1

Exploring Statistical Relationships with Maps and Charts

Jean-Benoît BOURON[1], Claire CUNTY[2], Hélène MATHIAN[3] and Myriam BARON[4,5]

[1] *Géoconfluences, DGESCO, ENS Lyon, France*
[2] *EVS, Université Lumière Lyon 2, France*
[3] *EVS, CNRS, Lyon, France*
[4] *Lab'Urba, Université Paris-Est Créteil, Créteil, France*
[5] *IDEES, CNRS, Université de Rouen Normandie, Rouen, France*

1.1. Introduction

Spatial analysis involves identifying, describing and explaining the spatial organization of a phenomenon. It uses data and statistical methods to shed light on these issues. Such an approach implies a dialogue between geographical and thematic dimensions, in which maps obviously play a central role. Either methods used are derived from classical statistics[1] and do not give a specific role to space – any spatial effects are revealed a posteriori by transferring the result onto the map – or methods are derived from spatial statistics and directly integrate the geometric component of data. Here, we select first methods, and more specifically, an exploratory and descriptive one, combining data processing with graphic and cartographic representations. We have chosen to base this approach on a single

1 The term "statistics" refers both to the disciplinary field that studies phenomena by analyzing data, and to data themselves, as well as to methods used to make these data understandable.

Handling and Mapping Geographic Information,
coordinated by Claire CUNTY and Hélène MATHIAN. © ISTE Ltd 2024.

scale, that of the world, which raises interesting questions about the definition of entities in relation to the two points of view we are interested in here.

From a thematic point of view, that is from a geographic and mapping perspective, a planisphere is a singular form. In the hierarchy of geographical units, it is first and foremost the macro or global scale. It is the only scale that enables us to apprehend all geographical units as a closed system. From a cartographic point of view, this scale imposes *projection* choices that do not necessarily arise at other scales. While the correct conservation of distances and areas is a well-known constraint (see Chapter 7), projecting a sphere onto a plane also transforms proximities at a scale where borders have a very strong meaning. We therefore need to think specifically about this. Finally, at this scale too, a recurrent issue in geography and cartography emerges, that of the unequal size of spatial entities (known as a modifiable areal unit problem [MAUP]; see introduction and Chapter 4).

From a methodological, that is, statistical, point of view, the entities studied (countries) form a whole that cannot be considered as a sample. Like all territories, of course, they have an existence independent of any experience. However, countries of the world are undoubtedly one of the most "familiar" sets of spatial entities, generally made familiar from an early age onwards as part of national education curriculums. Despite its complexity, this collective is understood with many a priori thematics: as such, it thus constitutes a set of "cases", rather than a set of "representations" part of a larger population. All this makes it more difficult to assimilate countries into statistical entities. Such a set is therefore particularly well-suited to illustrate and discuss the descriptive and exploratory (rather than a predictive and confirmatory) framework within which statistical methods will be mobilized.

In this context, this chapter will focus on the analysis and representation of relationships between variables. The observation of a geographical phenomenon can rarely be reduced to the analysis of a single measurement. Knowledge of any observation considered from the point of view of the "spatial" dimension implies starting by mapping a variable (an indicator). The latter must enable us to account for the distribution, in both spatial and statistical sense, of the theme (e.g. infant mortality rate by country; Covid-19 incidence rate by region; transaction prices by quarter; etc.). The map makes it possible to simultaneously apprehend the thematic and spatial variability of the variable's values, that is, both the form of heterogeneity of the values (are they concentrated around the mean value, or on the contrary, widely dispersed?) and the forms of organization that this heterogeneity draws in space: concentrations, regionalization, oppositions, discontinuities, gradients, etc. As such, one map calls for another, and we quickly enter a phase of exploration that leads us to question this variable in relation to others, trying to understand the simultaneous (or not) localizations: What is the birth rate in the least affluent countries? What is the Covid-19 mortality rate in the most densely populated regions? What is the income of residents in the most expensive neighborhoods? etc. By comparing two or three (or more) maps, we begin to analyze similarities and dissimilarities, and their geographies, which leads us to question the relationships between variables represented (Figure 1.1).

Geographic data exploration involves a sequence of processing and cartographic representations, combining phases of data comprehension, sometimes phases of transformation and phases of analysis using statistical methods to reveal and model relationships between different variables. In this exploratory approach, graphic and cartographic representations are an integral part of the analysis. Visualization stimulates questioning and drives reasoning: it is a major dimension of comprehension and explanation. The question of cartography itself would require us to revisit many basic notions of semiology and statistics: for this, we refer to some previous studies (Béguin and Pumain 1994; Cauvin et al. 2010a; Lambert and Zanin 2016).

In this chapter, we propose to discuss the steps involved in this data exploration process, using an example that draws on the global scale and, as we go along, points out the specific features of this scale: how to "divide" the world, according to the major wealth and development disparities? The thematic challenge associated with this example is as much scientific as it is pedagogical, given that major historical divisions (e.g. "Third World", "developing" countries, "North and South") have been progressively abandoned without being replaced by other terms (Bouron et al. 2022).

The aim here is essentially to outline the methodological tenants of this approach in order to show how data mining can shed light on these ways of divvying up the world. After addressing the cartographic issues involved in representing the world (section 1.2), we introduce the issues involved in exploring statistical relationships based on two variables (section 1.3). We then illustrate this question, generalized to several variables and focused on the challenge regarding "world dissection", by proposing an approach based mainly on the visualizations of statistical treatments employed (section 1.4).

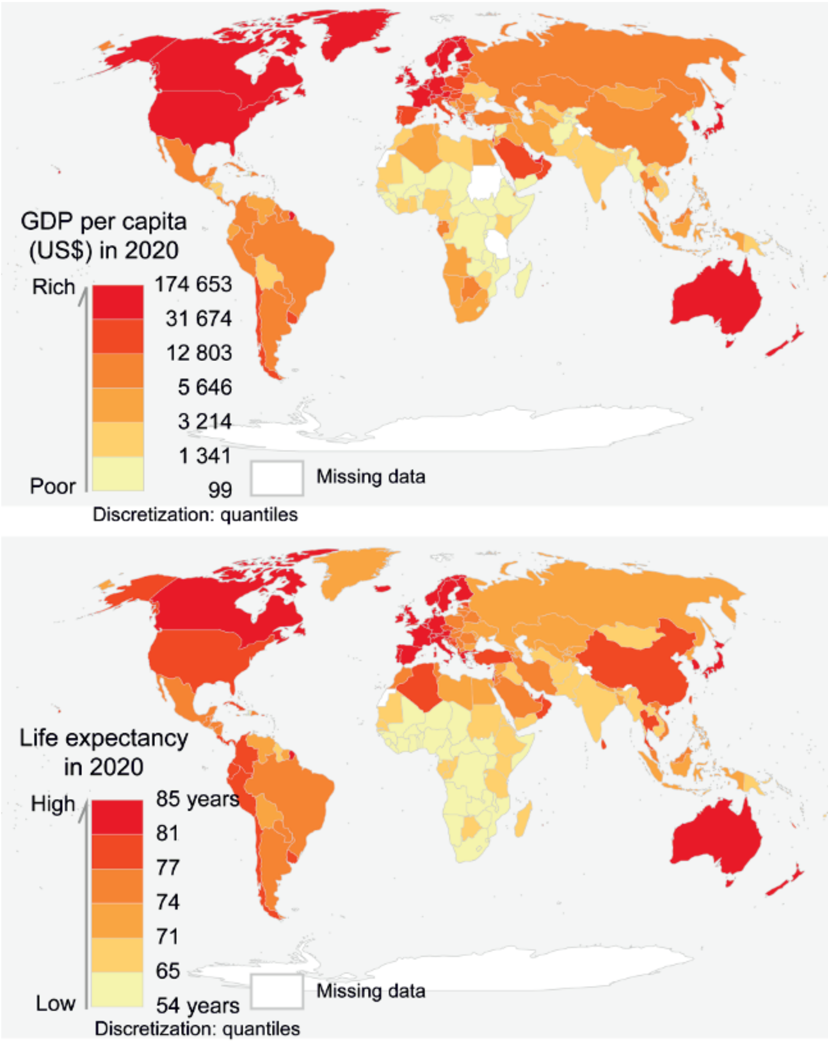


Figure 1.1. *Compare spatial organizations: GDP/capita and life expectancy in world's countries (sources: World Bank Eurographics; Software: R (mapsf) J.-B. Bouron, C. Cunty, H. Mathian)*

1.2. Mapping the world: which world? what data?

As with any cartographic approach, there are first of all the issues of the map base and its projection, the spatial grid and the associated data. We present these in the following three paragraphs and discuss their specificity considering countries of the world. First of all, the choice of projection that will allow for the transition from globe to planisphere is an important issue. As for the grids, although determined (countries, states), these can likewise be interrogated further, for instance, in terms of the different organizations of political systems (countries, federations of states). Finally, the national statistics that will be mobilized also reflect these differences and, as with any international analysis, raise the question of their thematic harmonization and synchronization.

1.2.1. *Planispheres, projections and centers*

The ways countries belong to bigger entities and how close or far they are from one another can be perceived differently depending on the projection chosen (Grataloup 2007; Bahoken 2022). The cartographic representation of data, depending on the choices made, orients mental images we have of the world's different countries (Lambert and Zanin 2019; Monmonier 2019). These countries can be seen, for example, as margins of the various oceans and seas, recalling that 80% of the planet is occupied by water, as shown by the Spilhaus projection. This thus seems to fit in with the continuity of the ancient world as defined by the Greeks and Romans around the Mediterranean. The question of the map's center also depends on the point of view proposed (or imposed) by the cartographer. Countries of the world can thus be viewed from the southern hemisphere, in particular Australia, providing an unexpected representation of the Indo-Pacific world for some of the inhabitants of the northern hemisphere, notably Europeans and North Americans (see Figure 1.2(a)).

When it comes to representing the countries of the world on a planisphere, a certain continuity can be demonstrated since the second half of the 19th century, with polar projections predominating. The latter were proposed as early as the 1860s, notably by Gustav Jäger, a zoologist keen to show ancient connections between great continental masses and, consequently, links that might exist between different species distributed across the world (Capdepuy 2013). The choice of the planisphere featured on the flag of the United Nations Organization (UNO) from 1945 to 1950 extends these proposals, as do the projections put forward by Jacques Bertin in the early 1950s, and even those of Christian Grataloup in the 2000s. This family of projections makes it possible to account for the interconnections between the different countries of the world, most of which are located in the northern hemisphere, without the a priori selection of a center that would orient the viewer's gaze. However, this last point has been debated, as illustrated by the successive planisphere projects for the UN flag. The American continent, or even its northern part, was first the selected focus, although it was Europe that was ultimately preferred for the final version, to better highlight the opposition between the two superpowers (Capdepuy 2011). In any case, a polar projection centered on the North Pole relegates certain countries to the periphery, in particular New Zealand, which in this case appears highly distorted. In all cases, there is no projection that can avoid focusing on one part of the world, without relegating another to the periphery.

The choice of a polar projection, to emphasize not only links between different parts of the world but also major geopolitical oppositions, almost makes us forget the deformations inherent in any projection, that is, the deformation of surfaces and/or directions when it comes to representing the Earth's geoid in two dimensions. This is particularly striking with the Gerardus Mercator projection (Figure 1.2(b)). Dating from 1569, it is widely (if not excessively) used, and not just for the navigational charts for which it was intended. It places great emphasis on spaces from mid-latitudes upwards, especially at high latitudes, showing them to be much larger than they are, like any conformal projection that preserves angles and therefore shapes to the detriment of surfaces. Russia thus appears as large as the entire African continent, whereas the latter, with its 30.4 million square kilometers, is in fact almost twice as big. Alternatives exist, such as Arno Peters' projection, presented in the early 1970s, which has many similarities with the one formalized by James Gall in the mid-19th century (Figure 1.2(c)). It corrects for differences in size between continents and countries, but at the cost of distorting the shape of the latter, like any equivalent projection. For some authors, this projection seems to favor an alter-globalist vision of the world, by considerably reducing the influence of the so-called northern countries, particularly those of Europe and North America. To remedy this, a cartographic representation of the countries of the world can be based on the projection of Richard Buckminster Fuller, created as early as 1946 (Figure 1.2(d)), which respects the surfaces and shapes of these countries. But on the other hand, we are faced with a "splintering of the world" according to Roger Brunet and Olivier Dollfus in their volume devoted to New Worlds in the latest *Géographie universelle* (1990), while Nicolas Lambert and Christine Zanin highlight the promotion of a "universal world without north and south" (2019). These examples of projection and their interpretations show that we are a long way from "cartographic production" assimilated to "a 'neutral' tool, independent of those who [make it]" and thus from an "objectivation of space" (Lévy and Lussault 2003). Mapping relationships between geographical data in an exploratory way thus begins with a comparison of the different projections likely to be selected and the identification of the ways in which they will modify the visualization of information. In this chapter, we shall be using the Equal Earth projection, which preserves surfaces and gives an idea of the Earth's rotundity. It is currently considered a good compromise for representing global-scale themes (Genevois 2018; Šavrič et al. 2019).

1.2.2. *Grid-related issues*

Observing the world at the level of countries[2] raises a number of questions relating to borders and the grid effects they have on cartographic images. The aim here is not to discuss territorial conflicts and/or disputed borders, even though these inevitably arise on a global scale, but to reflect in a general way on the various grids and their impact on cartographic representations. Grids are defined as a partition of space "into units whose shape and size may be regular or irregular" (Pumain 2008). In the case of national borders, and more generally in the case of administrative borders, boundaries separate irregular geographical units to which statistical variables can be associated, most often produced by state statistical agencies. These statistical variables are in fact derived from the aggregation of individual data, the aggregation criterion being the location within a geographical unit. However, for certain themes, the relevance of this aggregation can be questioned. On the one hand, these grids may differ from those of territorial governance. For example, how can we consider the United States as a whole, in the same way as France, when it is first and foremost a federation of states? On the other hand, observation at country level automatically leads to a comparison of characteristics from one country to another, considering each national characteristic as a homogeneous whole. However, in some cases, the differences observed within a country for a given variable can be much greater than the differences between countries. For example, differences in literacy rates among adult women in a country like Niger will be more marked between urban and rural areas within the country than with neighboring countries such as Chad or Mali. Geographers were quick to identify the problems associated with using and interpreting data aggregated within variable-sized grid cells under the name of modifiable areal unit problem, or MAUP (Openshaw 1983).

The grid is used to transfer thematic information to produce a map, providing information on the distribution and spatial organization of the phenomenon under study. In the case of choropleth maps, the variability of spatial units' sizes will lead to a bias in the perception of small countries, to the benefit of those with a larger surface area. Admittedly, the smallest countries in terms of surface area are often those with the fewest inhabitants: seven of the world's 15 least populous countries are also among the smallest; but these small countries are also frequently the most densely populated and the richest: the countries at the top of the various rankings by GDP per capita are all "small countries", in terms of surface area (e.g. Liechtenstein, Luxembourg and Monaco).

2 We use the term country here, even if we often implicitly refer to states. We will consider here the 193 countries currently recognized by the UN.

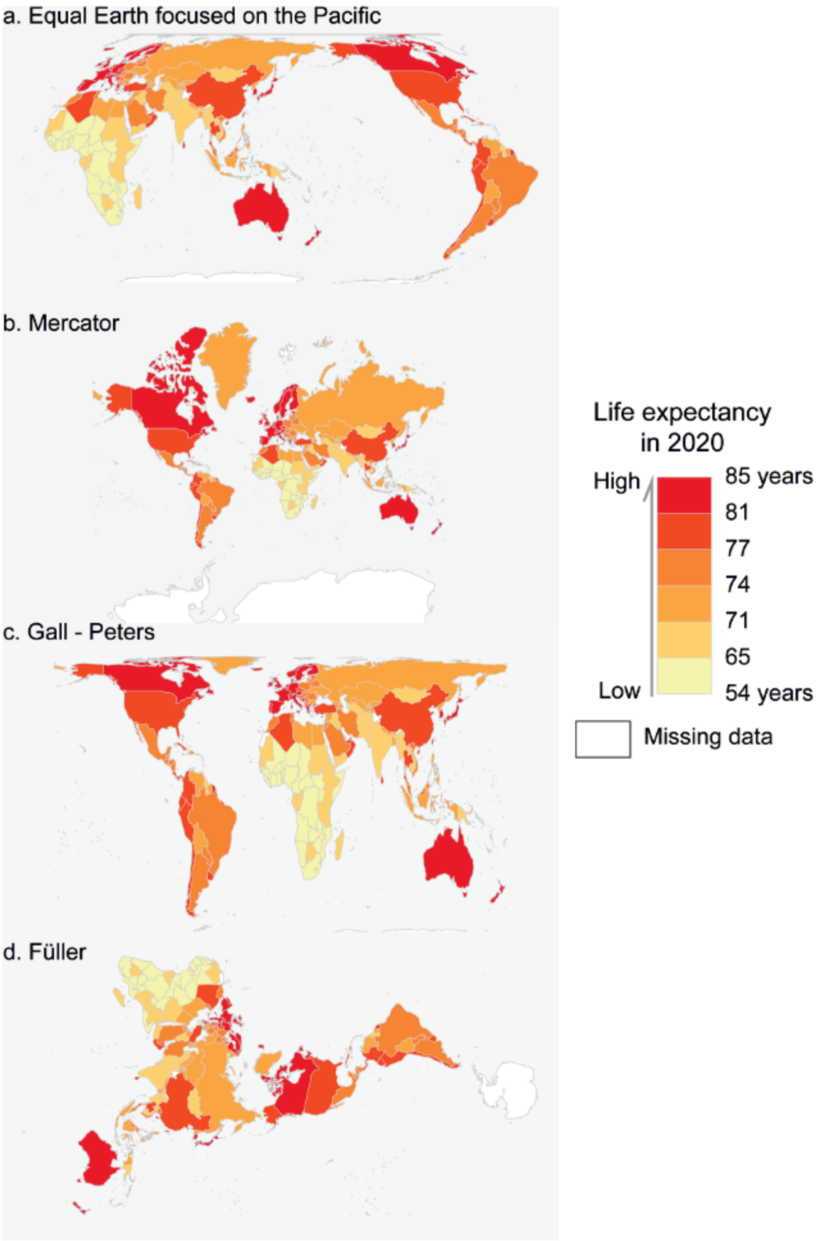


Figure 1.2. *The world in all its forms and projections. Variable represented: life expectancy (source: World Bank Eurographics; Software: R (mapsf) J.-B. Bouron, C. Cunty, H. Mathian)*

In this type of cartographic representation, space is seen as a support on which the cartographer simply transfers thematic information. The map thus plays the role of a revealer (Cauvin et al. 2010b, pp. 66–67) of spatial structures. When observing these structures, we need to keep in mind all the underlying elements mentioned above to avoid misinterpretation.

1.2.3. *Data at global scale: a composite material*

Thematic maps at global scale are well known in geography and cartography, and have been since the very first geography lessons at school or college, so much so that this type of representation can be taken as a given and not as a construct. Yet that is precisely what it is, and it is just as important to question the way in which this construct is produced as much as it is questioned at other scales. Collecting data on a global scale is far from self-evident and involves a number of pitfalls.

One of these difficulties lies in conditions under which data represented are produced. Most country-level data are collected by international organizations, the two main ones being the UN and the World Bank, plus the CIA, the US federal agency. These organizations have their own statistics departments, whose role is to collect data from the world's 193 States and numerous non-State territorial entities, and harmonize them in order to produce databases.

Let us take the most basic element of any national statistic: population. This is the basis for all data expressed on a per capita basis. None of the three agencies mentioned above has a world-wide census service. They therefore depend on national census services. At this stage, it is already clear that not all countries are in a position to carry out an exhaustive and reliable census. What are the working conditions of census-takers in failed States where the State no longer controls the administration? In informal settlements where the population does not necessarily wish to answer questions from public officials? In inaccessible mountainous areas? In war-torn regions? In regions where nomadism dominates? It is easy to see why statistics departments receive data of very uneven reliability. They are therefore obliged to correct them, based on previous censuses, compilation with other sources and their own knowledge of the local situation in each country. The question arises as to the circumstances in which these corrections are made, all the more so as estimates are added to these corrections: indeed, since each country has its own census time step, the data collected are at different dates. The United States has had a decennial census for two centuries. Some countries take their census when they can. In France, before the adoption of a rotating annual census system in 2004, intercensal periods varied between 5 and 9 years. The new French system, the only one of its kind in the world, is not without its critics in terms of reliability (Dumont

2018). In India, the census is decennial, but Covid-19 has delayed the 2021 census: in 2023 or 2024, data from the latest census will be available, more than 10 years after the previous census, carried out in 2011.

Usually, a thematic planisphere announces a variable at a specific date: for example, the number of cars per capita in 2020, for a map produced in 2022. This simple fact is already a bit of an arrangement with reality, since the 2020 population, on a global scale, combines results from recent censuses, corrected censuses, estimates from older censuses and estimates from corrected older censuses. This is not to claim that the world population figures by country are fanciful: these figures are not reality, but a reconstruction of reality.

And even then, we are only talking about the most common data, the total population. For other indicators, the quality of data collection is even more problematic. For example, the World Bank provides data on the ratio of girls to boys enrolled in elementary school at the latest available date. This date varies widely: 2020 for Algeria, 2015 for Angola, 1998 for Haiti and 1985 for Lebanon. Out of 220 territorial entities, only 104 have a figure for 2019 or 2020.

The situation is even more complicated for composite indicators, which are the result of a second statistical construction from raw data, which we have just seen to be of variable reliability. Therefore, the HDI (Human Development Index) depends on several factors, including the level of education, measured by the average number of years of schooling for adults aged 25 and over, and by the expected (and therefore projected) number of years of schooling for school-age children. We could similarly deconstruct GDP, another "totem" of international comparisons.

In our example, we have sought to use the crudest possible indicators, even if we have had to resort to composite indicators such as GDP per capita. By using these data, despite their inherent weaknesses developed above, we also want to affirm that these indicators remain, for the time being, the most acceptable way of describing the countries of the world according to a series of harmonized variables. They are the "least bad" tools for comparing countries. Our condition is that the use of these global indicators should be systematically combined with other indicators, measuring environmental impacts for example. Measuring a population's happiness, its footprint on the environment, does not yet give rise to the production of satisfactory global statistics: we still see maps of CO_2 emissions per capita based on domestic emissions, without taking into account "imported" emissions; China seems to have higher emissions per capita than Germany, without taking into account the fact that China's emissions largely fuel the consumption of countries richer than itself. Under these conditions, the choice of variables to classify the world's countries according to their level of wealth and development becomes fraught with challenges.

In our example, variables were selected by iteration, testing different combinations of variables. Eight variables were selected according to a number of criteria: in addition to their relevance to the question posed (the level of wealth and human development), these criteria mainly concerned the (relative) reliability of the producing organization, the availability of recent series and the low number of missing data (Table 1.1). The first phase of this work hence involved exploring the relationships between the variables.

Variable name	Variable label
Fertilr19	Fertility rate in 2019* – Source: World Bank
infmortr20	Infant mortality rate in 2020* – Source: World Bank
gdpHab2020	GDP per capita in 2020 – Source: UN (national accounts)
TcrGDP0020	GDP per capita growth rate between 2000 and 2020** – Source: UN (national accounts)
ConsoHholds	Household consumption per capita in US$ in 2020* – Source: World Bank
fcfHabave10yrs	Gross fixed capital formation in US $ per capita (10-year average 2010–2020) – Source: World Bank
T10B50	Income gap between richest 10% and poorest 50% – Source: Chancel Lucas et al. (2021)
Lifexp20	Life expectancy at birth 2020* – Source: World Bank
LogGDP2020	Logarithm of GDP per capita in 2020 calculated from GDP per capita (gdpHab2020)
*Data date. For each country, the most recent date from 2014 is used. Data prior to 2014 are considered missing. **Not to be confused with annual GDP growth. This is GDP per capita growth over 20 years.	

Table 1.1. *Data mobilized (blue: the data analyzed in section 1.4; italics: the other data mobilized in the exploratory phase)*

1.3. Exploring data and relationships with maps and charts

Analyzing a phenomenon generally requires the mobilization of several variables. This may involve analyzing the spatial distribution of an indicator constructed from several variables, as in the case of the HDI, widely used to assess the economic and human development of the world's countries. However, if we wish to examine the differentiation between countries on the basis of different perspectives, we need to call upon several variables and study them simultaneously. While the first step is obviously to map each variable separately, in order to integrate all similarities and any dissemblance between geographical distributions, this type of approach must be accompanied by statistical methods to help understand data.

Figure 1.1 illustrates the temptation to see spatial coincidences when comparing the two maps. However, this analysis is only visual and needs to be systematically confirmed. This is what the statistical visualization of this relation offers, in the form of a scatter plot (Figure 1.3(a)). By associating statistical graphs with map analysis, we can validate and confirm a spatial resemblance that is visually assessed in a more qualitative way ("countries in the south tend to be the least wealthy, with low life expectancy, while further north this is variable"). More realistically, associating maps with a statistical relationship makes it possible to specify the geography of this relationship, to which several spatial configurations may (or may not) correspond.

The aim of this section is not to go into specialized statistical treatments, which are the subject of other chapters in this book (see Chapters 3–5), but rather to introduce how statistical and cartographic representations form the basis of spatial analysis reasoning, enabling us to account for spatial forms such as gradients and regions. The multiplication of viewpoints, through various maps and numerous graphs, encourages an exploratory outlook.

This exploratory position is well known in geography and statistics (Tukey 1977), combining abduction, induction and deduction (Banos 2005). Observation enables us to identify a phenomenon through empirical exploration (induction), then to formulate hypotheses relating to the explanation of this phenomenon (abduction), and finally to formalize them at a theoretical level (deduction). In this context, we present a simple example of data exploration in an interactive environment (section 1.3.1). The identification of a relationship between two variables raises the issue of its generalization. This is the second stage of the approach (section 1.3.2) known as hypothetico-deductive. It leads to the formalization of the phenomenon (the model) and the calculation of new dimensions to be explored (i.e. any deviations from the model).

1.3.1. *Exploring statistical relationships and spatial organizations*

The fact that two variables show spatial coincidences visible on two maps (Figure 1.1) is interpreted as the fact that the two phenomena are connected, either directly (causal link) or indirectly (one or more common factors explain the two organizations). It may also be fortuitous (Charre 1995). The statistical graph used to represent a relationship is a Cartesian graph, called a “scatterplot”, which cross-references the values of each of the variables, and the intensity of the relationship is assessed by the significance of the correlation coefficient (R, Pearson’s coefficient). Figure 1.3 shows three emblematic cases of the existence or absence of a relationship: at country level, GDP per capita (in log) and life expectancy vary together and in the same direction (Figure 1.3(a)). In this case, the relationship is one of direct, multi-fold causality. Figure 1.3(b) shows a strong covariation between a mortality rate and life expectancy, one that is inverse (or negative). As the mortality rate is one of the components of life expectancy, this relationship reflects a form of redundancy. Finally, Figure 1.3(c) shows that there is no relationship between the level of GDP in 2020 and its growth rate between 2000 and 2020, illustrating the fact that the level of GDP has no influence on its growth, as economic dynamics are highly variable depending on the country’s economy, irrespective of the level.

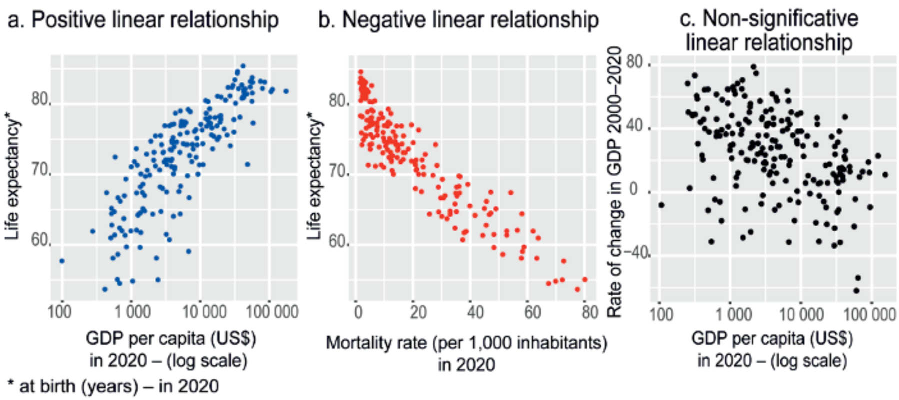
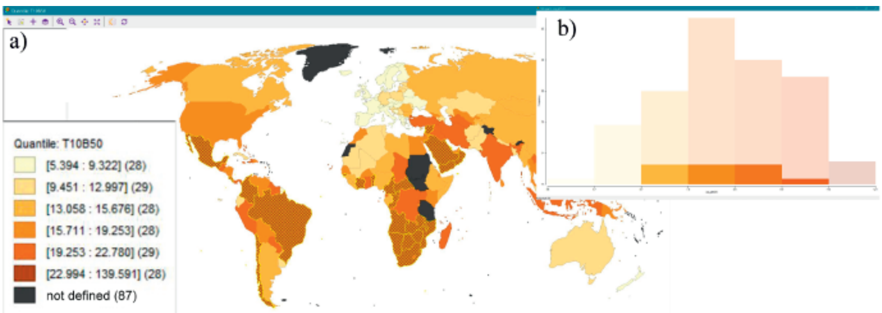


Figure 1.3. *Statistical linear correlations: three emblematic cases (sources: UN, World Bank, J.-B. Bouron, C. Cunty, H. Mathian)*

While these statistical graphs can be used to confirm covariation hypotheses, they cannot be dissociated from the maps. The latter reveals the spatial relationships, that is the configurations associated with this statistical relationship: a single relationship can correspond to several spatial organizations, and any spatial organization is the result of a process in which space is never neutral. Geographical proximity, contiguity, etc., explain the spatial interactions that shape the spatial configurations of the themes observed. Combining graphs and maps offers a wealth of possibilities for questioning data, stimulating questions at various levels: which countries, with equivalent and rather low GDPs, have different life expectancies?

What diversity of GDP growth rates does this group of countries of similar economic and geographical levels present?

In the trend initiated by Tukey (exploratory data analysis [EDA]), exploratory spatial data analysis (ESDA), adopted by a large number of quantitative geographers, has led to the development of geographical data exploration environments combining maps and graphs, visualizing geographical shapes and statistical relationships. By offering the possibility of interacting with graphics through selections that are visible on different cartographic representations, these environments allow for a differentiated understanding of relationships concerning space, the interrogation of data to test hypotheses, or the pushing of investigations to render unexpected observations. These environments incorporate a large number of general statistical and spatial methods for analyzing, simplifying and modeling data, thus helping to identify statistical relationships and explore their various spatialities, at different levels (Anselin 1995; Unwin 1998). Figure 1.4 shows an example of this multi-window approach, integrating maps and interacting graphics. A set of countries has been selected on the inequality map, in this case the most unequal countries (by clicking on the corresponding legend box). On the one hand, this selection makes it possible to visually assert the geographical organization of these inequalities by underlining them with yellow shading. On the other hand, this selection has repercussions on the other graphs, in this case the distribution of GDP by country (right-hand window), showing the diversity of wealth levels in the countries concerned by this level of inequality.



The selection of the most unequal countries on map (a) is reflected in the GDP distribution (the part of histogram (b) representing the selected countries is highlighted).

Figure 1.4. *Example of two interactive windows, dynamically linked in Geoda software (Anselin 2003)*

The selection of a set of geographically close countries could also have been selected and displayed on the map in order to visualize the specificity of this subset within the statistical distribution. These environments allow multiple windows to be opened, revealing "partial" or "regionalized" relationships. Figure 1.5 generalizes this type of approach to several windows and more than two variables. We can look at the variability of one variable, or a covariability of two variables, with the values of another fixed. Here, we select from the scatterplot the intermediate GDP values that present a wide diversity of life expectancy values (Figures 1.5(a) and (b)). Figure 1.5(c) shows the position of this subset within the distribution of a third variable (in this case, inequality), and the map shows the location of the selection (Figure 1.5(d)). In the same way, we can reason the other way round by selecting a geographical region and visualizing the positions of this subset in the statistical graphs. These statistical exploration environments make it possible to visually interrogate the associations between variables and their spatial translations, putting the map back at the heart of the analysis process, and with it, a reflection on space.

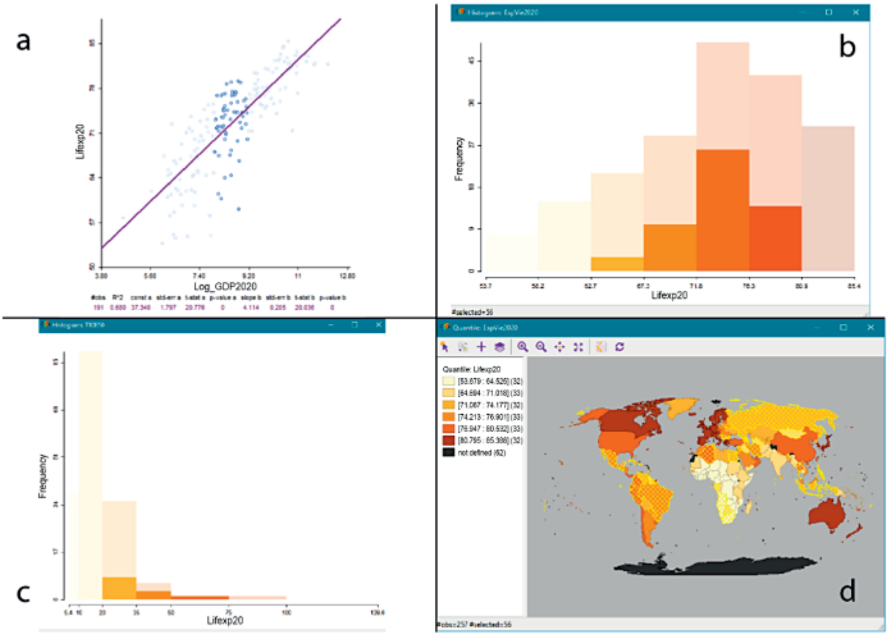


Figure 1.5. *Exploring relationships between several variables by selecting and visualizing positions from both a statistical and a geographical point of view*

REMARKS ON FIGURE 1.5.– *(a) Scatterplot showing GDP and life expectancy in 2020, (b) distribution graph showing life expectancy in 2020, (c) distribution graph showing the inequality indicator 2020 and (d) map showing life expectancy in 2020.*

1.3.2. *Formalizing a relationship: from a statistical model to the map of model deviations*

Moving from the stage of identifying a similarity or coincidence of spatial forms between two maps to that of identifying the relationship between the two associated variables amounts to generalizing a relationship observed across all the units studied. In our case, this translates, for example, from an empirical observation such as "the least wealthy countries have low life expectancies, while the richest have high life expectancies" to an assertion such as "knowing a country's GDP per capita, it is possible to deduce its life expectancy" (Figure 1.5(a)).

Given the *strong identity* of the entities we are working on, and the exploratory nature of our approach, a descriptive approach may suffice to account for the relationship. A *confirmatory* approach, associated with the formalization of a model (regression line), would be of interest for the purpose of evolution analysis (model parameters could be compared over time). Another interest in formalizing the relationship lies in the analysis of deviations from this model, and in particular their cartographic representation, which may reveal other explanatory factors. Here is an example of how a statistically processed graph can be used in conjunction with a map to illustrate both a relationship and that which lies outside it.

Among variables we used in this work on the division of the world, some were highly correlated, such as household consumption and GDP per capita ($R^2 = 0.86$)[3]. Identifying a strong linear correlation makes it possible to formalize a relationship (ConsoHholds = 0.47 × GDP + 1204), associate it with a quality of representativeness (R^2) and, depending on the case, infer this relationship on a larger population. In our case, the population (all countries in the world) is exhaustive, and the aim is not to predict consumption on the basis of GDP. However, the formalization of the model allows us to calculate the slope of the relationship (0.47), which can be interpreted as the average share of GDP in this consumption[4].

3 This means that 86% of differences in household consumption can be "statistically explained" by differences in GDP.

4 On average, household consumption (over US$ 1,204) accounts for 47% of GDP.

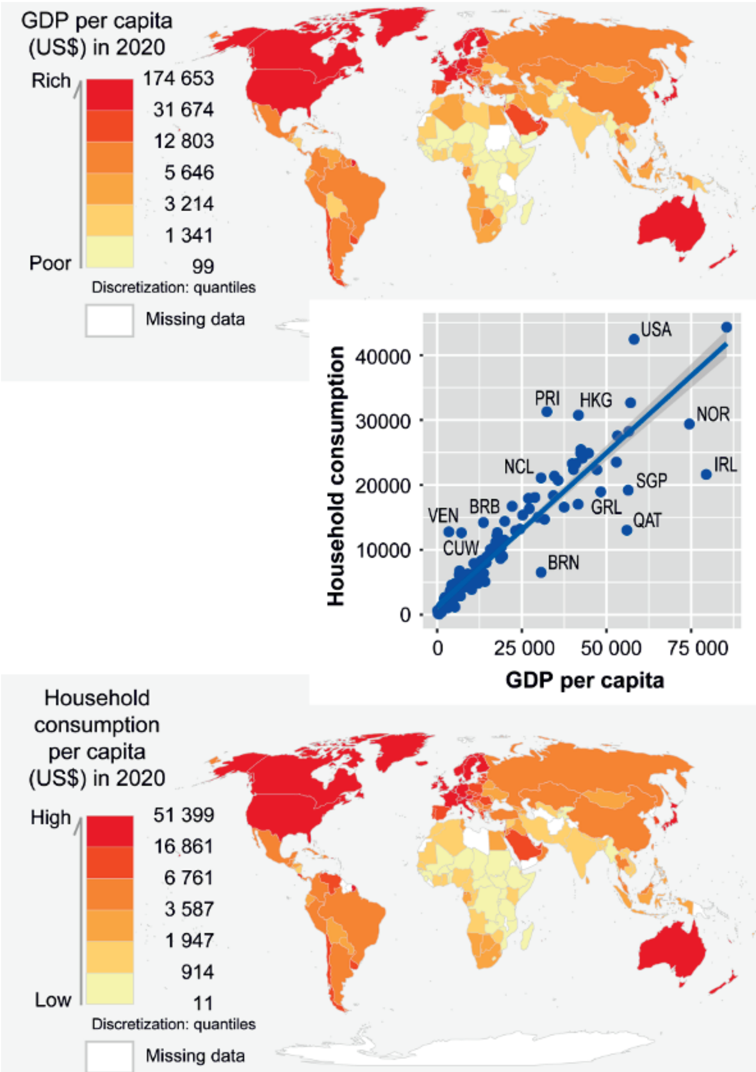


Figure 1.6. *GDP in 2020 and relationship with household consumption (sources: UN, World Bank Eurographics; Software: R (mapsf) J.-B. Bouron, C. Cunty, H. Mathian)*

Identifying deviations from this model is also a source of knowledge. For example, the two maps in Figure 1.6 are very similar. However, if we represent the deviations from this statistical relationship, we show a different geography (Figure 1.7): countries in the same class show the same consumption behavior relative to their GDP, whatever the level. Countries with values close to zero (light colors) are those that "consume" as their GDP would suggest, on average. More interestingly, countries with high positive or negative values (darker values) "consume" differently from what their GDP would suggest. Countries with negative values (in blue) "under-consume" in relation to their GDP/capita, while countries with positive values (in red) "over-consume" in relation to their GDP/capita. The values of these differences constitute a new variable, "de-correlated" from GDP, which can be interpreted as the share of consumption that cannot be explained by GDP.

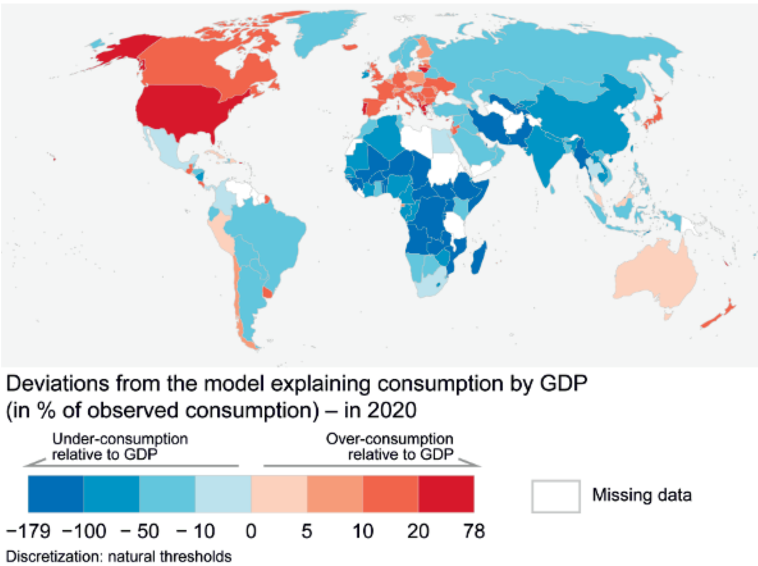


Figure 1.7. *Deviations from the linear model estimating household consumption as a function of GDP (sources: UN, World Bank Eurographics; software: R (mapsf) J.-B. Bouron, C. Cunty, H. Mathian)*

1.4. Describing statistical relationships between several variables

The dialogue between cartographic and graphical visualizations derived from statistical processing presented above enables us, through the multiplicity of viewpoints, to increase our knowledge of the geographical structuring of the phenomenon observed. We propose here to give another illustration based on a question requiring multivariate analysis.

Multivariate data processing is marked by graphics. These include Bertin's matrix experiments (1967), in which the analyst graphically reorganizes the data table, successively and iteratively permuting variables and individuals, to visually highlight the structure of the information. These are also data analysis methods, widely used by geographers since the 1970s–1980s, whose success is due to numerous graphic representations they produce: correlation graphs, factor maps, classification trees, etc. In this way, data analysis methods enable us to synthesize multivariate data, without any a priori assumptions, to identify structures of differentiation on the basis of relationships, to rank them and to categorize data. Applied to geographical data, they enable us to understand the structures of relationships between variables associated with the spatial organizations observed (Sanders 1990).

They are particularly well suited to the very general question presented here (section 1.2). Dividing up the world according to major wealth and development divides means classifying the world's countries according to a number of variables that reflect their level of wealth and development. The countries described by seven variables[5] (see Table 1.1) were then subjected to a hierarchical ascending classification (HAC). HAC is one of the so-called unsupervised classification methods, that is, structures emerge without any prior expert intervention, with the exception of upstream variable selection. The advantage of this method over others (e.g. k-means) is that it produces a classification tree (dendrogram) offering several levels of partitioning. The choice of a partition, based on this dendrogram, will be made according to various considerations: the number of classes, the proportion of information it represents, as well as our ability to interpret the differences or similarities it reveals and to distinguish them on a map. A partition into a large number of classes will bring greater precision, but the differences between classes will be more difficult to interpret.

A six-class partition was our first choice, as it offered a good compromise in terms of knowledge on differentiation factors (71%) and class structuring: too few classes reveal simple shapes; the more classes there are, the more difficult it becomes to communicate their differences and understand their distribution on a map. Once again, the result is interpreted by combining maps and graphs, that is, the statistical result (the classification tree, Figure 1.8) on which the choice of partition is based, and the associated cartographic result (Figure 1.9). By studying statistical profiles of the associated classes at each stage of the tree, we can identify major factors explaining branch separation (Figure 1.8).

---

5 The selection of the seven variables is the result of an analysis exploring two-to-two relationships and consistency choices.

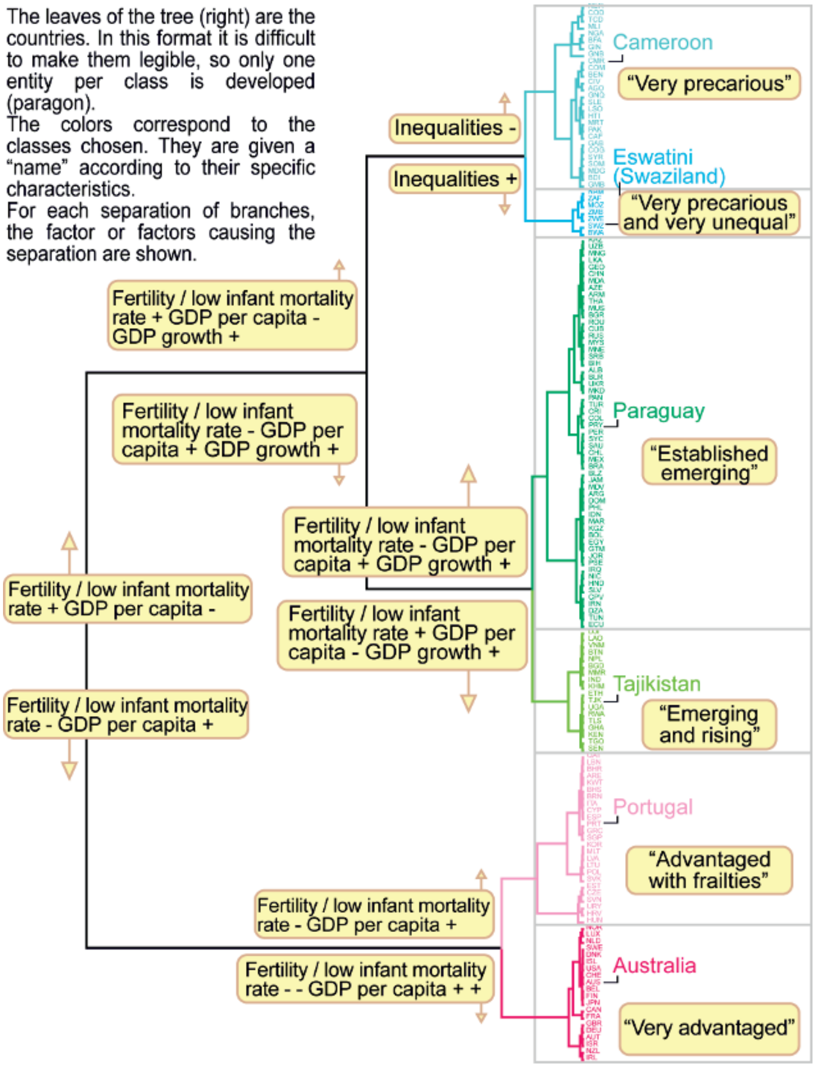


Figure 1.8. *Result of hierarchical classification: the dendrogram traces the structuring of classes from near to near*

REMARKS ON FIGURE 1.8.– *The classification tree is "cut" along a "cut line" running from left to right. The number of associated classes is given by the number of branches the line intersects (sources: UN, World Bank; production: H. Mathian, J.-B. Bouron, Géoconfluences, 2022).*

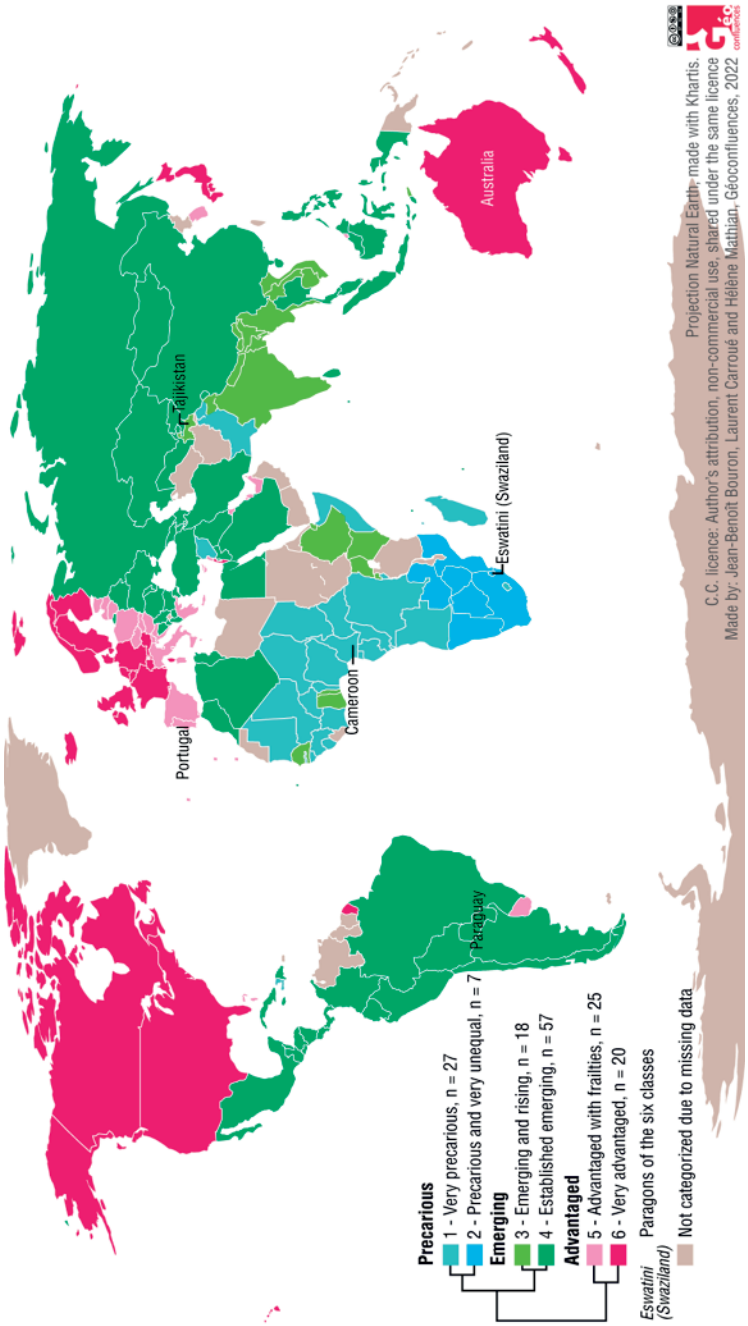


Figure 1.9. *World partitioning into six classes (illustrating 71% of the differences between countries)*

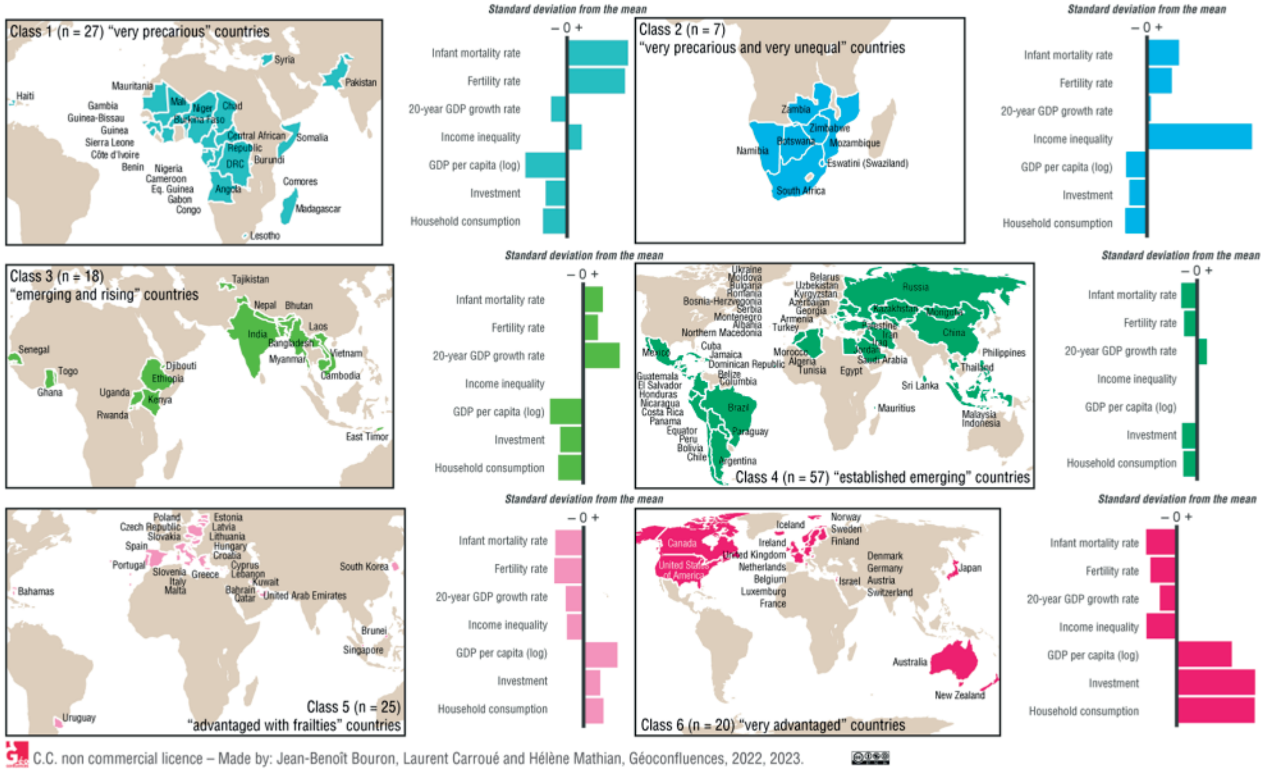


Figure 1.10. *Small-multiple maps of the six country classes (sources: J.-B. Bouron, L. Carroué, H. Mathian, Géoconfluences)*

Following the principle of multiple MacEachren maps, the map in Figure 1.9 can be broken down into six maps, as shown in Figure 1.10. Each map is associated with its average statistical profile relative to the seven variables employed. It is these profiles that allow us to interpret the "texture" of the class (Box 1.2). Isolating each class undoubtedly enables us to better grasp its geographical extent, and the multiplicity or otherwise of its locations. However, it is no longer possible to read the geographical proximities between classes, continuities or discontinuities, and this is what we are trying to understand by understanding social issues that may be linked to these spatial discontinuities. When reading a planisphere, we always come up against the issue of the great variability in the surface area of countries. In our case, several Class 4 countries are among the largest in the world (Russia, China, Brazil, Argentina, Mexico, Indonesia), which accentuates the visual importance of this class, reinforced by the fact that it is the one with the most countries. Conversely, the relatively small size of most of the countries in Class 5 makes it less visible than Class 6, which has five fewer countries, but some of which are very large (Canada, the United States and Australia). At first glance, you might think that all the countries in this class are in Europe, but Israel, Uruguay, Lebanon, Qatar and the United Arab Emirates are also in this class.

Description of classes

Analysis of bar charts (Figure 1.10), which show deviations from the mean for each variable, provides a detailed description of the classes. Based on the six average profiles, three main groups emerge. We present here only broad generalities, and refer to the article by Bouron et al. (2022) for further details:

– the first group of so-called "precarious countries": this group is made up of Classes 1 and 2, corresponding to two groups of poor countries with "very precarious countries" (27 countries in Class 1) and another class that is also precarious and characterized by high inequality (7 "very precarious countries and very unequal" in Class 2);

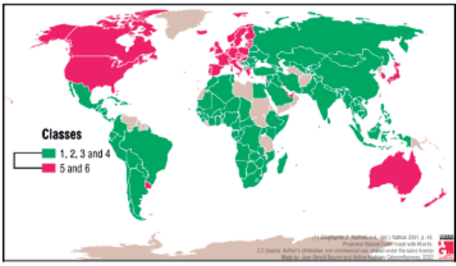
– a second group called "emerging" countries due to their characteristic as being in the process identified as emergent: rising per capita wealth, declining demographic markers of poverty, but continuing high levels of internal inequalities. It is made up of Classes 3 and 4, which represent 46% of the countries referenced in our study. Several economic and demographic giants are included (China, India, Russia, Brazil, Mexico, Indonesia, Malaysia, Thailand, Vietnam, Egypt, Turkey, etc.), accounting for two-thirds of the world's population and the bulk of per capita GDP growth. Class 3 is made up of 18 "emerging and rising countries". Class 4 is by far the most numerous: 57 of the 154 countries for which data are available. These are the "established emerging countries";

– a final group of "advantaged" countries is comprised of Classes 5 and 6. Class 5 comprises 25 countries that we have qualified as "advantaged with frailties" because their indicators are intermediate, between the very advantaged and the "established emerging" countries. Class 6 corresponds to the "very advantaged". These 20 countries have the lowest infant mortality rates, and the highest consumption, investment and GDP per capita values of the six groups.

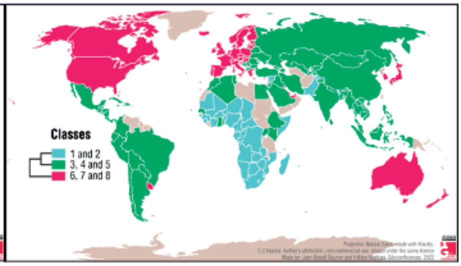
Box 1.1. *Interpretation of classes obtained*

However, it is not uncommon to try to classify a set of entities whose differentiations tend to be organized along a continuum: in particular, when working at aggregated levels such as territories, it is rare to observe phenomena that are structured a priori into classes, which are easy to detect. The classification tree finely reveals the way in which characteristics are hierarchized to differentiate this continuum. In an exploratory approach, these differentiations can be apprehended by declining the sequence of partitions, in keeping with the idea of multiple maps. Figure 1.11 illustrates four partitions derived from the upper part of the tree. These can be read kinetically, as if a ruler was running from left to right across the tree, showing as it goes along the spatial organization revealed by the partition. At each stage, a class splits into two, linked to a new differentiation feature that refines this division, both thematically and spatially. Figure 1.11(a) shows a world split in two. What is interesting about this first partition is that it can be compared to the "North-South" boundary that still appeared on many planispheres up until the end of the 2000s. If we are to believe the seven indicators, this boundary has shifted, with Russia in the "South" and Uruguay in the "North". Admittedly, in detail, the opposition between these two groups identified by Willy Brandt in 1980 has not completely disappeared. Nevertheless, this division into two classes only reflects 39% of differentiations between countries.

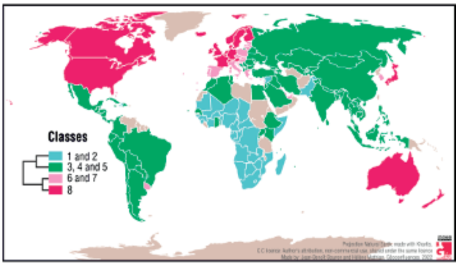
In three classes (Figure 1.11(b)), one group stands out (in blue), almost all of whose members are African states, reflecting its strong regionalization. This distribution into three groups can be seen as a map of emergence: when we subtract the advantaged (in red) and the precarious (in blue), we are left with countries (in green) characterized by a rise in GDP per capita coupled with a rise in inequality, as well as lower fertility and infant mortality rate than those of poorer countries. A breakdown into three classes illustrates 56% of the differentiations. In the next step (Figure 1.11(c)), the group of advantaged countries is subdivided, isolating a sub-group (light pink) on the basis of lower economic indicators but more evenly distributed incomes. This group includes ten former socialist countries that joined EU in 2004, as well as four southern European countries weakened by recent economic crises (Portugal, Spain, Italy and Greece). Finally, the last stage (Figure 1.11(d)) reveals an additional class: a small group of Southern African countries (dark blue) stand out very clearly from the group of "poor" countries (light blue), mainly due to a very accentuated indicator: among the highest income inequalities in the world.



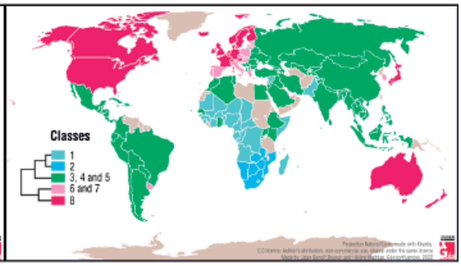
a. Partition into 2 classes
(39% of differentiations)



b. Partition into 3 classes
(56% of differentiations)



c. Partition into 4 classes
(62% of differentiations)



d. Partition into 5 classes
(67% of differentiations)

Figure 1.11. *Differentiation dynamics of successive splits in the hierarchical tree before the chosen partition into six classes*

1.5. Conclusion

This chapter could be considered as a kind of preamble to all other chapters of this book. It lays the foundations for an approach in which cartographic representations are both the product of statistical processing applied to geographic data, and key elements in a process combining exploration, confirmation and description phases. This chapter provides an opportunity to examine some of these stages through applied concrete case studies: the choice of representations and reflection on the association of geometries, the mobilization of statistical methods in an exploratory and descriptive framework and the production of indicators based on modeling. Introducing these insights by working on countries of the world places the approach in a special position, both statistically and geographically, and cartographically. Working on countries encourages us to return to the *identity* of entities, which is not the case when working on a set of sensors, or a set of pixels, or even a set of statistical census grids, such as community subdivisions.

In this approach, a map is therefore the result of numerous iterations alternating statistical processing and visualization, during which multiple choices and transformations are made, which are of course facilitated in interactive environments. However, this approach must be adapted to the geographic question at hand. In all cases, it allows us to enter into a virtuous cycle in which visualizations serve data and the understanding of the phenomenon.

1.6. References

Anselin, L. (1995). Local indicators of spatial association: LISA. *Geographical Analysis*, 27(2), 93–115.

Anselin L. (2003). GeoDa [Online]. Available at: https://geodacenter.github.io/.

Bahoken, F. (2022). Représenter la mondialisation par des flux, le rôle de la distance cartographique perçue. *Mappemonde*, 133 [Online]. Available at: http://journals.openedition.org/mappemonde/7558.

Banos, A. (2005). La voie de l'étonnement : favoriser l'abduction dans les Systèmes d'Information Géographique. In *Apport des SIG à la recherche*, Fotsing, J.-M. (ed.). Presses Universitaires d'Orléans, Orléans.

Béguin, M. and Pumain, D. (1994). *La représentation des données géographiques. Statistique et cartographie*. Armand Colin, Paris.

Bertin, J. (1967). *Sémiologie graphique*. Gauthier-Villars, Paris.

Bouron, J.-B., Carroué, L., Mathian, H. (2022). Représenter et découper le monde : dépasser la limite Nord-Sud pour penser les inégalités de richesse et le développement. *Géoconfluences* [Online]. Available at: https://geoconfluences.ens-lyon.fr/informations-scientifiques/dossiers-thematiques/inegalites/articles/rdecoupage-economique-mondial.

Brandt, W. (1980). *Nord-Sud : un programme de survie. Rapport de la Commission indépendante sur les problèmes de développement international*. Gallimard, Paris.

Brunet, R. and Dollfus, O. (1990). *Mondes nouveaux. Géographie universelle*. Hachette/Reclus, Paris/Montpellier.

Capdepuy, V. (2011). Le monde de l'ONU. Réflexions sur une carte-drapeau. *M@ppemonde*, 102 [Online]. Available at: http://mappemonde-archive.mgm.fr/num30/mois/moi11201.html.

Capdepuy, V. (2013). Le monde en étoile : genèse d'une projection cartographique. *Ma@ppemonde*, 111 [Online]. Available at: http://mappemonde-archive.mgm.fr/num39/lieux/lieux13301.html.

Cauvin, C., Escobar, F., Serradj, A. (2010a). *Thematic Cartography 1. Thematic Cartography and Transformations.* ISTE Ltd, London and John Wiley & Sons, New York.

Cauvin, C., Escobar, F., Serradj, A. (2010b). *Thematic Cartography 2. Cartography and the Impact of the Quantitative Revolution*. ISTE Ltd, London and John Wiley & Sons, New York.

Chancel, L., Piketty, T., Saez, E., Zucman, G. (eds) (2021). Rapport sur les inégalités mondiales 2022. Report, Word Inequality Lab.

Charre, J. (1995). *Statistique et territoire*. RECLUS, Montpellier.

Dumont, G.-F. (2018). Une exception française : son recensement de la population. Quelle méthode ? Quelles insuffisances ? Comment l'améliorer ? *Les Analyses de Populations & Avenir*, 3, 1–226.

Genevois, S. (2018). La projection Equal Earth, un bon compromis ? [Online]. Available at: https://cartonumerique.blogspot.com/2018/11/la-projection-equal-earth.html.

Grataloup, C. (2007). *Géohistoire de la mondialisation. Le temps long du monde.* Armand Colin, Paris.

Lambert, N. and Zanin, C. (2016). *Manuel de cartographie. Principes, méthodes, applications.* Armand Colin, Paris.

Lévy, J. and Lussault, M. (2003). *Dictionnaire de la géographie et de l'espace des sociétés*. Belin, Paris.

Monmonnier, M. (2019). *Comment faire mentir les cartes*. Éditions Autrement, Paris.

Morel, J. (2021). *Les cartes en question. Petit guide pour apprendre à lire et interpréter les cartes*. Éditions Autrement, Paris.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Geo Books, Norwick.

Pumain, D. (2008). Maillage [Online]. Available at: https://hypergeo.eu/maillage/.

Sanders, L. (1990). *L'analyse des données appliquée à la géographie*. RECLUS, Montpellier.

Šavrič, B., Patterson, T., Jenny, B. (2019). The Equal Earth map projection. *International Journal of Geographical Information Science*, 33, 454–465.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.

Unwin, A. (1998). Exploratory spatial data analysis with local statistics. *The Statistician*, 47(3), 415–421.