. Instead, they should be considered as **base layers on top of which the remaining layers are placed**.

With a similar reasoning, we can define the best way to order vector layers, placing polygons first, then lines, and then points at the top of the rendering order.

Sometimes, the rendering order **might be imposed by the meaning of layers**. For instance, if we are creating a map with a layer containing streams and another one containing roads, this last one should go on top of the first, since roads usually pass over streams and not the other way round.

A common functionality that most GIS have is the use of **transparency** for layers. It can be applied to both raster and vector layers. Figure 9.9 was created using this technique. The polygon that defines the boundary of the

watershed has a semi-transparent fill, which allows to see the shaded relief layer underneath. The result is a map in which the hydrological meaning of the watershed is much clearer and easy to understand.

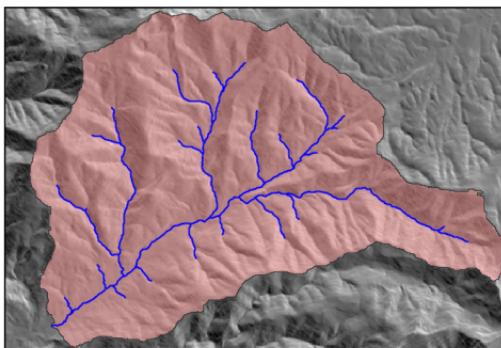


Figure 9.9: Combination of two layers using transparency.

In the case of raster layers, transparency can be applied partially, just rendering those cells that are within a given range of values.

If a variable is divided in many **separate layers** (horizontal division), the **same symbology** must be applied to all of them, in order to have a coherent map.

Particularities of on-screen rendering

Apart from the ideas that are applied to printed maps, additional ones must be taken into account when the visualization occurs instead on a computer screen. A printed map should not be designed in the same way as a map that is meant to be rendered on the screen.

The main elements to consider are the **low resolution of the screen** when compared to a printing device and the **interactivity** of the visualization itself.

When creating a printed map, resolution is not a problem, since printing devices offer a level of detail that goes beyond what the cartographer might need. However, screen resolution is much lower and certain elements might not be rendered with enough clarity. Although these elements can be used in printed maps, they should be replaced for on-screen maps. Among these problematic element we find **fonts with ornaments** such as shades, **fonts with serifs** (small lines attached to the end of the stroke to increase readability) or **texture fills of small size**.

Regarding interactivity, we must take into account that, unlike a printed map, an on-screen map is not a static element, but a dynamic one. That does not mean that the map changes by itself but instead that the user can alter it using the tools that we have already seen (zoom, pan, etc.)

Since the scale can be changed by the user, that might cause problem with certain elements such as symbols and text labels. If all elements are scaled proportionally, reducing the scale will make the labels too small and impossible to read. On the other hand, if the scale is increased, labels might be too large, as can be seen in figure 9.10.

A solution to that is to use an **absolute size** for those elements, so they always have the same size regardless of the scale. With lower scales, however, that might result in maps that are saturated, as can be seen in figure 9.11.

The ability of GIS to render layers at different scales can also cause **performance issues**. At low scales, the number of elements to be rendered can be too large and painting them on the screen might take too much time. To avoid these problems, a **multiscalar approach** can be adopted, in which, depending on the scale, different layers and elements are rendered. For the same information, different versions with different levels of detail can be used, each of them being used only at a given scale range.

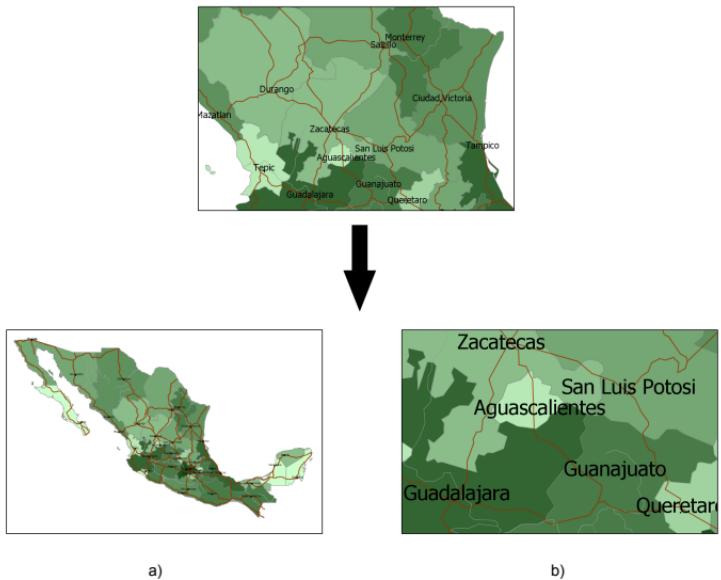


Figure 9.10: A scale change modifies the size of symbols and text labels, and it can make them too small (a) or too large (b).

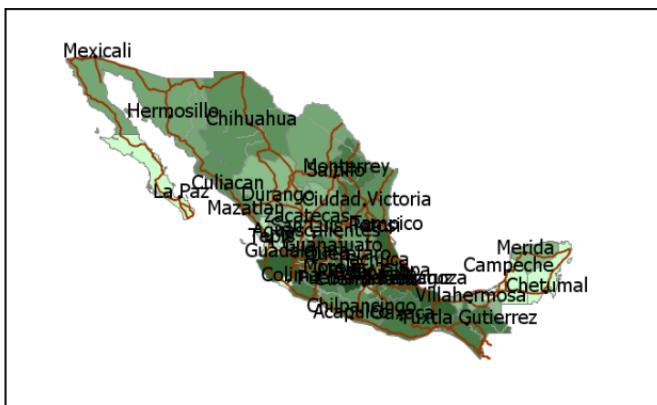


Figure 9.11: Saturated map caused by rendering elements at a fixed size at a low scale.

CONTENTS

1	What is GIS?	1
2	History of GIS	5
	The evolution of GIS as a discipline	6
	The evolution of technology	7
	The evolution of data	9
	The evolution of theories and techniques	10
3	Fundamentals of cartography and geodesy	13
	Basic concepts of geodesy	13
	Reference surfaces	14
	Coordinate reference systems	16
	Coordinate conversion and transformation	18
	Basic cartographic concepts	19
4	Geographical data	25
	Data and information. Types of information. . .	25
	Subdivision of information. Layers	27
	Geographical information models	29
	Raster model	30
	Vector model	31
	Raster vs vector	34

5 Data sources	37
Remote sensing	38
Electromagnetic radiation	39
Sensors and platforms	40
Photogrammetry	42
Printed cartography. Digitization	43
Digitization Quality	46
GPS	48
Voluntary Geographical Information	50
Metadata	51
Content of metadata. Metadata creation	53
6 Software and technology	55
Desktop GIS	55
Data input and output	56
Visualization	56
Analysis	57
Editing	58
Map design	59
Web mapping. Clients and servers	60
Some techniques related to GIS services	65
Standards	67
Relevant standards in GIS	70
Mobile GIS	71
7 Databases	75
Relational databases	77
Database management systems	78
Spatial databases	79
Queries	80
Spatial indexes	84
8 Spatial analysis	87
Some examples of spatial analysis	88
Spatial queries	88

Topological analysis	88
Measurement	89
Combination	89
Transformations	89
Terrain analysis	90
Descriptive statistics	90
Inference	91
Optimization and decision-making	92
Particularities of spatial data for its analysis	92
Scale	92
The Modifiable Areal Unit Problem	94
Spatial autocorrelation	95
Structure	97
Border effects	97
9 Visualization of geographical data.	99
Basic ideas about data visualization	100
Visual variables	100
Properties of visual variables	102
The perception of visual variables	105
Maps and cartographic communication	107
Types of information and their visualization	109
Map elements. Map composition	110
Types of thematic maps	113
Proportional symbol maps	114
Point density maps	114
Isoline maps	116
Choropleth maps	118
Visualization in a GIS	118
Combining multiple layers	118
Particularities of on-screen rendering	120

