GIS

A Computing Perspective

Third Edition

Edited by Matt Duckham, Qian (Chayn) Sun and
Michael F. Worboys

First published 1995

ISBN: 978-1-4665-8719-9 (hbk)
ISBN: 978-1-032-53973-7 (pbk)
ISBN: 978-0-429-16809-3 (ebk)

Chapter 8

Cartography and Geovisualization

DOI: 10.1201/9780429168093-8

The funder for this chapter is Matt Duckham

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

Cartography and Geo-visualization

8

Sections

Cartography and geovisualization is concerned with the design, evaluation, and implementation of effective user interfaces between people and GIS. Ultimately, GIS are only useful if they can support people in decision-making.This chapter is designed to help you develop the skills to:

- construct effective **data graphics** for presenting static, non-spatial data;
- understand, design, and critique **maps** used for presenting spatial data; and
- apply the principles of **interaction design** to GIS interfaces that go beyond presentation of spatial data to help in analysis, exploration, and decision-making with geographic information.

WHAT is your favorite map? Even if you have never thought about this question before, you probably already have some ideas. Some people prefer the detail and precision of a traditional topographic map, perhaps showing an area with which they are familiar (e.g., Figure 8.1a). Others prefer the simplicity and practicality of a schematic map, such as is often used for public transport maps (e.g., Figure 8.1b). Perhaps you are attracted by the handmade beauty of a historical map, complete with its distortions and ornaments (e.g., Figure 8.1c). Or maybe you thought of a digital map: one that you can access from a computer or mobile device, centered on your location, enabling you to zoom and pan, select layers, or be guided along a route.

This chapter is about the visual display of geographic information, such as illustrated in the maps in Figure 8.1. In Section 8.2, we focus specifically on the art and science of map design and making, termed *cartography*. First, though, Section 8.1 reviews the basic principles of data graphics, which underpin the visual presentation of any information. Display is only part of the story, however. This chapter is also about *interaction* with geographic information. Here again, we shall start with the basic principles of interaction, covered in Section 8.3, before moving onto what makes interaction with geographic information special, in Section 8.4.

cartography

a. 1944 1:62500 topographic map (Source: U.S. Geological Survey)

b. St Petersburg subway map (Copyright Alex Florstein)

c. Ortelius' 1570 Thietmarsia map (Source: U.S. Library of Congress)

Figure 8.1: What is your favorite map? Portions of three different types of maps

8.1 Data graphics

data graphics

Figure 8.2 shows two different graphical presentations of related data. Which do you prefer? Which do you think is better? For most people, the question of which graph is *better* is not so easy to answer as which is *preferable*. However, there are a number of principles one can apply to any graphical depiction of data for exactly this purpose. The term *data graphics* is used to refer to any graphical depiction of data, including graphs, charts, and maps.

8.1.1 Principles of data graphics

Based on Tufte's famous book *The Visual Display of Quantitative Information* (Tufte, 2001), it is possible to identify four main principles of data graphics: density, integrity, correspondence, and aesthetics.

data density

Data density We have all heard the adage "a picture is worth a thousand words." By the same logic, a data graphic that requires a description of *two* thousand words has a greater density of data than a similar graphic that can be described using only *one* thousand words. The principle of *data density* is concerned with presenting as much data as possible in the space available (see, for example, the "data-ink ratio" in Box 8.1 on page 310).

Data density can be increased by increasing the amount of data in the graphic; it can also be increased by decreasing the amount of redundant lines, text, and marks. Comparing the data graphics in Figure 8.2, Figure 8.2b has much higher data density than Figure 8.2a. Figure 8.2b shows data for more

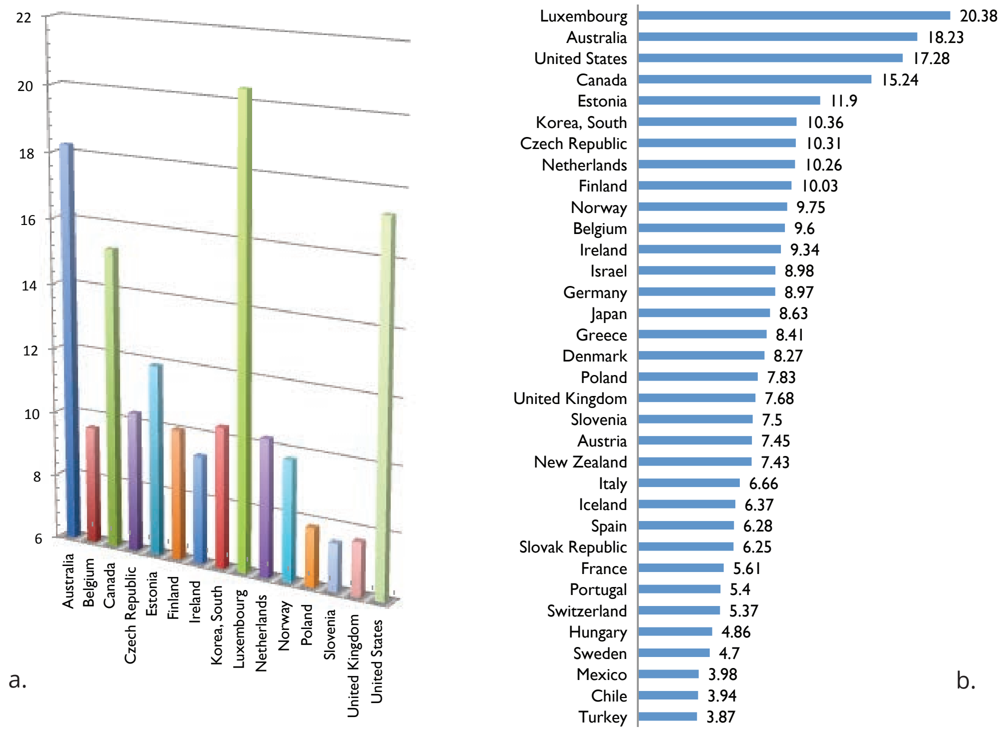


Figure 8.2: Two simple data graphics showing emissions of CO_2 (tonnes per capita) in some OECD countries

than twice as many countries as Figure 8.2a. The gridlines, ticks, box, shading, shadows, colors, and 3D effect in Figure 8.2a are all redundant: they can be removed without reducing the amount—and even increasing the legibility—of the data shown.

Data integrity Pictures can also lie. Any data graphic has the potential to distort data: to suppress or misrepresent facts and relationships in the data. The principle of *data integrity* is concerned with presenting the data itself, as far as possible, and avoiding embedding misleading messages in the data graphic.

data integrity

For example, closer examination of Figure 8.2 will reveal that the *y*-axis scale of Figure 8.2a begins at 6 tonnes of CO_2 per capita. This design choice tends to give the impression that countries like the United Kingdom have relatively low emissions, while in fact the UK is amongst the middle per capita CO_2 emitters in the OECD. The misleading impression is further advanced by the deliberate omission of the 15 countries with the lowest per-capita emissions.

Correspondence Data graphics are more than simple pictures that *show* something; data graphics should allow the viewer to compare, relate different items, and even infer causes and effects. The principle of *data correspondence* is connected with the aim of helping the viewer to understand patterns and form hypotheses about the relationships in the data.

data correspondence

Box 8.1: Data-ink ratio

A concrete example of the principle of data density is Tufte's *data-ink ratio*, in Tufte (2001). The data-ink ratio is the proportion of the ink used to print a graphic that is actually required to display the data, defined as:

$$\textit{Data-ink ratio} = \frac{\textit{Non-erasable data ink}}{\textit{Total ink required to print the graphic}}$$

Non-erasable data ink is ink that would lead directly to loss of data content in the image if removed (e.g., scales, shading, axis and grid lines, labels, and so forth). Although conceived in an era when most graphics were printed, one can easily imagine a translation of the ratio to the proportion of non-erasable pixels in a digital data graphic. Taking the data-ink ratio to extremes can lead to minimalist and even difficult to interpret data graphics. However, it remains an instructive exercise to try to maximize the data-ink ratio in your own data graphics.

In Figure 8.2b, for example, ordering the data by emissions, rather than alphabetically invites the viewer to identify high and low emissions, compare different countries, and contrast similar countries. The organization of data in Figure 8.2a makes identifying correspondences harder, such as identifying those countries with the highest per capita emissions.

aesthetics

Aesthetics Data graphics should be aesthetically pleasing. This principle of *aesthetics* is the hardest of our four principles of data graphics to assess objectively. It is possible you like the look of the data graphic in Figure 8.2a; I do not. However, aesthetics is not just about personal preference. Here again it is possible to identify and classify those features common to beautiful pictures and graphics. Evidence of care and attention to detail production; the use of alignment, centering, spacing; and proportion and balance in data graphics can combine to produce beautiful data graphics, in what Tufte (2001) referred to as "simplicity of design and complexity of data."

8.1.2 Visual variables

marks
visual variable

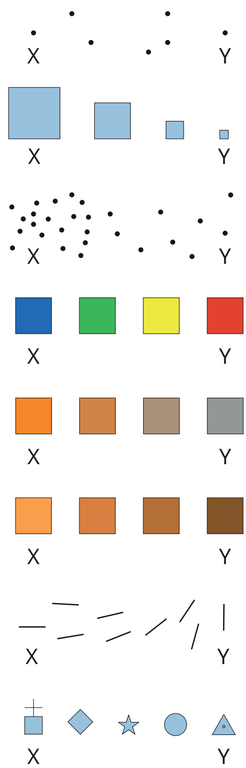
All data graphics, including maps, must be constructed from basic visual symbols, called *marks*. The different ways in which the characteristics of marks can be varied are termed *visual variables* or *graphic variables*. In this book we distinguish six visual variables: *position*, *size*, *pattern*, *color*, *orientation*, and *shape*, summarized in Table 8.1.

hue
value
saturation

Position concerns where the marks appear, both relative to other marks as well as in absolute terms—where "on the page" or the screen. Pattern can be decomposed into two elements: *texture* (density) and *arrangement* (e.g., ordered or random) of marks. Color is usually decomposed into three elements: *hue*, *saturation*, and *value*, known as the HSV model. Hue refers to the shade of a color, for example, the amount of "redness" or "greenness." Value refers to the brightness of a color: its lightness or darkness. Saturation refers to the purity of a color: how intense a color appears. Saturation is often used in combination with value, but it may be used independently to control the prominence or impression of certainty of symbols. Shape concerns our ability to distinguish marks based on their geometric shapes. The shape of marks can be *abstract* or *mimetic*, depending on whether they resemble the thing they represent (mimetic) or not (abstract). Orientation concerns the

Table 8.1: Visual variables

Visual variable	Example
Position	
Size	
Pattern (texture, focus, and arrangement)	
Color (hue)	
Color (saturation)	
Color (value)	
Orientation	
Shape	



the inherent directions in which a mark points for marks that possess strong rotational asymmetry (such as lines or arrows).

Interpreting the characteristics of marks, and the similarities and differences between marks, is the primary mechanism by which the viewer derives information from any data graphic. Our visual perception is capable of rapidly discerning fine distinctions between marks. Indeed, many distinctions can be detected by our visual system in a fraction of a second without even requiring our conscious attention. For example, you will immediately spot the "odd line out" in Figure 8.3a. Similarly, it is likely you are able to identify the odd line out in Figure 8.3b, even though the distinction between marks is fine (a difference in orientation of only 5 degrees). The characteristics of marks that "pop out" in our vision in this way are termed *preattentive features*, and include orientation, size, color hue and value, and many aspects of shape.

preattentive features

Other distinctions may require conscious attention to be discernible. For example, it will most likely take you much longer to identify the odd line out in Figure 8.3c. The combination of multiple distinguishing features of marks, such as color and orientation together, do not allow unique marks to "pop out" in the same way as single-featured marks.

Value message Certain visual variables can allow more than simple *difference* to be depicted; they may also allow *relative* values to be communicated. The

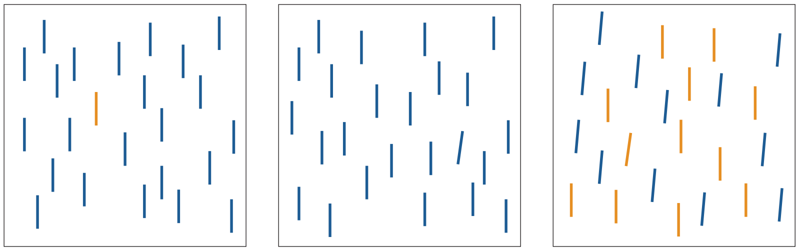


Figure 8.3: Preattentive features: Single and combined differences in visual variables

value message

association of a particular visual variable with greater or lesser relative value is referred to as the *value message* (sometimes called the *magnitude message*).

Some visual variables are best for communicating *ordinal* differences in value, allowing items to be ranked. For example, color saturation or value are especially useful for communicating relative ordering. Examining the color value example in Table 8.1, a natural interpretation is that there is more of something at Y than at X. The effect can be even more pronounced when combining changes in color value with reinforcing changes in saturation.

While color value and saturation can be used to communicate relative ordering, they are not especially good for communicating the *magnitude* of difference. A few visual variables can be used to communicate such *numerical* (i.e., interval and ratio, see Section 4.2.1) value messages—they can communicate not only whether there is *more* of something, but an impression of how *much* more there is.

Size is perhaps the best example of a visual variable associated with a numerical value message. Looking at the size example in Table 8.1, you can probably estimate how much more of something there would be at X than at Y. Most people's estimate of how *much* more is typically around 20–30 times more. In fact, the edge of the square at X is six times longer than that at Y, and so the area is 36 times larger. Humans tend to attach a value message in proportion to the length of lines, but in proportion to the area of areal symbols, such as squares (and in proportion to the volume of volumetric symbols, like cubes). However, the effects of such messages are somewhat muted because humans tend to consistently underestimate the relative size of areal and volumetric symbols, for example generally associating a value message of between 60–90% of the actual difference in area.

Table 8.2 provides a guide to the effectiveness of the six visual variables for communicating ordinal and numerical value messages, as well as their effectiveness for communicating simple difference without any value message (nominal level). Box 8.2 on the facing page also gives some background to another important principle in anticipating how people will interpret certain features of data graphics: Weber's law.

Semantics In addition to being used to distinguish different marks, variations in each visual variable are identified with common associations or

Box 8.2: Weber's law

Much of human perception is based on relative rather than absolute judgments. This principle is captured by the idea of a *just noticeable difference* (JND). The JND is the magnitude of the change in intensity of a stimulus (such as the length of a line) required such that the change can be perceived by a human. Perhaps surprisingly, numerous studies have shown that the JND depends not on the absolute size of the change, but rather on the size of the change relative to the original stimulus. Mathematically, if δI is the just noticeable change in the intensity of a stimulus I, then the proportion $\frac{\delta I}{I}$ is a constant! This means humans are able to perceive very small differences in very small stimuli, even though they may be unable to discern similar small differences in large stimuli. For example, most humans would be able to just discern apart a line of 10 mm length from a line that is 11 mm. However, discerning a millimeter difference in length in a line that is 1 m or even 10 cm is generally impossible for humans. Instead, one should expect it to require a line of 1.1 m or 11 cm, respectively, to ensure similar reliability in noticing the difference. Weber's law captures this general principle, that the perceived change in a stimulus is proportional to the intensity of the original stimulus.

	Numerical	Ordinal	Nominal
Position	Good	Good	Good
Size	Good	Good	Good
Pattern texture	Moderate	Moderate	Good
Orientation	Moderate	Moderate	Good
Color hue	Moderate	Moderate	Good
Color saturation	Moderate	Good	Poor
Color value	Moderate	Good	Poor
Shape	Poor	Poor	Good
Pattern arrangement	Poor	Poor	Moderate

Table 8.2: Effectiveness of visual variables for different measurement levels, after MacEachren (1995)

messages. For example, the color red is often associated with danger; green is for "go"; blue is associated with the sky or the sea. Shape is often used iconically: a square with a cross on top is conventionally used to indicate a church; a triangle with a dot in the middle is known by surveyors to symbolize a triangulation point (see Table 8.1).

a. Swiss flag b. the Red Cross c. first aid

Figure 8.4: Semantics and visual variables

The semantic associations with symbols can be strong. Figure 8.4 shows a cross symbol in three different hue combinations (red and white, white and red, and green and white). These marks will be immediately familiar to most readers as the Swiss flag, the Red Cross symbol, and the first aid symbol. Awareness of the semantic associations of marks is important because ignoring those associations can confound the intended interpretation of a data graphic. If you relabeled all the first aid kits in your building with the Swiss flag, there would be serious potential for confusion despite the fact that both are striking and immediately recognizable in an emergency.

Box 8.3: The Peutinger table

The image below shows an excerpt from the Peutinger table, a 14th century copy of a fifth century map (source: Wikimedia, a photographic reproduction of a work of art in the public domain). The excerpt is recognizably centered on Sicily and southern Italy, enclosed by parts of the Dalmation and north African coastlines. The map is designed to provide a faithful representation of road connections and distances, but not the shapes and areas of the land masses. The results are akin to today's schematic maps, commonly used for transportation networks.

8.2 *Cartography*

Cartography is the art, science, technology, and history of maps and map-making. While study of the design of data graphics is a relatively recent endeavor, the design principles that govern maps have evolved over centuries of map use. Mappiform (map-like) depictions of the world date back at least 3000 years. Some surviving ancient maps, such as the Babylonian Imago Mundi (circa 600BCE) and the Bedolina Map (circa 800BCE), are symbolic; others, such as the Turin Papyrus Map (circa 1150BCE) are clearly intended as accurate depictions of aspects of the geographic world (see also Box 8.3).

The key feature of a map, as opposed to a data graphic more generally, is that the relative positions of marks on the map are used to indicate some aspect of relative geographic location. Maps are frequently also associated with being amongst the highest data densities of any type of data graphic.

In the context of a GIS, it is worth highlighting an important distinction, not always made clear, between the *spatial* and *graphical* representation of objects. Spatial objects exist to directly model the application domain, while graphical objects are the presentational form of spatial objects. Traditional hardcopy maps conflate the data storage and data presentation functions of a GIS. A fundamental difference between maps and GIS, then, is that in a map the spatial and graphical representations of objects may be the same; in a GIS spatial and graphical representations are always separate. In the remainder of this section we are concerned solely with the graphical representations of objects in a GIS or on a map, but not the underlying representation of the spatial data itself, in the machine (already discussed extensively in Chapters 3–6).

8.2.1 Abstraction

Maps are effective tools for presenting geospatial data because they provide abstract graphical representations of the geographic world, highlighting important information and suppressing unimportant information. Map users are therefore able to focus on the salient relationships represented in the map without becoming distracted by irrelevant details.

It is worth dwelling a moment on this point: maps derive their usefulness and power from what they leave out, as much as from what they put in. The geographic world around us is so complex, the data we can derive from it so dense, that maps become useful precisely because they do *not* show everything.

Figure 8.5 shows an example of a topographic map alongside an aerial photograph of the same region, Orono, Maine in the US. A *topographic map* depicts information about the land surface, including relief and features like rivers, roads, buildings, and so forth. Topographic maps are amongst the most complex and sophisticated types of maps. Despite this complexity, the topographic map omits much of the detail contained in the aerial photograph.

topographic map

a. Aerial orthophoto

b. Topographic map

Figure 8.5: Abstraction in topographic maps (Source: USGS)

The map achieves this abstract form by reducing the detail of both the semantic and geometric information it depicts. Reduction in semantic detail is achieved through *classification*: selecting a small number of aggregate feature types to portray on the map. In the topographic map in Figure 8.5b, the colors indicate just four main classes of land cover: water (blue), woodland (green), built-up urban (gray), and everything else (white). Other classes of land cover that might be evident in aerial photographs, such as running tracks, yards and gardens, and swimming pools, are omitted in favor of the *selected* classes of interest. Further, those chosen classes of interest are themselves the product of *aggregation* of more detailed classes. For example, it would be possible to split woodland into, say, scrub land, mature woodland, and park and amenity woodland, but at the cost of increased map complexity. Finally, reduction in geometric detail is achieved by omitting many small buildings, individual trees, and the detailed shapes of the regions one might identify from the aerial photograph. This process of geometric adaptation is called *generalization*, discussed in more detail in Section 8.2.4.

Highlighting the important spatial objects and relationships, using classification and generalization, ensures the topographic map is much more useful than the aerial photograph for a wide range of tasks, such as urban planning, hiking, or simply walking around town. Abstraction is one feature of maps that distinguishes them from other data graphics. Irrespective of how good data capture and graphics technologies become, there will always be a role for the map even in a world filled with images.

8.2.2 Types of maps

We have already encountered examples of one of the most recognizable types of maps: topographic maps (e.g., Figure 8.1a and Figure 8.5b). While there is no consensus on a single classification of the types of maps, in this book we distinguish two more major classes of maps. There are myriad subclasses, some of which we shall briefly touch on.

thematic map

Thematic maps Thematic maps aim to present spatially varying data on some specific topic or aspect of a geographic area. Consequently, thematic maps have more in common with the examples of data graphics we have already encountered than topographic maps. Figure 8.6 shows an example of one of the simplest thematic map types: a proportional symbol map. The map shows the number and distribution of reported UFO sightings in Ontario.

Hopefully, given what we have already learned about data graphics, your design skills are already on alert. Do you think it is a good map or not, according to our principles of data graphics? Figure 8.6 offers moderate data density. Despite relatively low levels of redundancy (color and coastline vignette are arguably unnecessary), relatively few data points are shown, at least by the standards of many maps. The map does allow for some spatial correspondences to be identified. It seems immediately clear that UFO sightings are more clustered around the major cities, Toronto and Ottawa. The size of symbols is used to communicate the value message: larger symbols mean more UFO sightings, and the relative symbol sizes approximately accord with the relative numbers of sightings. However, here lies a crucial problem with the map: data integrity. The relative sizes of the major cities and the UFO sighting symbols have been deliberately chosen to exaggerate the importance of the UFO sightings. Further, the lengthy time period over which the sightings were made (many years) has been suppressed. The misleading effect is enhanced with the hue semantics: dangerous (red) aliens marauding across Ontario's verdant (green) countryside.

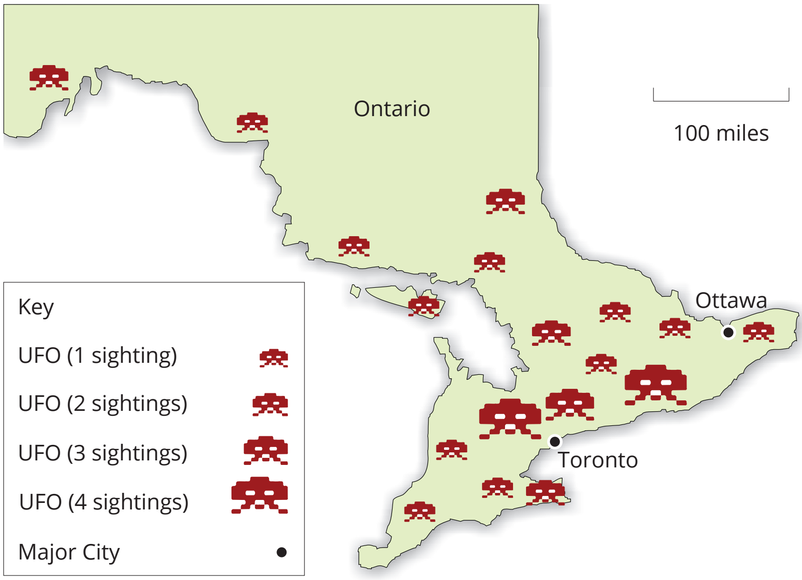


Figure 8.6: Proportional symbol map of UFO sightings in southern Ontario Province, Canada

An important lesson from Figure 8.6 is that, just like any data graphic, maps too can lie. Mark Monmonier's famous 1996 book *How to Lie with Maps* provides many examples of the manipulation and misrepresentation of information using maps.

One type of thematic map that is especially prone to mislead is the *choropleth map*. A choropleth map is a data graphic that presents data for defined geographic regions. Typically, choropleth maps use color or shading in depicting data values. For example, Figure 8.7 shows four related choropleth maps (a "small multiple" data graphic, see Box 8.4 on the following page) of the same data concerning carbon emissions in London boroughs. The maps visibly differ because of the different ways the underlying numerical data has been classified into five classes. The *equal interval* classification (Figure 8.7a) subdivides the entire range of the data into equally sized subranges. By contrast, a *quantile* classification (Figure 8.7b) generates classes that contain approximately equal numbers of members in the data. *Jenk's natural breaks* (Figure 8.7c) is a statistical procedure that identifies classes with high interclass but low intraclass variation. Finally, *mean-standard deviation* classification (Figure 8.7d) assumes the data is normally distributed, and it picks class breaks based on the mean and standard deviation of the data.

choropleth map

equal interval

quantile

Jenk's natural breaks

mean-standard deviation

Figure 8.7 is intended to highlight how the choices made by the map designer can communicate starkly different messages, even when showing the same underlying data. Some of the choropleth maps in Figure 8.7 give the impression of lower overall carbon emissions; some appear to highlight a concentration of high carbon emission regions in the north and west; some appear to single out particular regions as hotspots. Another important cause

Box 8.4: Small multiples

One technique for highlighting in correspondences in data graphics made popular by Tufte (1990) is the use of *small multiples*, also sometimes termed a trellis chart. A small multiple is constructed from an array of closely related data graphics, each usually using the same orientation, axes, scales, and symbology, but typically showing a different variable or alternative design. For example, looking ahead to Figure 8.7, the small multiple of four related maps of carbon emissions invites comparisons between the same regions in different maps, as well as between different regions in the same map. Aligning multiple related data graphics in small multiples is a powerful way to disaggregate complex data sets and help your audience identify relationships not easily discernible from a single graphic.

of misleading and mistaken maps is the so-called "modifiable areal unit problem" discussed further in Box 8.5 on page 321.

There are numerous other types of thematic maps. Figure 8.8 shows a famous example of a flow map: a type of thematic map that uses links to depicting flows between geographic locations. In the case of Figure 8.8, the flow map shows the exports of wine from France in 1864. Thicker lines provide the value message of greater exports. The map is carefully designed to avoid possible confusion from lines crossing each other or coastlines, and invite comparisons between countries and even continents.

cartogram

Finally, a *cartogram* is a type of thematic map that deliberately distorts geographic areas or distances in order to represent its thematic variable. Figure 8.9 shows a cartogram where the shape of each country is distorted in order that its area becomes in proportion to its population. The result is striking, highlighting in particular those densely populated "fat" countries (such as the UK, Japan) and sparsely populated "skinny" countries (such as Australia, Canada, and Russia).

Schematic maps Unlike the earlier maps in this section, the cartogram deliberately distorts geographic distances or areas in the pursuit of clearer communication of the map theme. A *schematic map* goes further still, suppressing metric geographic details in order to highlight *connectivity* between places. Because of the importance of connectivity, schematic maps are most familiar in the context of transportation maps.

schematic map

Figure 8.10 contains an example schematic map showing the Oslo metro network. The focus of such maps is the connections between stations. In common with most schematic transport maps, the lines are straight; the angles between are fixed to a small set of possibilities (in the case of Figure 8.10, restricted to only 90°); and the distances between stations are also generally fixed across the map, only lengthened for reasons of graphical clarity.

Schematic maps are excellent examples of the principle of abstraction. By deliberately omitting a wealth of metric details, schematic maps achieve much higher levels of usability for route planning than their planimetric counterparts. For contrast, Figure 8.11 shows the same Oslo metro network in planimetric view, with geographic distances and angles preserved. Even with

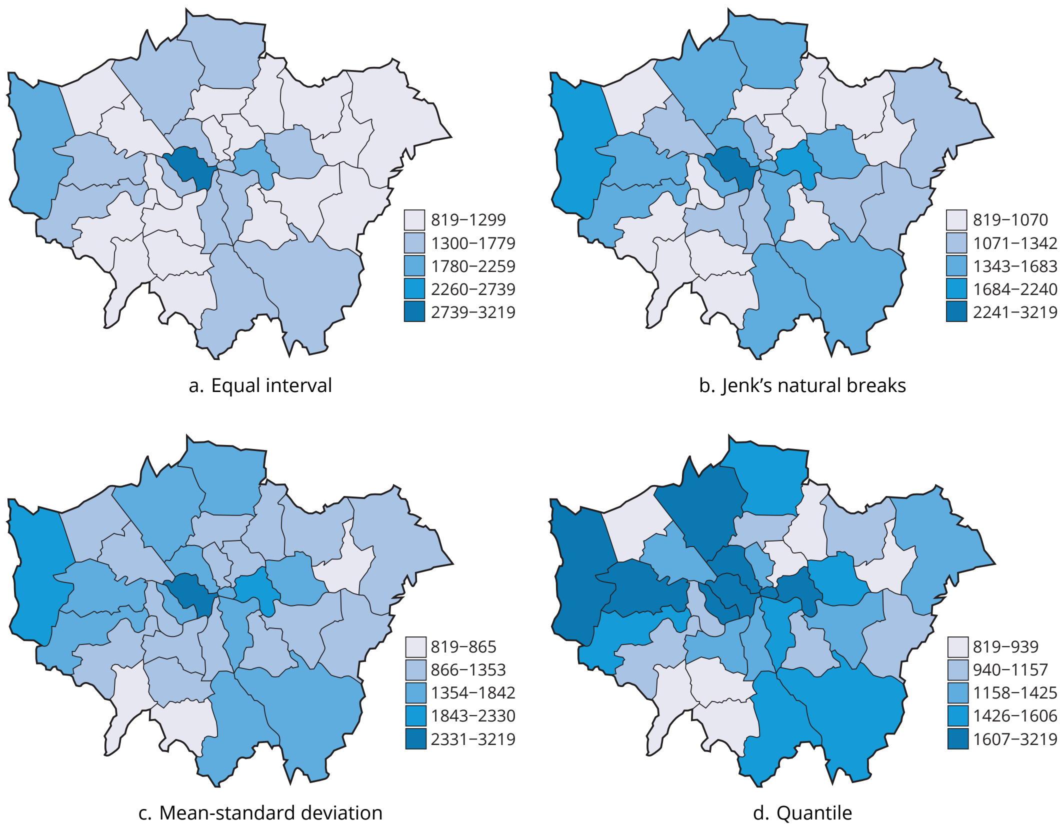


Figure 8.7: Choropleth maps of carbon emissions (kilotonnes of oil equivalent) in London, UK (Data source: london.gov.uk)

many stations and names omitted, the results would clearly in most situations be far less useful for getting around Oslo on public transport.

Schematic maps are sometimes also termed *topological maps*, because of their suppression of metric features in favor of the topological property of connectivity. However, it would not be true to say that schematic maps are purely topological. Comparing Figure 8.10 and Figure 8.11, it is clear that some metric properties are approximately preserved, such as the relative directions and broad shape of the network.

8.2.3 Map projections

Schematic maps and cartograms are examples of map types that deliberately distort metric features of geographic space in the service of clearer communication of data. But every map necessarily introduces some distortion of geographic space. This distortion arises because the Earth is not flat; maps are. It is impossible to project the surface of a sphere onto the plane without introducing some distortions. If you need to convince yourself of this, try to squash an orange peel onto a flat table surface: the results are always messy, requiring the peel to be stretched or ripped in some way.

Unlike the ad hoc and map-specific distortions of schematic maps and cartograms, study of the topic of *map projections* is concerned with the *systematic*

map projection

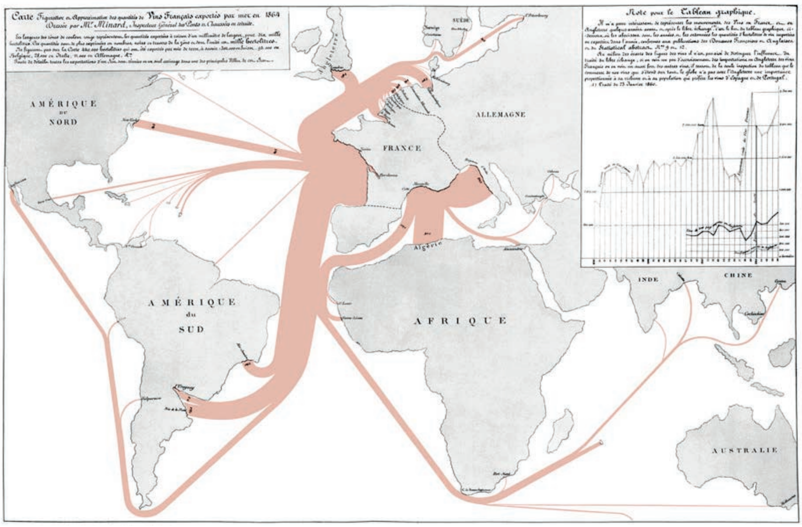


Figure 8.8: Minard's famous 1864 flow map of French wine exports (Source: Wikimedia, a photographic reproduction of a work of art in the public domain)

projective transformation of spherical coordinates (i.e., on the surface of the Earth) to planar coordinates (i.e., on a map sheet or a computer screen). Map projections were already encountered briefly in Section 3.2.3.

Figure 8.12 illustrates the three main types of map projection: projection onto a flat plane (azimuthal, Figure 8.12a); projection onto a cylinder (cylindrical, Figure 8.12b); and projection onto the curved surface of a cone (conical, Figure 8.12c).

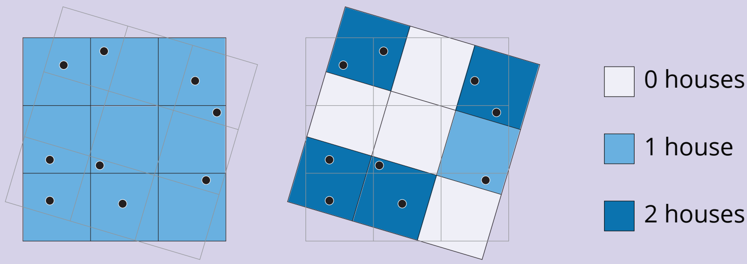
equidistant projection

Examples of each of the three types of map projection may be found in Figure 8.13. Figure 8.13a shows the globe projected using the Azimuthal Equidistant projection. A projection is termed *equidistant* if the projection preserves distances to some fixed point or line. In the case of the Azimuthal Equidistant, in Figure 8.13a, the distances from all mapped points to the map center (70°E, 20°N) are in proportion to one another. However, other geo-

Figure 8.9: Cartogram of world population by country (Copyright Mark Newman)

Box 8.5: MAUP: Modifiable Areal Unit Problem

Maps frequently aggregate information from across a region of space, either for reasons of clarity or because data is only collected or made available on an aggregated, per-region basis. Census data, for example, is commonly aggregated over arbitrary administrative regions in order to protect an individual's privacy. However, many other data sources, such as the environmental data in Figure 8.7, is frequently captured, communicated, or presented over aggregate regions. But what would such maps look like if we had chosen different regions over which to aggregate data? As an example, the configuration below illustrates how the appearance of a map may change dramatically with a relatively minor modification to the areal units. In the figure, the locations of households and classification scheme remain unchanged. However, a mere 16° rotation of the gridded regions leads to a remarkably different pattern in the apparent spatial distribution of households across the aggregate regions.



The *modifiable areal unit problem* (MAUP) summarizes this issue as follows: the results of any analysis upon spatial data that has been aggregated using arbitrary (i.e., "modifiable") regions (i.e., "areal units") is not guaranteed to be valid independent of those regions (i.e., "problem"). For example, in the illustration above, neither the apparently regular distribution of houses in the first grid nor the apparently irregular distribution in the second grid tells us anything about the actual underlying distribution of houses. A closely related problem to MAUP is the *ecological fallacy*: the mistaken idea that the properties of an aggregate region may be imputed to all the individual entities contained within that aggregate.

metric properties, including the shape and area of land masses, are distorted by the projection.

The Mercator map projection in Figure 8.13b exhibits different distortions. The Mercator projection is a cylindrical type of map projection. The Mercator projection does not preserve distances, and indeed it distorts the size and shape of larger objects. However, it does preserve angles and shapes locally, over small distances. Projections that preserve local angles are termed *conformal* projections. This property of conformality represented a breakthrough for maritime navigation in the 16th century, when the cartographer Mercator first introduced his eponymous projection.

conformal projection

The Albers Equal Area projection in Figure 8.13c is an example of a conical-type projection. As the name suggests, it is an *equal-area* projection. The Albers projection does not distort areas, with the sizes of countries and continents remaining in proportion to each other. Concomitantly, distortion will occur in shapes and distances using this projection.

equal-area projection

Map projections have a long history and have been topic of study and research in their own right. At the time of writing, there were more than 100 distinct projections documented and implemented in NASA's *G.Projector* software, a leading application for exploring different map projections. How-

Figure 8.10: Schematic map of Oslo metro network (Copyright Arsenikk)

ever, a more thorough investigation of map projections in this book would take us too far from our main themes. From the perspective of display and interaction, two messages are especially important to remember about map projections.

First, any 2D map must necessarily be subject to a map projection. Similarly, digital spatial data will either be supplied in spherical coordinates, in which case an appropriate map projection must be applied in order to display it on a 2D map; or spatial data will be supplied using projected coordinates, in which case the correct projection must be known and applied in order to relate that data to other data sets using coordinates in different projections.

Second, any map projection must necessarily introduce distortion into some aspects of the projected map, whether areas, distances, or shapes and angles. The principles of data integrity and correspondence demand that map projections are correctly applied to avoid unknowingly or misleadingly distorting maps.

8.2.4 *Map generalization*

map generalization

One further source of distortion in maps arises as a result of *map generalization*. Map generalization involves the distortion of geometric elements in a map in order to enhance its graphical clarity. The relationship between the size of a map and the size of the area that map depicts is captured using the

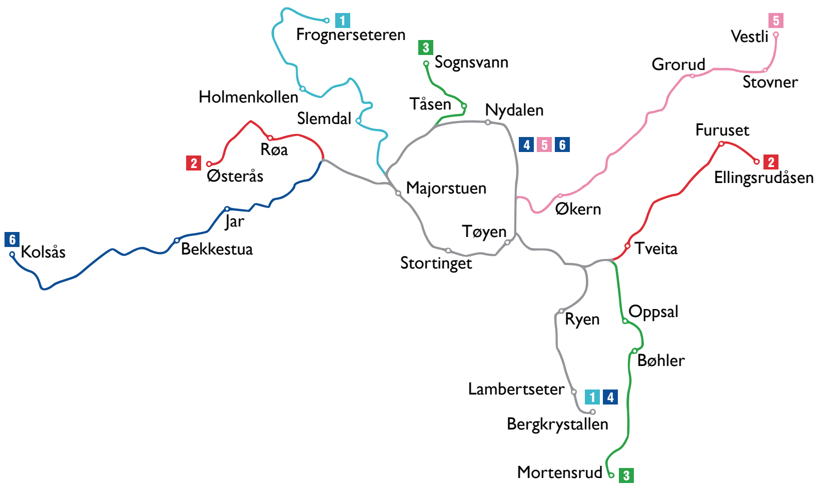


Figure 8.11: Planimetric map of Oslo metro network (Copyright Pneumaman)

representative fraction (see Box 8.6 on page 325). For example, a map where 25 m on the ground appears as 1 cm on a map has a representative fraction of 1:2500 (25 m = 2500 cm). In a map with a representative fraction of 1:63,360, each inch (2.54 cm) on the map represents a mile (approximately 1.6 km) on the ground (there are 63360 inches in a mile).

representative fraction

Maps must necessarily be much smaller than the real geographic spaces they portray. Consequently, distortions in the geometries of the features on the map are unavoidable if the map is to be legible. It is worth noting that because of the distortions introduced to at least some mapped distances by any map projection (even equidistant projections, see above), the representative fraction may vary across a map, especially a map with larger national or global extents.

There a many different map generalization procedures that have been used to enhance visual clarity. Here we categorize these into five distinct classes of map generalization operation:

- *Reduction*: Reducing geometric detail is perhaps the most fundamental class of map generalization operation. One example of reduction is *line simplification*. Recall that line simplification is the process of reducing

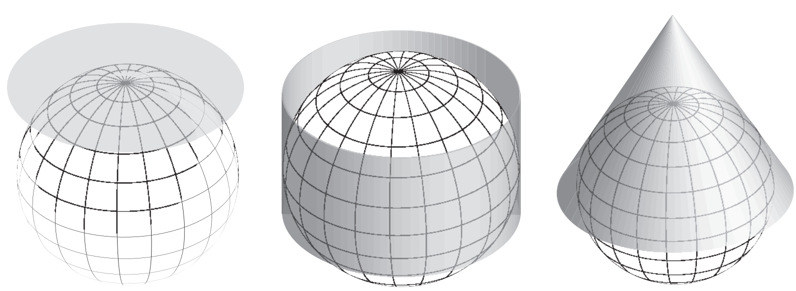


Figure 8.12: Types of map projection: azimuthal, cylindrical, and conical

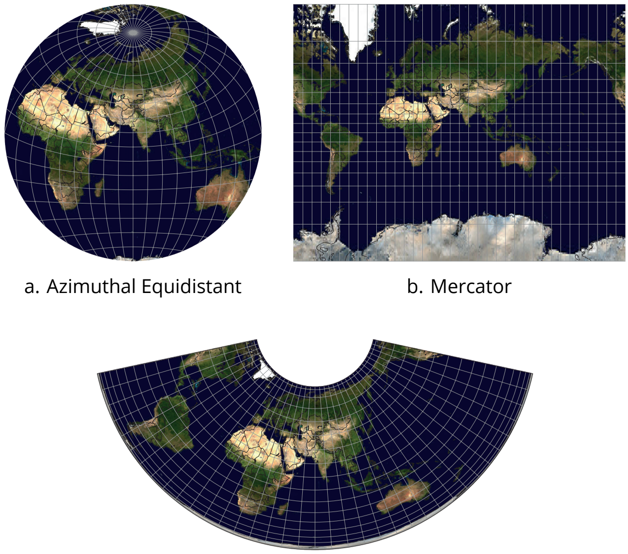

a. Azimuthal Equidistant
b. Mercator
c. Albers Equal Area

Figure 8.13: Example map projections

weeding

the level of detail in a polyline while retaining its essential character. The Douglas-Peucker algorithm is the most famous example of a line simplification algorithm (see Figure 5.7). Figure 8.14a shows another example of line simplification, termed *weeding*, resulting from removing all but each *n*th vertex in the polyline.

- *Merging*: Merging involves combining multiple distinct geometries into a single, clearer geometry. Figure 8.14b shows a group of buildings and the result of aggregation upon those buildings. Other types of merging operations found in the literature include *aggregation*, *agglomeration*, *amalgamation*, and *combination*.

displacement
enhancement
typification

- *Characterization*: Characterization involves changing the shape, size, or location of geometries in order to enhance clarity. Examples of characterization operations include *displacement* (moving geometries to ensure adequate visual separation between features, Figure 8.14c); *enhancement* (making geometries larger to ensure visibility, Figure 8.14d); and *typification* (changing or exaggerating the shape of a feature to enhance its symbolic or iconic characteristics, Figure 8.14e),
- *Collapse*: Collapse concerns the change in dimension of a mapped geometry, from 2D area to 1D line to 0D point. Figure 8.14f shows the runways at an airport collapsed from a polygonal area to two lines. Further collapse would result in the airport being marked by a point symbol.

elimination

- *Elimination*: Elimination involves the removal of whole geometric features in the pursuit of clarity (Figure 8.14g). Elimination is sometimes also referred to as *filtering*, *refinement*, or *selection*.

These five basic generalization operations are named for moving from more-detailed to less-detailed map features. For example, "simplification"

Box 8.6: Scale

The term "scale" is used in at least four distinct senses in geographic information science.

1. The term "scale" is frequently used as a synonym for "representative fraction": the ratio of the size of features on a map to the size of the same features in the world.
2. A related use of "scale" occurs in the phrases "large scale," "small scale," and so forth. In these cases, "small" and "large" refer to "small" and "large" *fractions*. Hence, a 1:1,000,000 map is usually termed "small scale" (small fraction), whereas a 1:1,000 map would be termed "large scale" (large fraction). Somewhat confusingly, this entails *large* denominators (the number of the right of the ":") equate to *small* scales; *small* denominators equate to *large* scales.

Together, these first two senses might more properly be termed "map scale."

3. Another sense of the term "scale" is as a synonym for relative (spatiotemporal) "extent." For example, the phrase "El Niño is a large-scale geographic phenomena" has no relation to maps; it is purely a statement about the spatial and temporal extents of El Niño (in implicit relation to other weather systems, say, "small-scale" rain storms). In natural language "scale" is often used in this sense, to refer to the relative size or extent of something.
4. Finally, and perhaps most importantly, "scale" is also used to refer to the dependence of phenomena upon level of detail. Many geographic phenomena and processes look different depending on whether they are being observed in more or less detail. For example, a topographic feature that looks like a "peak" (higher than its neighbors) at a fine level of detail may instead look like a "pit" (lower than its neighbors) as we "zoom out" to a coarser, more granular level of detail as neatly illustrated in Jo Wood's 1996 PhD thesis.

The key message is that the term "scale" should be approached with caution. Our advice is to avoid confusion by restricting your use of the term "scale" solely to its fourth and final sense above: dependence on level of detail. The first and third senses have ready synonyms to hand ("representative fraction" or "map scale", and "extent"), and the second sense is best avoided altogether, as it anyway frequently leads to confusion, even in discussions amongst experts.

suggests moving from more complex lines to ones with simpler geometries. Similarly, "collapse" suggests moving from higher dimension features to lower dimension features. However, it should be noted that there are a few occasions when specific generalization operations proceed in the opposite direction, generating more-detailed features from less-detailed features. Enhancement, for example, may in some cases be used to add more detail to an important feature to maintain clarity, even as the overall level of detail in a map may be being reduced (see, for example, the discussion of fractals in Section 3.4.3).

Naturally, it is also possible to combine multiple operations in a single map, or even in a single cartographic generalization process. Figure 8.15 shows two USGS maps of the same location in the docklands of Boston, Massachusetts. Looking at the map, one can see many examples of generalization, to assist with maintaining the clarity of the map when moving from the more-detailed 1:25000 map to the less-detailed 1:100000 map.

8.2.5 Map labels

One other essential feature of topographic maps in particular, but important on many other types of map, is the labels. Labels are used to identify particular instances of types of features (e.g., "Hospital," "Cemetery," or "Radio tower") or *toponyms*—names for places. Thus, the location of the label is con- toponym

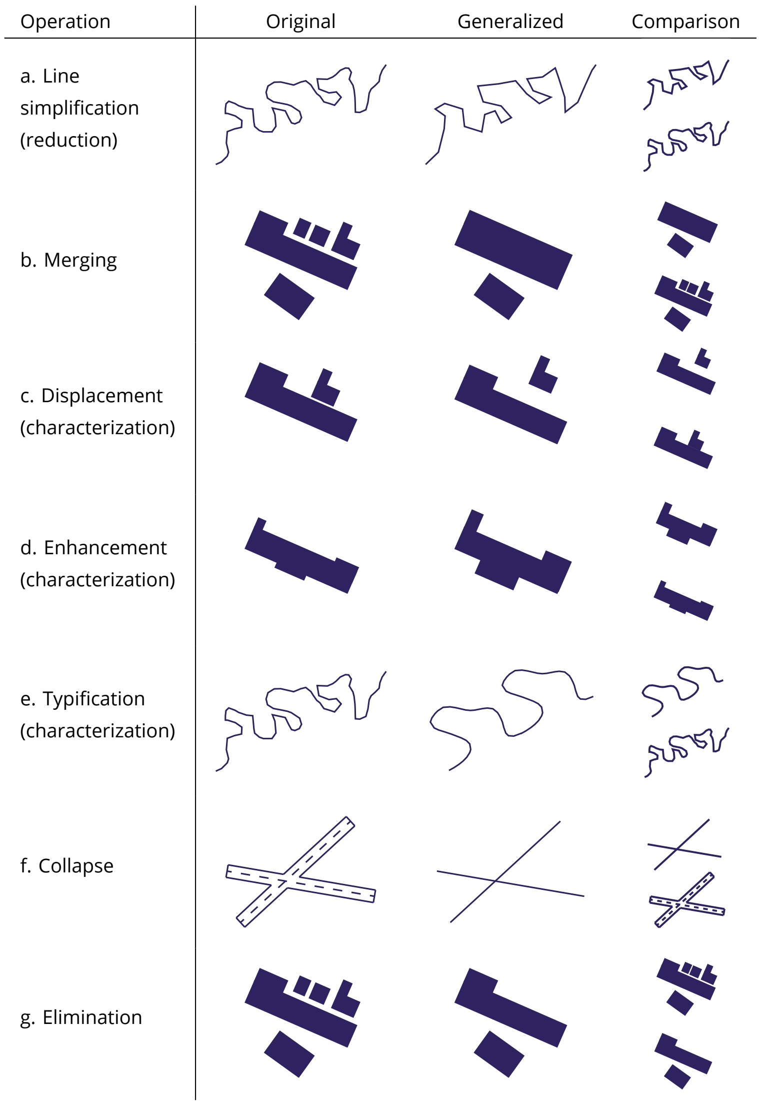


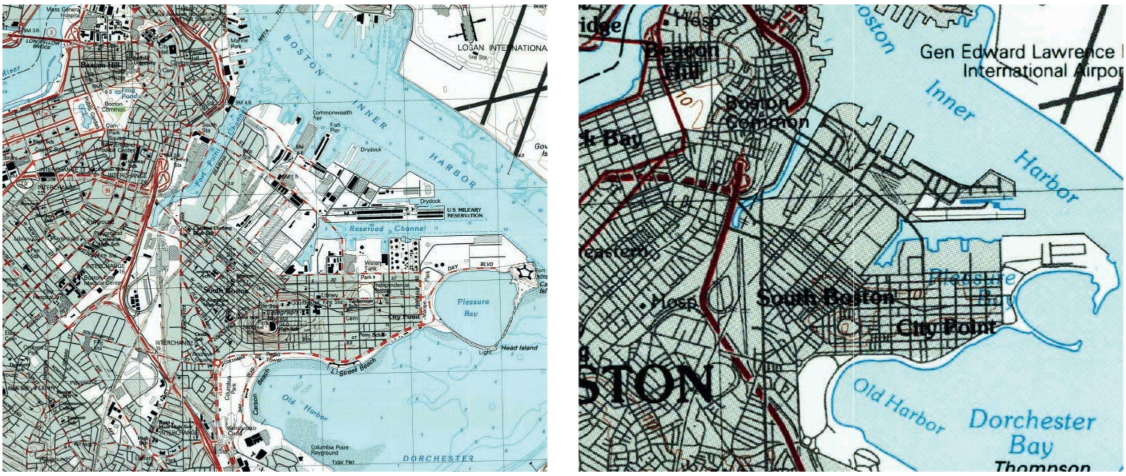
Figure 8.14: Map generalization operations

strained by the location of the feature or place it identifies. These constraints operate in three main ways:

- Point labels, where the label identifies a small point-like feature on the map, such as "Frog pond" in the northwest of Figure 8.15a or the "Light" on the harbor wall in the southeast. By convention, point labels are preferentially positioned slightly above and to the left of the point itself. However, other positions are also frequently used where other map symbols, labels, or geometries would interfere with preferred label position.

- Linear labels, where the name identifies a linear feature, such as a river or road. In Figure 8.15a, "Carson Beach" and "Street Beach" in the south of the map identify linear beach features. Linear labels typically run alongside the feature they label, preferentially conforming to the shape of the feature and avoiding backward sloping or upside-down text.

- Areal labels, where the name labels areas that are larger than can be handled with a simple point label. Areal labels are typically placed inside the areal feature itself, with the type spaced so as to provide the visual impression of "filling" the area. For example, "Pleasure Bay" and "Old Harbor" in the southeast of Figure 8.15a are toponyms that designate areas.



a. 1:25,000 map

b. 1:100,000 map

Figure 8.15: Example maps of Boston, Massachusetts, at different representative fractions (map scales) (Source: U.S. Geological Survey)

Some of the problems facing map labeling are similar to those facing the presentation of geometry or symbols. The labels must be legible and distinguishable, and positioned and sized such that each label is neither obscured by nor obscures other map features. When positioning a large number of labels, such as on a topographic map, no single, perfect solution will exist. Instead, clear and readable results require many compromises be struck between the positions and sizes of labels, symbols, and geometries across the entire map. Just as maps in general rely on selection, suppressing less-important features, so labels must be selected to identify only those objects and toponyms of greatest importance (cf. Figure 8.15a and b).

Other problems of map labeling are specific to the characteristics of text. The choice of type, for example, will have important impacts on the legibility and semantics associated with labels. In topographic maps, for example, the weight, size, width, slant, and typeface (e.g., Times versus Arial typefaces) are all carefully chosen to reflect categorical and hierarchical distinctions in the features labeled. Larger or more weighty type will carry a value message, such that features identified with bold type are more important than those identified with regular or lightweight type, for example. The use of capital letters similarly conveys a value message (e.g., "RIVER" versus "Stream"). Other aspects of shape provide both visual and semantic distinctions in type. For example, serifs (small lines attached to the ends of strokes in a letter, such as this text) tend to have connotations of heritage or formality. Sans serif fonts (used in figures and captions in this book) have connotations of modernity, and they tend to be more legible on maps (compare "Castle" and "Avenue"). The overall map style and cartographic identity will be strongly influenced by the choice of type. Most national mapping agencies have their own distinct and immediately recognizable type styles and conventions.

8.3 Interaction

Traditional maps and data graphics are mechanisms primarily for *presenting* information, for example, via digital display devices or hard copy maps. However, today's technology provides near-ubiquitous opportunities for *interacting* with data graphics and maps in ways not possible at the advent of GIS.

human-computer interaction

The term "human-computer interaction" (HCI) was first used during the 1980s to describe the interaction between computer systems and people. The term is also used to refer to the field of HCI: the study of the design, evaluation, and implementation of the interfaces between computing devices and people. Figure 8.16 summarizes the key components of HCI: the human, the computer, and the interaction between them.

Figure 8.16: Components of human-computer interaction

dialog

A *dialog* is a process of interaction between two or more agents, whereby agents cooperate to resolve conflicts and complete some task. For example, humans commonly use dialog to communicate with each other. In a human conversation each interlocutor will listen to the other, responding to what has previously been said. The process of interaction between a computer and its human user also normally takes the form of a dialog, although HCI dialogs usually have a much more rigid structure than human conversations.

user interface

The dialog between humans and computers is mediated by a *user interface*. A user interface in HCI is the mechanism by which a human user accesses the functions offered by a computer.

8.3.1 Input and output

A plethora of devices exists to enable humans to interface with computers: keyboards, mice, trackpads, microphones, monitors, and speakers are amongst the most familiar. In order to make sense of all these devices, we must first note that humans and computers are able to send and receive information in several different modes, termed *input-output* (IO) channels. Computers output on IO channels that are detectable by human senses, such as sight, hearing, and touch.

IO channel

Humans have five senses that receive information from outside the body, termed exteroceptors: *sight* (visual sense), *hearing* (auditory sense), *touch* (termed "haptic", meaning related to touch), *smell* (olfactory sense), and *taste* (gustatory sense). A fire alarm, for example, provides us with auditory information about a dangerous fire, ideally long before we receive similar information with our visual or olfactory senses. Humans have two other senses relating to the movements of the body, termed *proprioceptors*: balance (*vestibular* sense) and *kinesthesia*, the ability to sense our own bodily movements and tensions.

exteroceptor

haptic

proprioceptors

Humans can input information to the computer on these same IO channels. For example, your computer receives information from you primarily on the haptic IO channel, usually via the keyboard, mouse, or touchscreen. In return, information is sent from your computer to you primarily on the visual IO channel, via its display screen, or perhaps on the auditory IO channel via its speakers. Visual information is the most commonly used sense in user interfaces devices, such as through computer monitors and screens. Auditory and haptic information often plays a secondary role in user interfaces, reinforcing visual information, for example, through speakers, microphones, keyboards, touchscreens, and vibrating tactile devices (but see Box 8.8 on page 332). Olfactory and gustatory senses play the least significant role in user interfaces. Proprioception is also used in human user interface devices, although normally in support of information on other IO channels, especially visual IO. For example, flight simulators can be mounted on a mobile base that mimic the movements of a real aircraft, stimulating the vestibular sense. Figure 8.17 summarizes all these concepts and terms.

The same IO channel can often be used for sending information in both directions—from computer to human and from human to computer—at the same time. For example, auditory information can be sent via a computer speaker and received by human hearing, or sent by the human voice and received by a computer microphone. To avoid confusion, the term *display* is often used to refer to output from a computer to a human for any IO channel, while *input* is normally reserved for input to a computer from a human (Figure 8.17).

display

input

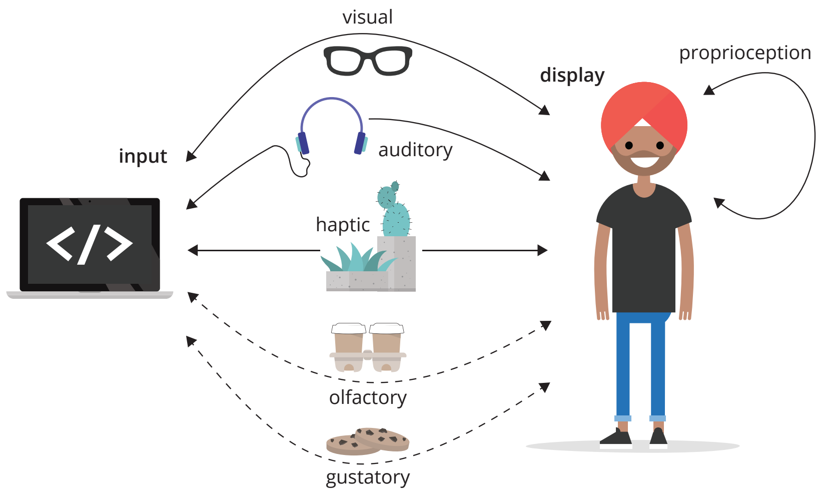


Figure 8.17: IO channels and human senses

explicit input

implicit input

A distinction is often made between *implicit* and *explicit* input. By default, user input is often *explicit*, with input entered consciously by the user via devices such as a keyboard or mouse. By contrast, *implicit* input occurs where the system automatically interprets information about a user's context or actions as input. For example, context-aware systems, such as LBS (introduced in the previous chapter), frequently make use of implicit input. A traveler wishing to find out about bus arrival and departure times might explicitly input their query to a transport authority website via their keyboard before leaving home. Alternatively, a location-aware mobile device could automatically interpret the user's location near a bus stop as an implicit query for information about the time of the next bus at that stop.

multimodal

Some devices are capable of simultaneous input and display, frequently on different IO channels. Touchscreens, for example, receive input on the haptic IO channel at the same time as displaying information on the visual IO channel. Systems that enable display or input of information on more than one IO channel (mode) at a time are termed *multimodal* (Box 8.7 on the facing page).

Irrespective of which IO channel a device uses, it must convert information between digital and analog formats. Humans are organisms that send and receive information in continuously varying analog formats; computers are digital machines that require discrete digital information to operate. For example, a computer mouse or keyboard converts the continuous analog movements of our fingers, hands, and arms into a discrete digital format that can be used by the computer. A speaker converts discrete digital audio information stored or generated by a PC into audible analog sound waves.

8.3.2 Intuitive and expressive interfaces

intuitive interface

Two important characteristics of any user interface are how *intuitive* and how *expressive* is the interaction it supports. An *intuitive* interface is easy to use,

Box 8.7: Multimodal display

Most GIS and mapping interfaces rely heavily on visual IO as the primary display mode. There are many situations in which visual display—or visual display alone—does not perform well in an interface. Haptic display, for example, is familiar in our vibrating mobile phones and computer game controllers, as well as used in some more specialized domains, such as Braille displays and embossers for the visually impaired. Using multiple IO channels, such as vision, haptics, and sound together for multimodal display carries three main advantages. First, multimodal interfaces may be accessible to a wider variety of users than unimodal interfaces, including visually or hearing-impaired users. Golledge, Rice, & Jacobson (2006) provide an overview of multimodality, and accessibility more broadly, in the context of GIS. Second, multimodal interfaces are able to operate in a wider variety of conditions than is possible with a unimodal interface (see Oviatt, 1999). For example, in vehicle navigation a visual map interface may be essential for certain situations, but when driving it would be dangerous to distract the driver's visual attention from the surroundings. Doyle, Bertolotto, & Wilson (2008) evaluated a multimodal GIS interface in connection with interaction errors and robustness. Finally, multimodal interfaces enable users to access complementary information on different IO channels concurrently. Experiments have demonstrated that such multimodal interfaces are preferred by users for more complex tasks, and they can lead to increased efficiency as well as decreased errors in user interaction, for example, as in Oviatt, Coulston, & Lunsford (2004).

requiring minimal effort to learn to operate. An *expressive* interface enables users to achieve specific tasks efficiently. Expressive interfaces allow users to specify commands more precisely, using a range of options that modify a command's behavior. The ideal interface style is both expressive and intuitive. Unfortunately, it is not usually possible to achieve both these goals. Different interface styles must strike a balance between being expressive and intuitive.

expressive interface

For example, a *command entry* interface allows a human user to issue commands to the computer directly. Common command entry interfaces include the R statistical language, the terminal in OSX and Linux operating systems, or the command prompt in Windows. Command entry interfaces are expressive because they typically offer many different options to modify the precise behavior of commands. However, such interfaces are rarely intuitive, because the human user needs to learn, master, and remember large numbers of commands and command options. At the Windows command prompt, for example, the command "`dir`" will result in the contents of the current directory being listed, while the command "`dir /a:d /p`" will list only the subdirectories of the current directory one page at a time.

command entry

By contrast, natural language interfaces enable computers to interact using written or spoken language. Natural language interfaces (such as Siri, Alexa, and Google Assistant) can be highly intuitive once some basic interaction principles have been understood (such as the need to employ the correct "wake word" to initiate interaction). However, natural language interfaces offer low expressivity in their limited ability to correctly interpret highly specific or complex commands. In addition to the many technical challenges in parsing human speech, human natural language is highly ambiguous. For example, the sentences "Time flies like an arrow" and "Fruit flies like a banana" are at first glance structurally similar. Only once the meaning of each sentence is grasped does it become clear that the two sentences are ambiguous and can assume quite different structures depending on the interpretation.

Box 8.8: Sound display

While most GIS and map interfaces rely heavily on visual display, sound can still play an important role in GIS and its applications. The process of using sound to represent data is termed *sonification*. *Sound symbols* are abstract sounds used to represent and distinguish information. For example, most people will be familiar with the Geiger counter often seen in movies. The Geiger counter sonifies the levels of ionizing radiation, with more frequent, louder, and higher pitch clicks indicating increasing levels of radiation. In fact, there are a variety of *sound variables* that may be used in sound symbolization, including loudness and pitch. Krygier (1994) provides an overview of sound symbolization for geographic information and enumerates the key sound variables. Sound is inherently dynamic, so many of the dynamic visual variables have counterpart sound variables, including duration, order, frequency (in this context meaning the periodic recurrence of sounds), and rate of change. An example of the use of symbolic sounds in a GIS interface to represent the reliability of geospatial information can be found in Bearman & Fisher (2012). Sound symbols can also be used metaphorically, just like visual symbols. Sound icons, often termed *earcons* (Gaver, 1986), are sounds that provide a metaphor for a command or function. For example, in desktop computing, the action of deleting a file is often accompanied by the sound of paper being scrunched up as if it were being discarded in a waste paper bin. Spoken language too is an increasingly common feature of many intuitive interfaces (see Section 8.3.2). Our sense of hearing also allows us to locate the source of sounds in 3D space. Slight discrepancies between the sounds heard in each ear enable us to determine the direction of a sound source, analogous to retinal disparity in vision. Other audio cues, like loudness, help to determine distance. Many digital sound systems take advantage of this facility and produce sounds that appear to be located in space. However, the precision with which humans are able to determine position in space, termed *spatial acuity*, is lower for hearing than for vision.

In the most commonsense interpretation, “flies” acts as a verb in the first sentence and a noun in the second sentence, while “like” acts as an adverb in the first sentence and a verb in the second sentence.

Figure 8.18 summarizes the balance between intuitive and expressive interfaces with some familiar examples. One of the most familiar examples is the *WIMP*—windows, icons, menus, and pointers—interface found on most personal computers today (see Box 8.9 on the next page).

WIMP

8.3.3 Designing intuitive interfaces

We have seen that interfaces must strike a balance between intuitiveness and expressiveness. More expressive interfaces can typically only be achieved by decreasing the level of interface intuitiveness. However, the converse is not the case: decreasing the expressiveness of an interface will not, on its own, necessarily lead to a more intuitive interface.

An important question, then, is how do we design interfaces that are more intuitive? This question is not as easy to answer as one might at first think. Even for the most familiar and everyday physical objects, such as doors and light switches, poor interface design is all too common. The bathroom tap in Figure 8.19, found in Canberra Airport, is sufficiently unintuitive that it is accompanied by four lines of operating instructions. If everyday objects such as bathroom faucets—invented thousands of years ago in Roman times—can still confound interface designers today, how much more challenging may we expect GIS and digital mapping interfaces to be?

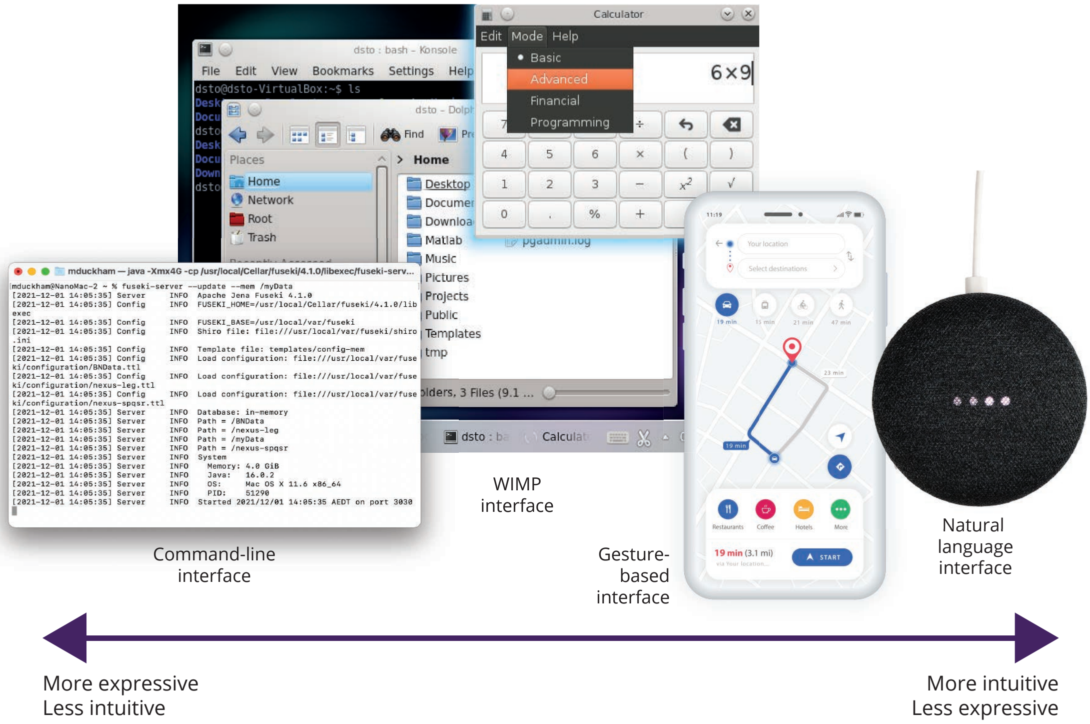


Figure 8.18: Intuitive versus expressive user interfaces

Hence, the question of what features of an interface lead it to be more or less intuitive lies at the core of HCI. In his classic book, *The Design of Everyday Things*, Don Norman sets out his influential approach to analyzing those features of any user interface—not just computer interfaces—that make them intuitive (Norman, 1988). This approach can be summarized into five fundamental principles of intuitive interfaces, summarized in Figure 8.20: visibility and affordances, mappings and metaphors, simplicity, feedback, and error handling.

Visibility and affordances Intuitive interfaces are "self-describing": simply by their form one should immediately see the main components and gain an

Box 8.9: WIMP interfaces

Originally developed in the 1970s, WIMP remains the most common interface style for todays desktop and laptop computers. WIMP stands for *windows*, *icons*, *menus*, and *pointers*. Windows are used as independent containers for particular processes or applications. Icons are small pictures that symbolize a command or function, such as a pair of binoculars or magnifying glass icon to symbolize a "search" command. Menus organize commands into logical groups and hierarchies (submenus) that are easier for users to access and remember. Finally, pointers enable users to directly manipulate objects, such as opening a file by clicking on it, or deleting a file by dragging it to a trash icon. WIMP interfaces have been successful because they are moderately intuitive at the same time as being moderately expressive. While WIMP is still common today, many devices offer other, more intuitive direct manipulation interaction styles, such as gesture-based interaction familiar in mobile devices such as phones and tablets.

Figure 8.19: A bathroom faucet (tap) as an interface. The instruction reads: "Sensor tap: This tap is 'hands free,' operated by sensor. To activate, please wave hand in front of the sensor, located to the right of the spout below"

intuition about what the interface does and how the components operate. *Visibility* is connected with ensuring that an interface makes the most important components most visible. *Affordances* on the other hand provide the user with clues about how to operate the components. In the physical world, chairs are for ("afford") sitting on, ladders are for climbing, handles are for grasping. However, even in the world of computer interfaces similar affordances can be found: buttons are for pushing, boxes are for checking, and sliders are for sliding.

visibility

affordance

The red button in Figure 8.20 is both highly visible and has the right affordances for pressing. By contrast, the mechanism for operating the faucet in Figure 8.19 is completely hidden from view and affords little aside from the incorrect action of grabbing the long spout.

Mappings and metaphors Users of an interface are not empty vessels. Users bring to any interaction a wide range of expectations drawn from experiences of other interfaces and of the physical world around them. Thus, intuitive interfaces should be compatible with such common experiences and harmonize with the mental models users bring to the interaction.

mappings

Mappings concern the relationship between interface controls and the physically or conventionally expected effects of those controls. For example, when viewing a map, pressing the down cursor or swiping up on the touchscreen might be expected to display an area to the south of the current view. Using a less obvious mapping (such as the keyboard "s" key) or even a counterintuitive mapping (such as the left cursor key or a pinch gesture) is guaranteed to make the interface less intuitive.

The control knob in Figure 8.20, for example, accords with a common mapping: turning clockwise to increase, counterclockwise to decrease. Again, the faucet in Figure 8.19 provides much less intuitive mappings, relying instead on a hand wave to start the flow of water.

Figure 8.20: Five fundamental principles of intuitive interfaces

An important mechanism for supporting mappings is the use of *metaphors*. In language, a *metaphor* is the use of a word or phrase denoting one kind of idea or object in place of another, in order to make a figurative comparison between the two (see, for example, Box 4.6 on page 158). Metaphors are everywhere and are widely used to help humans understand an unfamiliar target domain with reference to some familiar source domain. The ingrained nature of metaphoric understanding in humans makes metaphors a powerful vehicle for human-computer interaction. For example, your personal computer relies extensively on the *desktop metaphor*. The desktop metaphor suggests a likeness between the computer interface and an office desktop, with windows taking the place of sheets of paper, calculators, clocks, and so forth on a physical office desktop. Using the desktop metaphor with a WIMP interface (Box 8.9 on page 333) helps make coordinating and switching between different software applications more intuitive.

metaphor

Indeed, many GIS interfaces are built around the use of the *map metaphor*: they are designed to exhibit map-like characteristics, utilize map symbolization, and echo the physical (static) characteristics of hardcopy paper maps. Maps and map symbolization are immediately familiar to many people. Consequently, adopting the map metaphor usually leads to intuitive GIS interfaces, at least for those people who are already comfortable with conventional maps (see Figure 8.21).

map metaphor

Simplicity Intuitive interfaces are simple. One characteristic of simplicity is providing only those components in an interface that are strictly necessary to the functions of a system, and providing only those functions that are strictly necessary to anticipated user needs. The term *feature cascade* is sometimes used

feature cascade

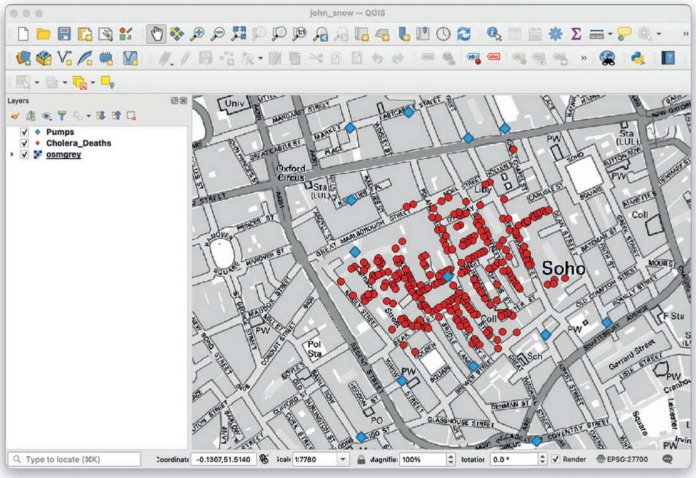


Figure 8.21: The map metaphor is common in GIS interfaces

to describe the undesirable tendency of more and more features to be added to interfaces, making them more and more complex and unwieldy.

The instructions accompanying the faucet in Figure 8.19 reveal an underlying complexity beyond what should be necessary simply to start and stop the flow of water ("... *wave* hand *in front* of the sensor, located *to the right* of the spout *below*"). While the control panel in Figure 8.20 is much less simple than the individual buttons, it also gives access to a much wider range of (labeled) actions.

feedback

Feedback Feedback is the process of accepting and responding to a user's actions with information about what a user has done, and what has been achieved. Feedback can be output on any IO channel, not only the IO channel used for the action. A visual confirmation popup, a satisfying sound, or a reassuring haptic buzz can all be used individually or even in combination to provide immediate feedback in response to a user's action, for example. The elevator button in Figure 8.22 (and Figure 8.20) immediately lights up (visual IO channel) in response to being pressed (haptic IO channel). A lack of clear feedback can be most frustrating, as will be familiar to any one with experience of ineffectually pressing an elevator button with a faulty light or fruitlessly attempting to activate a faucet sensor, such as in Figure 8.19.

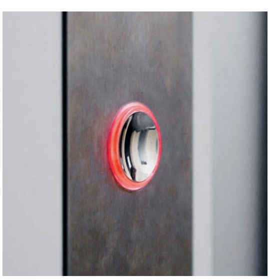


Figure 8.22: Lift button, from Figure 8.20

To ensure the user's reasoning processes can continue uninterrupted, feedback needs to be rapid. As a rule of thumb, a response time of about 0.1 seconds or less is perceived by users as instantaneous. A response time of 1–2 seconds is still fast enough for a user's flow of thought to be uninterrupted but is a noticeable delay lessening the impression of interactivity. A response time of 10 seconds is about the limit for feedback, as a user's attention is likely to be transferred to other tasks if delays exceed this time.

Error handling Computers are machines and do only and exactly what they are instructed. Humans, however, are fallible and will make errors. Intuitive interfaces must allow, account for, and even anticipate mistakes by users. The emergency alarm button in Figure 8.23 (and Figure 8.20), for example, has an inbuilt reset function (twisting the activated button) to reset a mistaken activation.

Figure 8.23: Emergency alarm, from Figure 8.20

Forgiveness is the principle that things that are easy to do should be easy to undo. Conversely, things that are hard to undo should necessarily be harder to do in the first place. For example, an interface that allowed the irreversible deletion of all stored data from a computer's memory with a single button press would fail the forgiveness principle. In a similar vein, mistyping a word into a document is easy to do, and so it should be easy to undo (for example, using the delete key or an easily accessible "undo" menu item, gesture, or keystroke).

forgiveness

Mistakes can also be an opportunity for the user to learn about a system. Intuitive interfaces do not penalize or sanction users when they make mistakes, but instead they gently help users understand what to do differently next time.

8.3.4 Usability engineering

The process of developing interfaces that maximize usability is termed *usability engineering*. In general, all usability engineering techniques emphasize the importance of considering usability at every stage of system development. Developing usable interfaces is not an exact science: experience and judgment play a vital role in successful interface development. Most importantly, interface and interaction design is an iterative process, requiring continual cycles of prototyping, testing and critical evaluation, and redesign (see Box 8.10 on the next page).

usability engineering

The usability of an existing interface can be assessed empirically by evaluating its performance either in the laboratory or in actual use in the field. Interface usability may sometimes be quantified, especially where objective evidence is required for upgrading or replacing an existing interface. Five basic criteria are often used to quantify the degree of intuitiveness and expressiveness of an interface.

Time to learn How much time does it take for users to learn how to use a system?

Speed of performance How quickly can users carry out benchmark tasks?

Rate of errors How many and what type of errors do users make in carrying out benchmark tasks?

Retention over time How well do users retain their skills and knowledge over time?

Subjective satisfaction How much did users enjoy or dislike performing benchmark tasks?

In addition to empirical user evaluation, interface designers often use more qualitative analytical techniques to ensure usable interfaces. Box 8.10 on the following page briefly discusses two of the most common approaches to

Box 8.10: Design methodologies

A effective and easy-to-learn method to help novice interface designers develop innovative data graphics and information visualizations is the five design-sheet (FdS) methodology of Roberts, Headleand, & Ritsos (2015). The FdS method begins with a "brainstorm" sheet, where the designer focuses simply on quantity of designs, aiming to develop as many possible options as possible. Sheets 2–4 develop three promising but completely different design options in greater detail. Finally, the fifth sheet combines, refines, and realizes a single preferred design that aims to take advantage of the best balance of features of all the options explored. Another common interface design methodology is GOMS. GOMS recursively decomposes high-level interface functionality into smaller component functionality. In a GOMS model, the user *goals* are specified in terms of the basic *operations* a user can perform. GOMS also models the different *methods* for achieving a particular goal and attempts to predict which particular method will be *selected* and in which cases. Taking a design approach to problem-solving more broadly—not only interface design problems—is often referred to as "design thinking." Design thinking is usually explained with reference to five stages: first, *empathize* with the humans whose problem you are trying to solve; second, *define* a precise problem statement based on the information gathered in the first stage; third, *ideate*—generate ideas and look for alternative ways of approaching the problem (as for FdS "brainstorming"); fourth, *prototype* and experiment with scaled down versions of the ideas; and fifth, look for ways to *test* these solutions out. The design thinking process has direct analogy to the stages of the modeling process introduced in Section 1.3 (see Figure 1.15) and the five main stages of the system development lifecycle (summarized in Box 1.3 on page 19: requirements, analysis, design, implementation, and validation). Like modeling and system development, the design thinking process is never linear, with many cycles of iteration essential to success.

interface design—the five design-sheet method and goals, operations, methods, and selections (GOMS)—as well as to problem solving with design more broadly, often called "design thinking." In addition to the iterative nature of any design process, *prototyping* is one of the simplest but most important usability engineering techniques. Prototyping involves developing a "mock-up" (prototype) of the target interface in order to reveal problems that may be hard to detect with pen-and-paper designs. Prototypes may be used simply to test design ideas, termed a *throw-away* prototype, or developed as a preliminary version of the actual interface, termed an *evolutionary* prototype. Prototyping is the most informal and low-cost usability engineering technique and can be valuable for even the smallest projects.

prototyping

8.4 Geovisualization

Geovisualization (also termed *geographic visualization*) is the process of using computer systems to gain insight into and understanding of geospatial information. Experimental evidence indicates that humans use two fundamentally different types of information in thought processes: verbal and visual. *Verbal thinking* is important for reading and writing, engaging in conversation, and logical thought. *Visual thinking* is important for reasoning about groupings, parts, and spatial configurations of objects.

geovisualization

verbal thinking

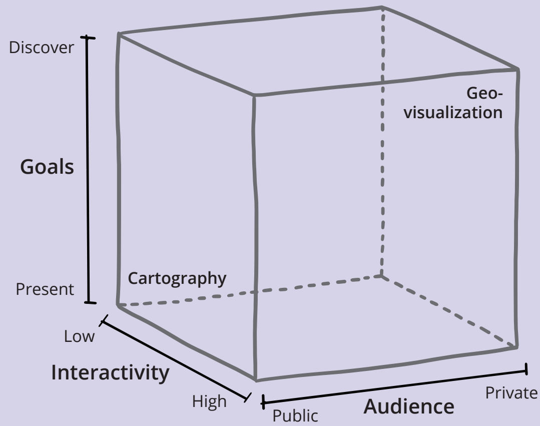
visual thinking

Scientific visualization is the process of using information systems to represent and interact with information in a way that enhances visual thinking. It is important to emphasize that the "visual" in visualization refers to "visual

scientific visualization

Box 8.11: The cartography cube

The *cartography cube* has been used as a way of conceptualizing the relationship between conventional map use and geovisualization (MacEachren, 1994b). The cartography cube has three axes: *interactivity*, *goals*, and *audience*. Interactivity is the degree to which users can manipulate and redefine a map. Goals refer to the degree to which a map is designed to help users discover new information. Audience refers to the degree to which a map is targeted at a specialized audience. In general, conventional maps are located in the lower left-hand corner of the cube, presenting known information to a public audience using low levels of interactivity. Conversely, geovisualization techniques are located in the upper right-hand corner of the cube, using high levels of interactivity to help private audiences discover new information. For example, a conventional topographic map might be useful in *presenting* known information, such as the route to the top of a mountain, but on its own it is less likely to help users *discover* new information, such as understanding the land-forming processes within that region. A conventional topographic map is also *public* in the sense that anyone, from hill walkers to town planners, might use it. Many of the views of information we encounter using a GIS are *private* in the sense that they might only exist on our computer screen for a few seconds (or less) before being altered or replaced.



thinking," and not necessarily to the visual IO channel (visualization may be based on any IO channel, e.g., visual, auditory, haptic).

Geovisualization is a branch of scientific visualization that deals specifically with geospatial information (see Box 8.11). Geovisualization design principles emphasize the importance of using the dynamic, interactive, and multimedia capabilities of GIS to help users gain insight into geographic problems, sometimes referred to as *geographic thinking*. Geovisualization techniques can extend the map metaphor—which treats GIS and geospatial data as digital analogs of hardcopy paper maps—in several ways. Breaking away from the familiarity, but also the constraints of the map metaphor, can result in GIS interfaces that are more expressive and arguably more intuitive too. The following sections examine in more detail three important visualization techniques with particular relevance to geospatial information: animation, 3D displays, and geovisual analytics.

geographic thinking

8.4.1 Animation

Conventional paper maps are static and, aside from wear and tear, their appearance is fixed at the time of manufacture. In contrast, using the computer to display a series of static images can convey the impression of motion or change over time, a technique known as *animation*. Each static image within the sequence that makes up an animation is called a *scene*.

animation

scene

Animation may be used simply to highlight or emphasize features on a static map, for example, by using a flashing arrow to indicate "You are here!" on an animated tourist information map. However, the most important function of animation is for visualizing *change* in geographic phenomena. Static maps can sometimes be used to depict change, for example, using static symbols such as arrows to represent flow or movement. Animation is a more intuitive mechanism for visualizing complex dynamic phenomena than static maps. Weather reports, for example, often use a chronological series of maps of atmospheric conditions to show the movement of a storm front.

fly-through

attribute re-expression

Animations are not restricted to representing change over time. Animations can also be constructed from any sequence of spatial or attribute changes. One example of an animated spatial change is *fly-through* animations, such as those often used in architecture to show building designs, in which the user's perspective on a static environment or data set is gradually moved. Attribute change, also called *attribute re-expression*, involves an animated logical sequence of scenes constructed from ordered attributes. Figure 8.24 illustrates the difference between animations of chronological change and attribute change. In Figure 8.24a, the animation can be interpreted as showing a gradual northerly migration of the population over a 30-year period. In Figure 8.24b, the animation can be interpreted as highlighting a spatial variation in population age, with younger people comprising a greater proportion of the population in the southern regions.

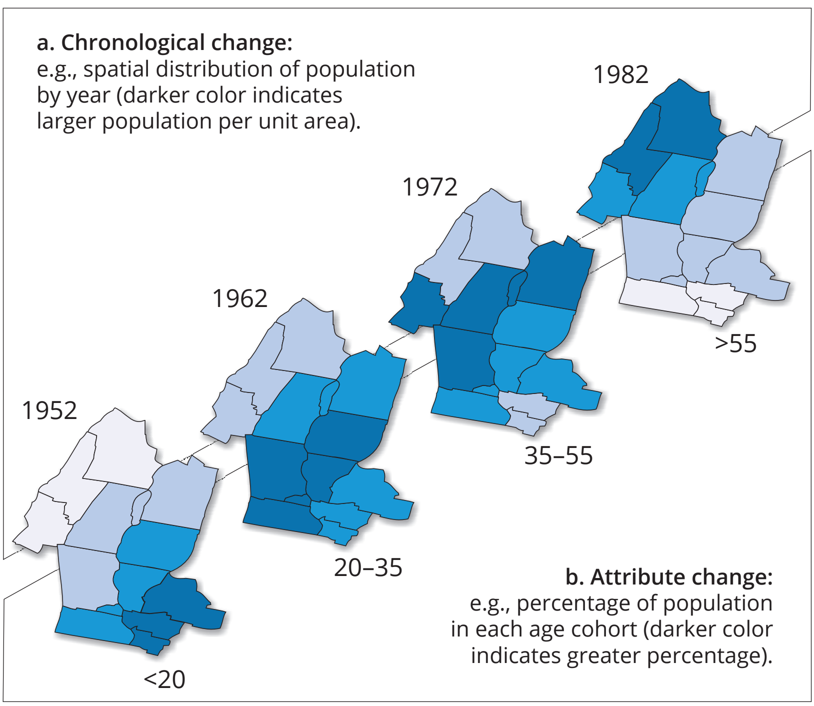


Figure 8.24: Chronological and attribute changes in animation

Animations are built from sequences of static scenes, so it follows that animations of geospatial data may use all of the static symbols in conventional maps, encapsulated by the six visual variables discussed previously. In addition, there are six *dynamic visual variables* that may be used to distinguish

dynamic visual variable

different features in an animated map: moment, frequency, duration, magnitude of change, order, and synchronization (Di Biase, MacEachren, Krygier, & Reeves, 1992).

The point in time at which a change occurs in an animation is referred to as *moment*. Moment may be thought of as the temporal counterpart of the static visual variable, position. *Frequency* is the rate at which a change occurs in an animation, and it is akin to the static visual variable *pattern*. *Duration* refers to the length of time that each static scene is visible. Duration gives an animation its pace. Scenes that are visible for longer move at a slower pace and appear to have greater importance or relative value when compared with those faster-paced scenes that are visible for a shorter length of time.

moment

frequency

duration

The *magnitude of change* is the amount of change that occurs in moving from one scene to the next. Sequences of scenes in which the magnitude of change from one scene to the next is small yield smooth animations. Sequences in which the magnitude of change is large yield abrupt or jumpy animations. The ratio of magnitude of change to duration in an animation is termed the *rate of change*. The fifth dynamic visual variable is the *order* in which scenes appear. The order of scenes in an animation is normally dictated by the chronology of the data (for example, in animating the spread of an epidemic over time) or the natural order of the attribute used to construct the animation (for example, the age classes in Figure 8.24b). Finally, *synchronization* is closely related to order and refers to the relative timing of changes in two or more phenomena represented in an animation.

magnitude of change

order

synchronization

Order and synchronization are particularly significant in animation, as they can be used to suggest a causal relationship between phenomena. As long as the magnitude of change is sufficiently small, users may interpret features in one scene as being the cause of features in a subsequent scene. In an animated map of the spread of an infectious disease epidemic, for example, a region of disease that enlarges through a sequence of scenes would normally be interpreted as contagion, in which individuals with the disease in one scene are the cause of further cases of infection in subsequent scenes.

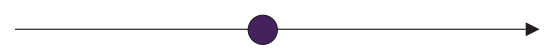
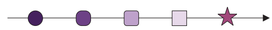
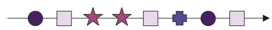
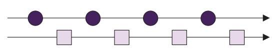
Table 8.3 summarizes the dynamic visual variables by showing changes as shapes located on an animation timeline, with time moving forward from left to right.

Animation in visualization should be used sparingly and with caution, however. Human visual perception is especially attuned to movement, and so incautious animation can be irritating and distracting to users. Further, humans are prone to pay attention only to certain parts, not all of an animation—an instance of a general feature of human perception termed *selective inattention* by psychologists. Selective inattention can cause users to miss important transitions in animations altogether, termed *change blindness* (explored in the context of animated choropleth maps by Fish, Goldsberry, & Battersby, 2011, for example). The sequential and timed nature of scenes can also make it difficult for users to gain an overview and identify insightful moments from longer animations. A 2002 review by Tversky, Morrison,

selective inattention

change blindness

Table 8.3: Dynamic visual variables, after MacEachren (1994a)

Dynamic visual variable	Example	Explanation of example
Moment		Moment of change is indicated with a circle
Frequency		Changes increase in frequency, from low to high
Duration		Result of square change lasts longer than circle changes
Magnitude		Change is gradual (circle to square) then abrupt
Order		Order of changes is apparently random
Synchronization		Changes in circle and square phenomena are closely related

& Betrancourt indicated that in practice animation fails to outperform a well-designed static data graphic in most cases.

8.4.2 3D displays

Conventional maps and most visual computer displays today are limited to two spatial dimensions; yet, the geographic world has three spatial dimensions. *Contours*, which connect points of equal height, are the most common mechanism for representing a topographic surface on a conventional 2D map. An example of contour lines is shown in Figure 8.25a, along with an example of a *hypsometric map* in Figure 8.25b. A hypsometric map uses a logical sequence of colors to indicate elevation. Such maps are only useful for displaying surfaces, not true 3D shapes like buildings, cliffs, or caves, because they can only show one (elevation) value at each location.

contour

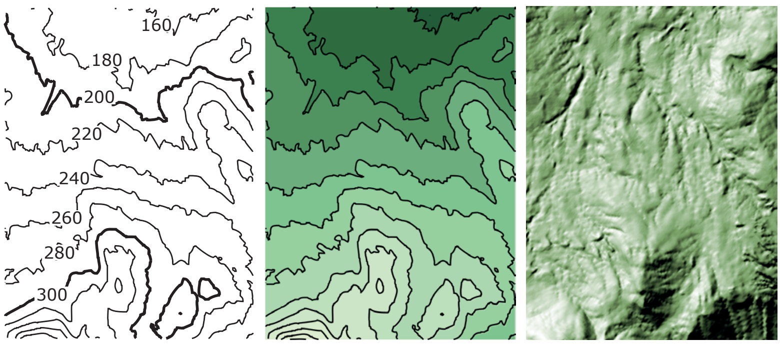
hypsometric map

Contour and hypsometric maps provide an abstract view of a 3D surface, with which a skilled mapreader is able to gain an excellent understanding of the shape of a surface. These types of maps are expressive but not particularly intuitive. Contour maps can be difficult for the novice mapreader to interpret. A more intuitive representation of 3D surfaces involves using psychological cues to give the impression of depth. There are several such *depth cues* that may be exploited in a 2D map or computer display. *Shading* may be used to give the impression that an object with height is casting a shadow. Figure 8.25c gives an example of hill shading, common in both conventional maps and computer displays. For topographic surfaces, the correct effect is most likely to be achieved when the source of illumination appears to be positioned in the northwest. Other sources of illumination are more likely to result in misinterpretation of the shape of surface, with valleys being mistaken for ridges, and vice versa.

depth cue

Box 8.12: Stereoscopic displays

A *stereoscopic* display takes advantage of retinal disparity, by showing precisely controlled different images to each eye using special hardware, such as a headset, stereo glasses, or stereo display screen. The need for specialized hardware means that stereoscopic displays are used less frequently than conventional visual displays. Further, a variety of eye-related conditions that can interfere with stereo vision (including losing sight in one eye) mean that a significant minority of people lack the ability to see binocular depth cues. Happily, in most interfaces monocular depth cues are amply able to provide the required 3D effect. Arguably, the sporadic popularity—and more tellingly the quotidian unpopularity—of stereo 3D movies is a reflection of the adequacy of monocular depth cues for understanding 3D scenes in most situations, without the need for binocular depth cues.



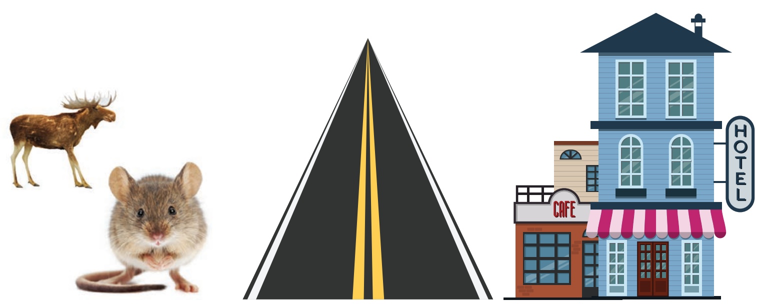
a. Contour b. Hypsometric c. Hill shading

Figure 8.25: Three representations of a terrain surface in two dimensions

Other depth cues include the *relative size*, *linear perspective*, and *interposition* of the objects depicted, illustrated in Figure 8.26. Familiar objects with a smaller relative size than expected appear to be farther away. For example, we would normally interpret Figure 8.26a as a moose that is farther away than the mouse, rather than as a mouse that is much larger than the moose. *Linear perspective* concerns the effect that parallel lines appear to converge into the distance, such as the edges of the road in Figure 8.26b. *Interposition* concerns the effect whereby distant objects are occluded by nearer ones, such as the cafe partially obscured by the hotel in Figure 8.26c.

linear perspective

interposition



a. Relative size b. Linear perspective c. Interposition

Figure 8.26: Relative size, linear perspective, and interposition depth cues

The impression of depth can also be achieved using animation, in which closer objects appear to move faster than more distant ones. This effect, which you can observe out of the window of any moving vehicle, is called *motion parallax*.

motion parallax

monocular

The depth cues described above are termed *monocular*, because they are effective with only one eye. Two other depth cues rely on *both* our eyes func-

binocular
convergence

tioning correctly (*binocular* depth cues). First, *convergence* occurs as our eyes move to fixate on near objects, such as your finger in front of your face. Convergence only influences depth perception for relatively close objects (less than a few meters away); eye convergence is negligible for objects in the middle to far distance. Second, as a result of the slightly separated positioning of our eyes, each eye receives a slightly different view of the world around us, termed *binocular* or *retinal disparity*. These slight discrepancies are used by the visual system to help determine how far away an object is, termed *stereopsis*. About 1 in 50 people, however, lack stereopsis abilities (termed *stereoblindness*) and many more (in some studies more than 1 in 5) have poor stereo vision. As a result, binocular depth cues should not be relied upon in geovisualization interfaces (Box 8.12 on the preceding page).

retinal disparity
stereopsis
stereoblindness

Both monocular and binocular 3D effects are commonly found in virtual and augmented reality interfaces. Virtual reality (VR) aims to immerse the user in a simulated or digitally generated interactive environment. VR environments are typically based on realistic 3D graphics and often delivered using stereo goggles or headsets. To complete the immersive effect, virtual reality environments usually also support visual display with other multimodal interfaces, in particular sound and haptics. In contrast to virtual reality, augmented reality aims to mix digital objects, accurately registered in 3D and in real time, into our view of the world around us. Augmented reality is more frequently unimodal visual display, because the sounds, smells, and feel of the world around us help make the experience more immersive.

virtual reality

augmented reality

Displaying physically 3D phenomena, such as topographic surfaces, is one obvious application of 3D interfaces in GIS. However, not all (or even any) of the dimensions need to be spatial to take advantage of 3D interfaces. In relation to geospatial information, the third dimension is often used to represent some other non-spatial data, a process known as *spatialization* (Skupin & Fabrikant, 2003).

spatialization

8.4.3 Geovisual analytics

visual analytics

Visual analytics is concerned with using interactive visual displays to support humans to identify and understand patterns and relationships in data. *Geovisual analytics* is the special case of visual analytics that deals with *spatial* data.

geovisual analytics

From a user's perspective, the purpose of all the different techniques encountered in this chapter can be seen as supporting a continuum of uses, from presentation of data at one extreme to supporting exploration and gaining insights into data at the other (cf. Box 8.11 on page 339). If conventional data graphics and traditional cartography have the primary purpose of "presenting knowns" in data, then geovisual analytics sits firmly at the opposite end of that continuum: "revealing unknowns" in data. When describing the goals of geovisual analytics, authors often use the terms *exploration* and *insight*, as well as terms like *discovery*, *hypothesis generation*, and *sensemaking*.

A range of different tools and techniques have been proposed to enable geovisual analytics. There are, however, a core of established interactive functions that underpin most geovisual analytics interfaces. The five most common basic interactions are discussed below, illustrated with a European Commission (EC) visualization tool for exploring European demographic statistics, the Eurosat RCI (regions and cities illustrated) geovisual analytics interface in Figure 8.27. The interface combines an interactive map view (left-hand side) with reconfigurable data graphics (right-hand side), such as scatter plots and bar charts. The key interactions the user can engage in are:

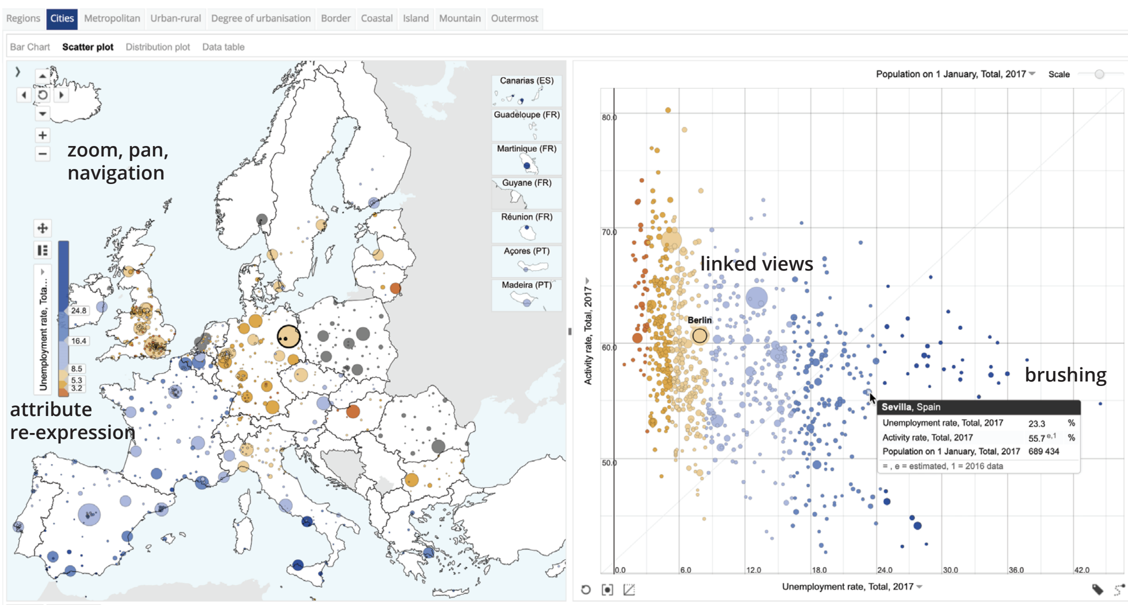


Figure 8.27: Eurostat RCI interface showing unemployment rates (color hue) and population (circle size) in European cities. The annotations highlight the geovisual analytics functions: panning, zooming, and navigation; attribute re-expression; brushing; and linked views (Source: European Union)

Panning, zooming, navigation Panning (change of viewpoint) and zooming (change of level of detail) are basic interactions for both spatial and non-spatial data. However, the ability to explore different views and pan and zoom around data can also be disorientating for users. Consequently, panning and zooming need to be paired with a high-level overview or a "home" function to ensure users can always re-orient themselves before they get lost in data. In Figure 8.27, pan and zoom can be achieved using the interface buttons in the top left corner or using familiar mouse gestures. Pan and zoom are complemented with a "reset" button that returns the user to the default view at any time. The scatter plot on the left-hand side also has its own independent rescaling functions, along with its own "reset" button.

Attribute re-expression Attribute re-expression is the ability to interactively change the way selected data is presented in the interface. For example, we have already seen in Figure 8.7 how changing the way data is classified can profoundly change the patterns that are evident. In Figure 8.27, the classification of unemployment rate (color hue) in the legend on the left-hand side can be altered interactively, with the user sliding class boundaries up and down. Doing this can be highly exploratory, highlighting previously unseen patterns, breaks, and relationships in data.

brushing

Brushing *Brushing* is a form of implicit input where a user's mouse hover is interpreted as a query for further information about any features located under the mouse pointer. In Figure 8.27, the mouse pointer's location above the scatterplot point related to Sevilla automatically brings up a tooltip box that contains more detail about the data for that city. The user need not explicitly type in a query or perform a mouse click. Simply lingering with the mouse automatically brings up the more detailed numerical data.

linked view

Linked views Linking multiple views involves changes to highlighted information in one view being mirrored by corresponding changes to other views. Linked views are especially well suited to exploring spatial data, with map views linked to non-spatial data graphics, such as bar charts or scatter plots. In linked views, a mouse click to highlight a feature in the map will instantly highlight the corresponding bar in the bar chart, for example. Similarly in Figure 8.27, highlighting Berlin's data point in the scatter plot immediately causes the corresponding data point in the map view to be highlighted with a thicker border.

Animation Animation has already been encountered in detail in Section 8.4.1. Animation can help users identify subtle patterns and interpret important changes not always visible in static views. As discussed, animations are often chronologically ordered but can also be constructed from other orderings, such as changes to attribute values (see Figure 8.24). Although no animation options appear in the interface in Figure 8.27, an animation toolbar does appear in the interface when displaying any time-varying data sets, such as population over time.

Reflections

The principles behind data graphics and HCI also lie at the foundations of cartography and geovisualization. Well-designed maps must also be well designed data graphics; effective GIS interfaces must additionally encompass good interaction design. In addition to the classic texts by Norman (1988) and Tufte (1990, 2001) already cited, there are some excellent textbooks on HCI and visualization topics from a computing perspective, with Munzner (2014), Preece, Rogers, & Sharp (2015), and Ward, Grinstein, & Keim (2015) especially recommended.

At the same time, geographic information scientists have unique expertise and perspectives to contribute to design conversations. Cartography draws on the world's oldest design traditions; maps were the first data graphics. Under the heading of "interfaces," cartography and geovisualization were identified back in Chapter 1 as one of the five key themes at the core of GI science. Unsurprisingly, then, there are a range of authoritative books and introductory texts on both cartography (e.g., Griffin, 2017; Peterson, 2020) and geovisualization (e.g., Andrienko & Andrienko, 2006; Çöltekin, Bleisch, Andrienko, & Dykes, 2017).

The essence of this design expertise is an understanding of the rules governing good information and interaction design combined with a spark of creativity and innovation. While there are few *inviolable* rules in design, there *are* correct and incorrect design decisions, good and bad designs, as we have seen. For example, Figure 8.28 shows one all-too-common mapping mistake—using color hue to represent a value message, in this case a spectral color ramp to depict elevation in a DEM. Hue is a poor choice for communicating numerical values such as height, as we have seen in Section 8.1.2. Not only does hue fail to communicate a value message (does red indicate a ridge or a valley, show more or less height?), but also the transitions between hues creates the impression of false boundaries in the surface (such as between blue and green and between green and yellow). Penn State University Professor Cynthia Brewer is the preeminent authority on color usage in maps, and her 2016 book on map design is essential reading for anyone planning to use color in maps.

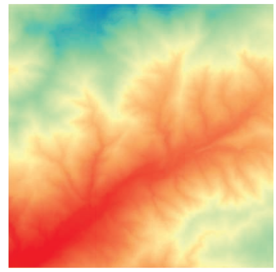


Figure 8.28: Don't do this! A spectral color ramp incorrectly used to depict height in a DEM

Some rules are, however, made to be broken. Generations of geography students have been drilled on the need for "map furniture," for example. However, rigid adherence to the map furniture mantra—"scale bar, legend, north arrow, title"—is certainly not essential to effective communication with maps, and indeed it is more likely to leach the beauty, creativity, and joy from a design. A combination of proficiency and panache, finesse and flair is a hallmark of great design, and the most technically adept designers in GI science are frequently also the most creative (enjoy, for example, Andrienko et al., 2010; Bleisch & Hollenstein, 2018; Goodwin, Dykes, Slingsby, & Turkay, 2015; Harrower & Brewer, 2003).