GIS GEOGRAPHICAL INFORMATION SYSTEMS SERIES

Cartography and the Impact of the Quantitative Revolution

Colette Cauvin
Francisco Escobar and Aziz Serradj

iSTE

WILEY

Cartography and the Impact of the Quantitative Revolution

To Waldo Tobler

who developed the concept of transformation
and opened up for us so many new paths in cartography

To Jean-Claude Müller

who regularly made us realize the benefits
of the latest technologies for cartography

To Henri Reymond

who helped us put these new paths to work,
offering guidance and scientific support to our reasoning

The authors would like to thank the Laboratoire Image et Ville (UMR 7011, CNRS) and all the people who, in one way or another, have helped in the production of this book. We would like to mention in particular, Jean-Philippe Antoni and Hélène Haniotou who have made a tremendous contribution to the creation of the figures in all three volumes, as well as Jimena Martínez who created the website (http://www.geogra.uah.es/carto-thematique-hermes/).

Thematic Cartography

volume 2

Cartography and the Impact of the Quantitative Revolution

Colette Cauvin
Francisco Escobar
Aziz Serradj

iSTE

WILEY

First published 2010 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc. Adapted and updated from two volumes *Cartographie thématique 3 et 4* published 2008 in France by Hermes Science/Lavoisier © LAVOISIER 2008

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Cataloging-in-Publication Data

Cauvin, C.
Thematic cartography and transformations / Colette Cauvin, Francisco Escobar, Aziz Serradj.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84821-109-4 -- ISBN 978-1-84821-110-0 -- ISBN 978-1-84821-111-7 --
ISBN 978-1-84821-112-4
1. Cartography. 2. Visualization. I. Escobar, Francisco. II. Serradj, Aziz. III. Title.
GA108.7.C38 2010
526--dc22
2009048722

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-84821-109-4 (Set of 3 volumes)
ISBN 978-1-84821-111-7 (Volume 2)

Printed and bound in Great Britain by CPI Antony Rowe, Chippenham and Eastbourne.

FSC
Mixed Sources
Product group from well-managed
forests and other controlled sources
Cert no. SGS-COC-2953
www.fsc.org
© 1996 Forest Stewardship Council

Table of Contents

General Introduction

In the first volume of this book on thematic cartography we established the essential elements of making a map and advocated the concept of transformation, introduced by W. Tobler [TOB 61]. A map is thought of as a result of a transformation process. The cartographic reasoning associated with this idea can be adapted for the production of any type of map, regardless of its aim or the study phase for which a given map is intended. In order to be able to make the necessary decisions, a cartographer needs to keep in mind all the different stages of the mapmaking process.

Nevertheless, only those transformations that are indispensable in making a map were described and explained previously. We now turn to the more inventive transformations, to which the second volume is devoted. This volume consists of two large parts, and aims to show the contributions of different manners of processing the attributes, and those representations which are difficult to create without the help of a computer.

The first part concerns the stage T2b, which was not discussed in the first volume of this book. Thus, in this part we address the processing of the attributes *[Z]* and the role of quantitative methods in cartography. Indeed, whilst for a long time cartographers have been superimposing and juxtaposing variables (often making the maps illegible), the inclusion of statistical tools in order to process data before representing them produced a fundamental revolution in the discipline. From now on, maps visualize the results of data processing, which summarize the available information or stress a particular feature of the studied phenomenon. Depending on the phase in the study for which the map is required, and also depending on whether the map is needed for a preliminary exploration or for a verification of a proposed assumption, the data processing can be very basic or very complex. It ranges from simple structuring of a single variable to combining *k* variables or creating a model, which may or may not incorporate the geographic space explicitly.

The second part puts forward the transformations connecting the coordinates *[XYZ]*. The principles of these transformations have been known for a long time, but their application was difficult, if not impossible, without a computer. This part examines the techniques which were long known but later renewed thanks to the computing revolution. These include cartographic transformations of position on the one hand, and 3D representations on the other. The former are more often encountered under the name of anamorphosis. These are original models, often revealing the underlying structure, which is not visible directly. The latter (3D representations) are characterized by the presence of a variable which is expressed vertically. These representations comprise several distinct categories with different meanings: 2.5D, 3D and virtual reality.

It is certain that this second volume will leave unanswered some questions about the future and the new opportunities in cartography. Therefore, the goal of the last volume will be to describe the contributions of new technologies. Although cartographers should always be open to these, it is important to judge them critically in order not to end up with aberrant maps and not to make ill-advised decisions.

PART I

Transformations of Attributes *[Z]* and Use of Quantitative Methods: Generalization and Modeling

Part I

Introduction

As was mentioned in the Introduction, this part will concern the transformation of attributes (T2b), a stage which is too often overlooked and often not entirely used. It is indispensable, however, for achieving a legible and useful map. Once the attributes have been collected, they are transformed into information which can be more or less elaborate depending on the assumptions and the recipient's requirements (see Volume 1, Chapters 2 and 5). Whatever the technique, it is only rarely that this information can be turned directly into a map. The map aim is to reveal the underlying structures relating to the attributes and their locations, to understand the hidden relationships between the data and to discover whether a global or local organization exists with spatio-thematic subsets. Therefore, to achieve these aims it is necessary to proceed with a more or less pronounced abstraction, in order to reveal "the essential features" by transforming the variables either graphically or mathematically.

Various transformations are possible and they can be grouped into three large families corresponding to each of the phases of mapmaking and thus to a particular type of request by the user (Figure I.1). The first family is *description*. It contains all the procedures which aim at describing the data with the aid of statistics and graphical diagrams. This family unequivocally belongs to the domain of Exploratory Spatial Data Analysis (ESDA) pioneered by J. W. Tukey [TUK 77] as Exploratory Data Analysis (EDA) and later changed to ESDA by L. Anselin [ANS 88]. The techniques of this family help the map author to learn about the phenomenon and the characteristics of its variables, in order to avoid inadequate representations and processing later on. These techniques constitute an indispensable step in the mapmaking process, and in recent years they have been completely modernized by the introduction of ESDA.

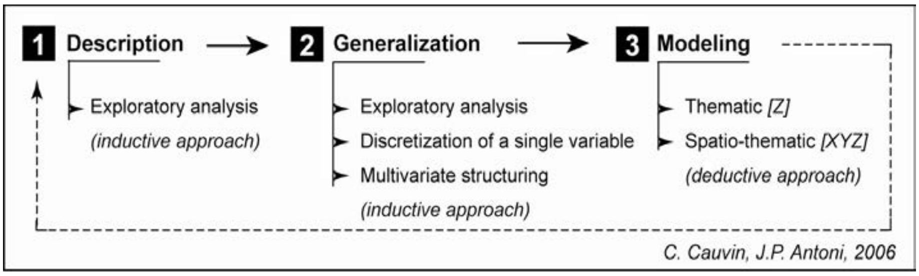


Figure I.1. *Families of the attribute transformations*

The second family is *thematic generalization*. It is used to simplify the data and to reveal the spatio-thematic subsets using techniques which allow the mapping of groups of individuals instead of each and every one. This reduces the number of variables compared to the initial one. In fact, some very specific cases aside, representing all the values and all modalities present in the variables always seems difficult, even though it is now technically possible. It creates a risk of producing a completely illegible and incomprehensible document. Even recently, this point was the subject of many discussions [CRO 95], but some authors, such as W. Tobler [TOB 73c], proposed more exact, unclassed maps, which are legible despite containing all the information [PET 79b]. It is certain that modern-day techniques facilitate the representation of all the values. Interactivity makes it possible to introduce the values into a map progressively, in an increasing or decreasing order. At the exploratory level, such a technique turns out to be very interesting, as testified by the works of O. Klein [KLE 07] on flow representations. Nevertheless, when it comes to presenting the results to a less expert public, classes are indispensable and this point will be considered in detail in Chapter 1.

Regardless of the decision concerning classes, the generalization stage belongs to a time when the assumptions are implicit and we are looking to isolate the spatial characteristics from the thematic phenomenon, by reducing the information. Hence, the processing amounts to a thematic generalization, and its principal goal is to reduce the number of objects (or groups of objects) to represent, and then create, classes. Processing can be graphical or mathematical. This choice depends primarily on the number of variables and on their measurement level. In each case we are trying to create a graphical sign or a "composite" variable obtained after processing which "sums up" the initial information. In light of the fact that all the processing methods lead to the construction of a single "synthetic" attribute, the techniques for creating classes – called discretization – will be considered in Chapter 1. They enable the cartographer to complete the description of the variables within ESDA. The second chapter is devoted to working with multiple data and places greater emphasis on quantitative methods rather than on the graphical solutions abundantly described in a number of works.

The third family is *modeling*. It can be employed only if the map to be made is at an advanced stage in the study of a phenomenon, when the assumptions are explicit and can be subjected to verification. The results of various modeling techniques or those concerning their steps are the subject of cartographic representation. Some models take only the attributes into account. The resulting values are subsequently transferred onto the cartographic support. Others involve both the attributes and the spatial components, producing a map directly. The aims of the map, its recipients and the phase of study also play a fundamental role in the choice of a model. This will be explored in Chapter 3.

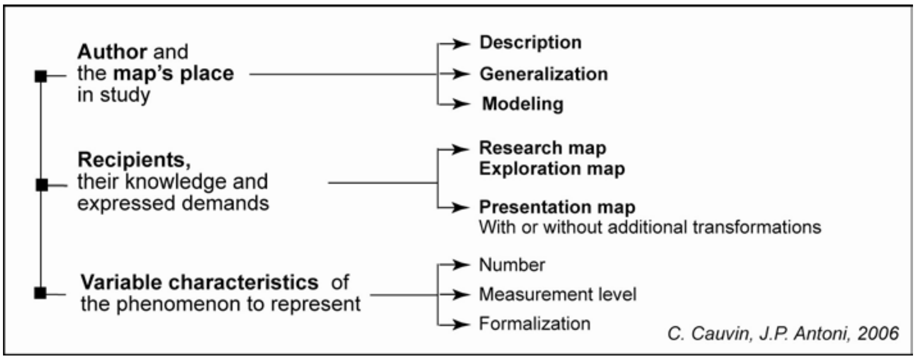

Author and the map's place in study
Description
Generalization
Modeling
Recipients, their knowledge and expressed demands
Research map
Exploration map
Presentation map
With or without additional transformations
Variable characteristics of the phenomenon to represent
Number
Measurement level
Formalization
C. Cauvin, J.P. Antoni, 2006

Figure I.2. *Attribute transformations: the choice criteria*

There are several criteria for choosing a particular attribute transformation technique from the wide range of possibilities (Figure I.2). The *first* concerns the map author and the stage in the study for which the map is required. It implies that the processing is performed within an exploratory, inductive or hypothetic-deductive approach. The *second* has to do with the requirements for the map, hence with the recipients. Depending on their knowledge and their goals, the recipients may need a map for research, exploration, reflection or presentation. The latter may call for further simplification and then successive transformations of the attributes. Finally, the *third criterion* relates to the variables characterizing the represented phenomenon. Their number, level of measurement and formalization will play different roles depending on the phase of the study. It is quite obvious that the general framework of map production is also important. When making choices, we should never forget the place which the map occupies in the scheme developed in Chapter 3 of Volume 1.

Attribute transformations are obviously based on statistics and spatial modeling, using the associated software which has developed alongside the progress of the computer-science revolution. It is in relation to this revolution that the integration of quantitative methods has taken place in cartography, as well as in other sciences (geology, botany, geography, etc.). It has brought about a revitalization of the attribute processing phase within the cartographic process.

Chapter 1

From the Description to the Generalization of an Attribute Variable *Z*

A thematic variable *[Z]* can be represented on a spatial support *[XY]* in various ways and with various visual results, depending on the processing to which the variable is exposed, especially if it is a quantitative variable. The cartographer may decide to generalize the map excessively and, for example, only distinguish the values of *Z* above and below the mean. Conversely, the cartographer may remain closer to the initial data and preserve a very large part of their values. The resulting maps will necessarily be different and the commentaries will be very dissimilar to each other. How best to make a decision? Why prefer one solution to another?

It is questions like these that the present chapter attempts to answer and to suggest an approach and criteria on which cartographers may rely in making their choice. As we wrote in Chapter 5 (Volume 1), thematic data can have several forms and can be obtained from various sources or generated in diverse ways [ROB 98]. The basic information, for example, can be encrypted if it comes from official sources (for instance, INSEE[1]), or from questionnaire surveys, or else if it is a field survey. Therefore, in order to make justifiable and consistent choices it is crucial to:

– Know the characteristics of the variables which are about to be represented. At this stage, exploratory analysis retraces the statistical approaches and constitutes a significant help.

– Choose how to process the variable *[Z]*, that is, choose the discretization mode.

– Validate the choices with the aid of tests, plots or indices.

[1] National Institute of Statistics and Economic Studies in France.

These three facets of transforming the variable *[Z]* are meant to generalize the data and in this way obtain a representation which highlights the important features of the studied phenomenon according to the issue considered.

1.1. Preliminary data analysis: a crucial step

An analysis preceding the representation is essential since it helps the cartographer in making decisions at different levels. It creates the variables whose characteristics can be found with the help of statistics, for as A. von Humboldt wrote in 1811 [FUN 37], "Statistical projections which speak directly to the senses without fatiguing the mind, possess the advantage of fixing the attention on a great number of important facts." Thus, plots present a certain interest, renewed and strengthened with the development of *exploratory data analysis* (EDA).

1.1.1. *From classical description to exploratory data analysis (EDA)*

The statistical description of data supplies the cartographer with a certain number of indices whose meaning it is important to know. It enables comparisons to be made between the variables, and their range and dispersion to be known. But it does not supply a global and immediate vision of their distribution. *Exploratory data analysis* (EDA) is one of the possible ways of deciphering what the data say using mathematics and, most importantly, simple graphical diagrams.

1.1.1.1. *History and relevance*

The idea of data exploration as a useful and irreplaceable statistical technique is attributed to J. W. Tukey [TUK 77]. Data exploration was subsequently elaborated and improved with the developments in mathematical visualization and dynamical graphics to the point of considerably changing the analysis of quantitative [MAC 92] and even qualitative data. The analysis of localized data is a conceptualization in which the central element of the broad investigation process is the connection between the data and the technique, involving the supporting theory and interpretation of the results [ROB 98]. In all cases, as we will see below, the links connecting the theory, data construction, analysis and interpretation are neither simple nor static.

Researchers' and cartographers' analysis of gathered data has always been associated with the application of analytical techniques such as plots, diagrams, and the like. The use of EDA revitalized this approach and presented a double benefit. On the one hand, it facilitates learning of the data analysis and statistics; on the other hand it raises important questions about the formal representation of data and thus leads to the emergence of new working hypotheses.

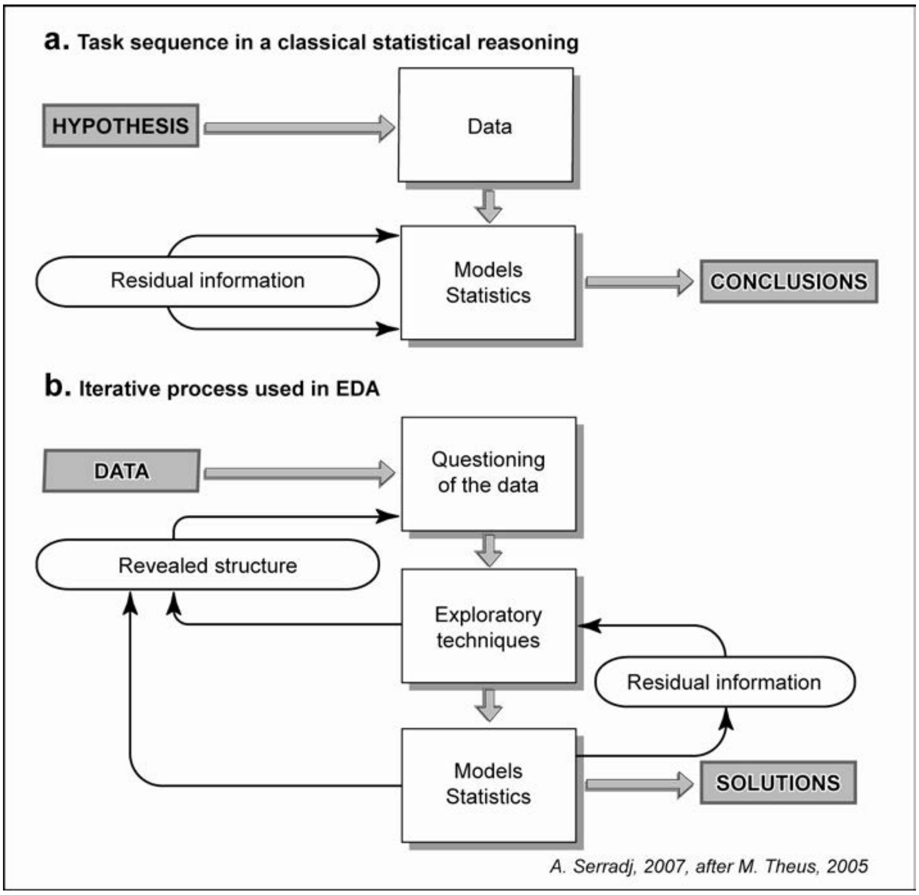
1.1.1.2. *Approach and definition*

Research work starts with a theory and ends with the presentation of results. Although at a glance its steps are clearly distinct or sequential, in practice it is rarely so. In its broadest sense, the analysis may include:

– initial evaluation of the data;

– creation or construction of the data;

– application of specific analytical techniques to probe the data.



A. Serradj, 2007, after M. Theus, 2005

Figure 1.1. *Linear statistical reasoning and EDA iterative process*

While the classic statistical reasoning is based on linear thinking (meaning that the tasks are performed in a sequential manner), EDA adopts an iterative process at several analytical levels. This is shown in Figures 1.1a and 1.1b, produced based on the proposals of M. Theus [THE 05]. This author stresses that EDA does not begin with a predetermined set of tools and that graphical exploration may be of interest for some data but not for all.

Naturally, geographers already used simple graphical and statistical methods before the *quantitative revolution*. But it was only recently that they began to use these methods widely. The methods are grouped under the name *exploratory data analysis* [BAR 79], or EDA. Instead of following the usual assumptions this approach to data analysis is restricted to finding a model which explains the data behavior. It is a more direct approach which helps to reveal the underlying structure of the data and their model. EDA is not a simple collection of techniques but a philosophy on the subject of dissecting a data series by asking the following questions: what are we looking for? What do we see? How can we interpret it? EDA achieves its goal using a broad range of techniques called “statistical plots” which can be interpreted to open new research directions and not just to illustrate the numerical statistical results as in the traditional approach.

In the original version of EDA the geographic component was not directly integrated into the exploratory analysis. With the advent of geographic information systems, EDA evolved towards exploratory spatial data analysis (ESDA). As M. Theus mentions [THE 05], ESDA is a natural extension of EDA in the way it treats thinking and methods concerning geographic problems. ESDA experienced rapid development after EDA was implemented on computers at the end of the 1980s. Among others the works of M. S. Monmonier [MON 88, MON 89] show the usefulness of combining scatterplots[2] with interactivity. In fact, selecting a set of points by scatterplot brushing has the advantage of automatically updating different plots and thus highlighting the correlations that exist between pairs of variables. Thus, in one example given by M. S. Monmonier [MON 89], it becomes clear that some common *a priori* regarding the cause and effect relationships between thematic variables have to be revised. It seemed natural and logical to assume that in each of the fifty United States of America the proportion of households with a cable TV connection depends on the household income and on the proportion of urban population within the state. The use of scatterplot brushing shows that this is not so. It turns out that the lowest cable-connection rate is in the most urbanized states with a high per capita income, such as Maryland and New York. The author explains that the delay in the penetration of cable television is certainly related to local restrictions and the huge capital requirements. In this category we also find the Midwest states where dispersed farms are not easily served by cable systems. The map obtained from this analysis suggests the cartographer or researcher needs to

[2] A plot combining two thematic variables.

look into some additional factors. This example clearly shows that interactive statistical plots bring questions to the surface rather than giving definite answers.

1.1.1.3. *EDA: predominance of graphical techniques*

The majority of EDA techniques are graphical, accompanied by some quantitative (statistical) methods. The reason for the importance of plots is related to the fact that the main task of EDA and ESDA is the impartial exploration of plots which give the analyst an incomparable power of discovery. Not only do the data reveal their structural secrets, but also the inherent data elements – often previously unsuspected – become visible. Combined with our natural capacity for form recognition, plots offer us an unparalleled power for extracting information.

Graphical techniques used in EDA and ESDA are often rather elementary. Variables can be represented by simple, cumulative or two-dimensional histograms. The usual statistical parameters are added to this: means and medians, standard deviations, coefficients of variation, etc. Only the "boxplot" (also known as a "box-and-whisker plot") is really specific to EDA. When these various representations are united in the same document they allow the optimal use of our shape recognition to extract information concerning the data characteristics.

1.1.2. *Exploratory data analysis and graphical representations*[3]

The purpose of EDA in analyzing the data to be represented is to synthesize, structure and summarize the information contained in the data. In this way it highlights the properties of the thematic attributes and suggests hypotheses.

A long time ago, Horace (65 – 8 BC) said that what we hear excites us less than what we see. Humankind knows just how true this is. The graphical representation of ideas is one of the most ancient and universal features of human activity. The oldest known language consists of ideographic drawings left to us by cave dwellers. The writings of the Egyptian, Babylonian and Maya civilizations are to a large extent pictorial symbols and hieroglyphs. The American Indians used pictorial methods to communicate their thoughts and ideas. It is true that the pictographic techniques of cavemen and the hieroglyphic writing of the Egyptians disappeared because of their inferiority, or at least ceased to be the sole means of communicating

[3] The origin of systematic data representation and analysis is attributed to W. Playfair (1759 - 1823), who utilized "bar charts, time-series plots, proportional-symbol displays and pie charts" (or circular diagrams). Subsequently, graphical representations were employed by C. J. Minard and E. J. Marey in the 19th century. In the 20th century, however, this form of analysis was overshadowed by the development of mathematical statistics.

ideas. Nevertheless, drawings persisted and graphical representations can be found in any age, in one form or another.

In the last 30 years, the use of visualization materials to present ideas has increased greatly. Nowhere is this trend better illustrated than in statistics, where experts have developed a very widespread use of plots. They are so important that we could say that the graphical method is quickly becoming the universal language [FUN 37].

Despite the robustness and variety of statistical tests and computational techniques, the power of graphical representation as an analysis tool for supporting a hypothesis or explaining a phenomenon cannot be denied. A deep analysis based on a graphical representation is often considered only as preliminary. But it enables us to reveal important information and trends with a higher accuracy than a simple numerical-data table.

As we mentioned earlier, once the numerical data are collected the first step is to look for their characteristics. For a cartographer this preliminary investigation may mean not only examining the spatial distribution of the data but also representing the data graphically. Graphical processing became a “device for showing the obvious to the ignorant” until the 1970s when J. W. Tukey revived cartographers’ interest in graphical representations by developing new forms of data presentation and by computer technology which facilitated this kind of representation [ROB 98].

Thus, it becomes clear that visualization is a permanent component of EDA. It is of help in the process of detecting the properties of data, which remains the fundamental goal of EDA. Thanks to certain software, a large number of data graphics can be produced. They are used as simple finished products intended in theory to communicate the identified characteristics of processed data to novices. They are used quite frequently, and some of these graphical representations will be described in the sections below in order to illustrate their importance in data visualization and to show those data characteristics which it is possible to extract.

We will see how useful different graphical representations are for data which are originally in the form of statistical tables. Obviously, reading a table of numbers does not give a clear and quick visualization. Nor does it let us understand the relative positions of the presented values or see the data set as a whole. Cartographers are therefore led to use particular representation types where they can translate the “statistical magnitudes” into “geometrical figures” which are much more evocative to the reader [TAV 83].

1.1.2.1. *Non-mathematical graphical representations: linear diagrams*

When the number of observations is not very large they are easy to represent by a point or a mark on a normed linear axis. This representation gives a practical means of examining the dispersion of two or more sets of data on either side of the same axis, and also to compare them. Figure 1.2 shows this using the data from Luxembourg[4].

1.1.2.2. *Common graphical representations*

Conventional graphical representations help to visualize the form of a statistical distribution of a quantitative variable. They are applied to the quantitative data prepared in the form of a table with modalities grouped into classes.

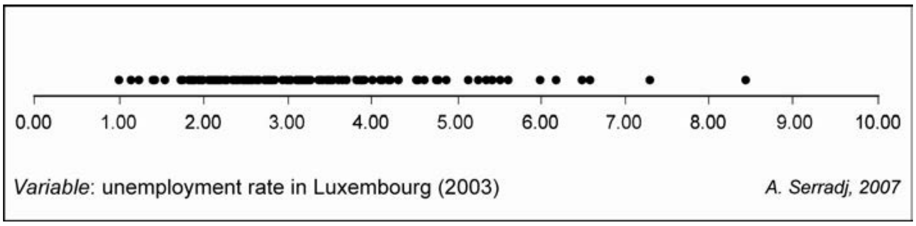


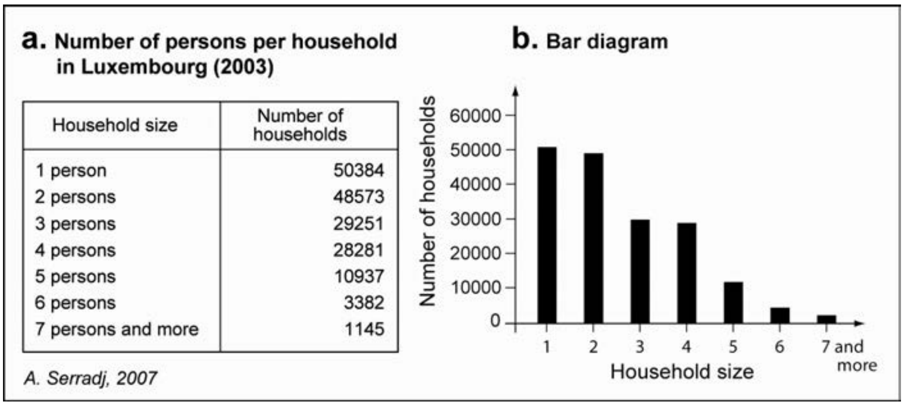
Figure 1.2. *A line plot*

1.1.2.2.1. Bar plots

The study of the spread of values in a statistical series often starts with a plot of the distribution of the data, as shown in Figure 1.3. The plot normally takes on one of the following forms: a bar plot, a histogram or a plot (smooth or not) of the frequencies or the cumulative frequency. If the data consists entirely of integer – and therefore discrete – values, the bar plot illustrates the frequency of a given value. Thus, the frequency bar-plot is a graphical representation of the distribution of the values of a variable. The horizontal axis corresponds to different discrete values (or modalities of the variable) and the vertical axis shows the total number of occurrences (or the frequency) of each modality. For each value Z_i of the variable there is a vertical segment (a bar) whose length is proportional to the number of occurrences n_i, or to the frequency f_i of the value Z_i[5].

[4] All the graphical representations in this chapter are constructed with the statistical data from Luxembourg.

[5] It is important to mention that n_i represents the absolute frequency while f_i stands for the relative frequency.



a. Number of persons per household in Luxembourg (2003)

Household size	Number of households
1 person	50384
2 persons	48573
3 persons	29251
4 persons	28281
5 persons	10937
6 persons	3382
7 persons and more	1145

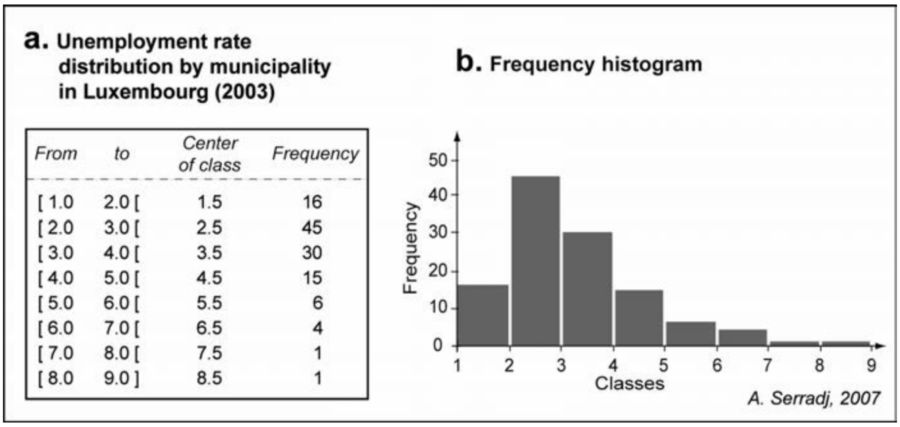
A. Serradj, 2007

Figure 1.3. *A bar plot*

1.1.2.2.2. Histograms

Histograms are employed to represent continuous data such as altitude, age, agricultural productivity, precipitation, etc. These data can take any value within a given interval on a continuous scale. In order to construct a histogram we need to break the data values down into classes. The difference between the upper and lower limits of a class is called the class interval, or class size. It can be varied to show the specifics of the distribution. The number of classes should be chosen with the desired level of detail kept in mind. The higher the number of classes, the more details of the distribution will be visible. Conversely, if the number of classes is small the distribution will suffer a strong generalization and its fine details will not be observable.

A histogram is a set of rectangles arranged side-by-side. The base of each rectangle corresponds to the class interval. Typically classes in the same distribution have equal intervals. The case of different class sizes will be considered further on. The rectangle constructed for each class has an area proportional to the total number of values in this class. The total area of the histogram is therefore equal to the sum of all the rectangles. This area corresponds to the total number of values in the studied data set. For example, Figure 1.4a shows the distribution of unemployment rates by municipality in Luxembourg. By dividing it into ten classes we obtain the histogram shown in Figure 1.4b.



a. Unemployment rate distribution by municipality in Luxembourg (2003)

From	to	Center of class	Frequency
[1.0	2.0 [	1.5	16
[2.0	3.0 [	2.5	45
[3.0	4.0 [	3.5	30
[4.0	5.0 [	4.5	15
[5.0	6.0 [	5.5	6
[6.0	7.0 [	6.5	4
[7.0	8.0 [	7.5	1
[8.0	9.0]	8.5	1

b. Frequency histogram

A. Serradj, 2007

Figure 1.4. *Dividing the data into classes of equal size and the corresponding histogram*

Proportionality is the essential feature of histograms. It makes their construction a little more complex if the class intervals vary. In fact, it often happens that the distribution of a studied phenomenon makes no sense unless it is divided into "useful" classes, in other words classes which are meaningful to the user. There are many examples of this: road slope, size of agricultural fields, household income, city size or rent in Luxembourg. We will use the last example for an illustration. Given that the areas of the rectangles are proportional to the frequencies, the histogram in Figure 1.4 respects this proportionality. In fact, if we consult the table in Figure 1.5a we will see that the frequencies in classes 1 and 4 are almost identical and that the corresponding areas in the "correct" histogram in Figure 1.5c are also very similar, while those in the "incorrect" histogram in Figure 1.5b are not. Nevertheless, it is worth mentioning that the frequency axis in the "correct" histogram requires special attention now that the correction has been performed. This axis should be removed and replaced by the actual class frequencies written underneath each rectangle, as shown in Figure 1.5d.

1.1.2.2.3. Cumulative frequency curves

A cumulative frequency curve represents the distribution function *F(z)* of a continuous variable *[Z]*. This function is defined for each value of *[Z]*. It is constant within each interval separating two consecutive possible values. Thus, it allows us to determine the frequency density in a given interval [CAL 73]. The convenience of cumulative frequency curves is in knowing "how many individuals have a character value below or above a certain threshold". If we plot the relative cumulative frequencies, they will range between 0 and 1. Hence, they have the meaning of probabilities.

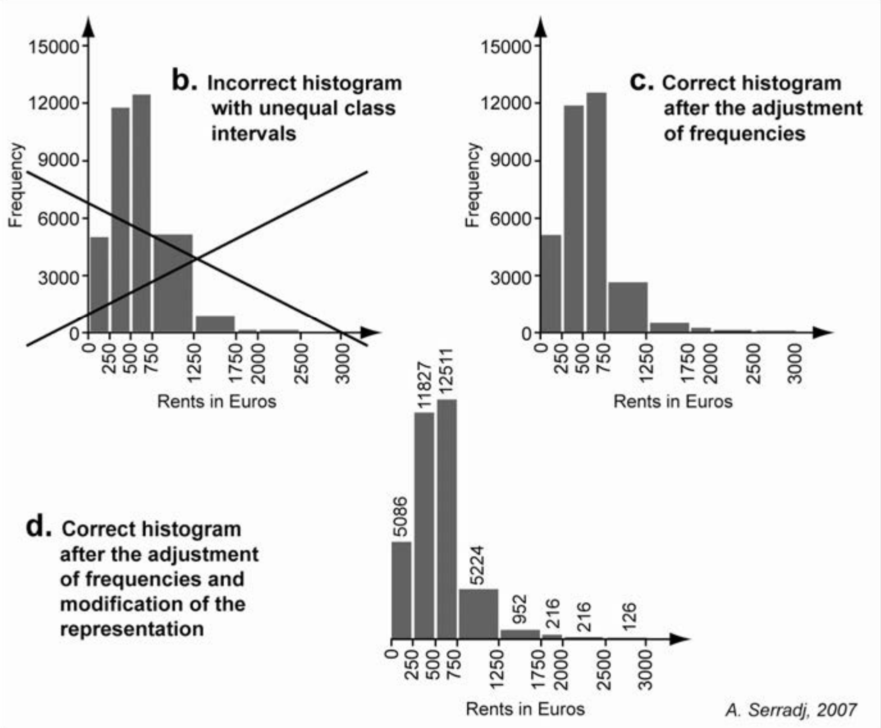
a. Preparation table

	Class limits		Statistical parameters			
Class	Min rent in €	Max rent in €	Frequency (f)	Class interval (e)	Frequency correction coefficient (250* / e) = (ccf)	Adjusted frequency (f x ccf)
1	0	249	5086	250	1	5086
2	250	499	11827	250	1	11827
3	500	749	12511	250	1	12511
4	750	1249	5224	500	0,5	2612
5	1250	1749	952	500	0,5	476
6	1750	1999	216	250	1	216
7	2000	2499	216	500	0,5	108
8	2500	2999**	126	500	0,5	63

* Net number of the histogram

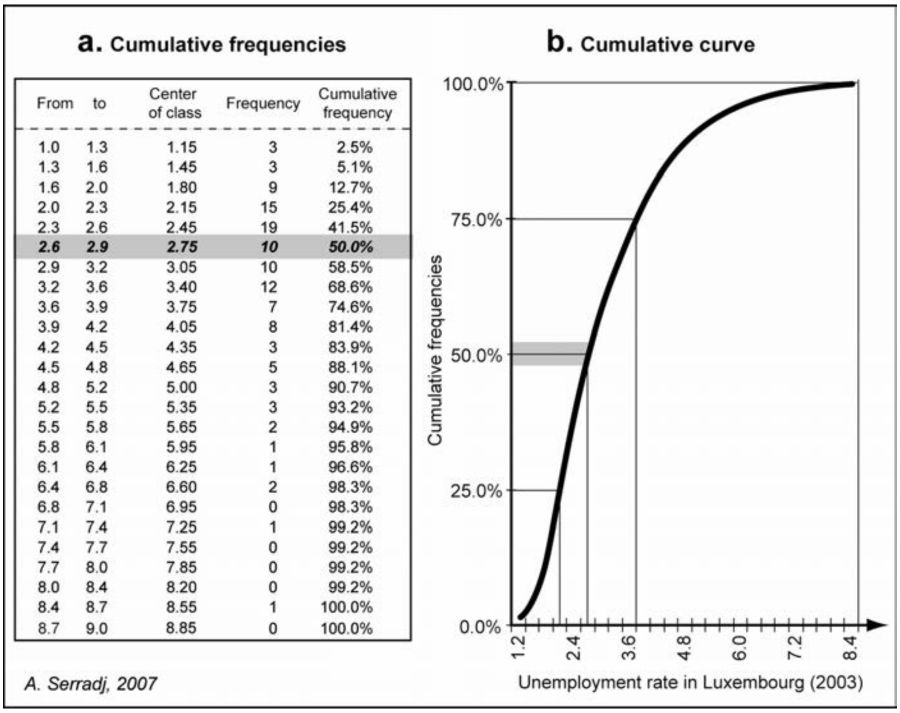
** For practical reasons, we limited this class by 2999 although it was indeterminate (2500 € and higher)

Rent data in Luxembourg (2001)



A. Serradj, 2007

Figure 1.5. *Construction of a histogram*



a. Cumulative frequencies

From	to	Center of class	Frequency	Cumulative frequency
1.0	1.3	1.15	3	2.5%
1.3	1.6	1.45	3	5.1%
1.6	2.0	1.80	9	12.7%
2.0	2.3	2.15	15	25.4%
2.3	2.6	2.45	19	41.5%
2.6	***2.9***	***2.75***	***10***	***50.0%***
2.9	3.2	3.05	10	58.5%
3.2	3.6	3.40	12	68.6%
3.6	3.9	3.75	7	74.6%
3.9	4.2	4.05	8	81.4%
4.2	4.5	4.35	3	83.9%
4.5	4.8	4.65	5	88.1%
4.8	5.2	5.00	3	90.7%
5.2	5.5	5.35	3	93.2%
5.5	5.8	5.65	2	94.9%
5.8	6.1	5.95	1	95.8%
6.1	6.4	6.25	1	96.6%
6.4	6.8	6.60	2	98.3%
6.8	7.1	6.95	0	98.3%
7.1	7.4	7.25	1	99.2%
7.4	7.7	7.55	0	99.2%
7.7	8.0	7.85	0	99.2%
8.0	8.4	8.20	0	99.2%
8.4	8.7	8.55	1	100.0%
8.7	9.0	8.85	0	100.0%

A. Serradj, 2007

Figure 1.6. *A cumulative frequency curve*

In fact, if a relative cumulative frequency equals 0.5 we can say that there is an equal chance of finding an individual with a value less than or equal to the value which corresponds to 0.5. As shown in Figure 1.6, the value for which the cumulative frequency equals 50% is 2.8. It corresponds to the center of the sixth class. Therefore, there exists a 50% probability of finding a municipality in Luxembourg where the unemployment rate is less than or equal to 2.8%.

1.1.2.3. *"Boxplot": a specific representation of EDA*

The idea of representing the central tendency and the dispersion of a statistical data set by "box and whiskers" was stated by J.W. Tukey in 1977. The statistical parameters for graphical representation are easy to obtain from the data set once the values are ranked (in increasing or decreasing order). We can extract the extreme values (minimum and maximum), the median[6] and the first and third quartiles. These parameters can then be represented in a simplified "box and whiskers" form, as shown in Figure 1.7b1. The median together with the first and third quartiles form the "box", while the extreme values (minimum and maximum) form the "whiskers". A more elaborate form of this representation requires the calculation of the hinge values, which become the "whiskers" (Figure 1.7b2). The calculation of the hinge values is done by subtracting one and a half times' the value of the interquartile range (Q3-Q1) from the first quartile (for the lower hinge value) or by adding the same quantity to the third quartile (for the upper hinge value). The observations located outside the hinge values are considered as exceptional values. We can note that for large data sets, deciles or even centiles are used instead of quartiles, depending on the number of observations.

Representations of this type allow us to "see at a glance" the features of the distribution (symmetry, asymmetry or dispersion, among others) thanks to the positioning of the characteristic statistical values. The summary of five numbers in Figure 1.7a shows this. What is more, these representations help to compare different distributions in time and/or space (Figure 1.7c).

G. M. Robinson notes [ROB 98] that this form of representation allows a better comparison of the characteristics of various statistical data sets. D. Sibley [SIB 90], cited in [ROB 98], improves the comparison even further by modifying the technique of the "boxplot" representation. The modification consists of centering the data with respect to their median value. This operation shifts the values and their distribution around the median, so that the median value is equal to zero. The operation of centering the values also eliminates size effects. If the centered values are divided by their interquartile range we obtain standardized values which are very convenient for comparing several data sets with each other (Figure 1.7d).

The various graphical representations described above explore the data characteristics, such as the shape of their distribution, which is essential for the purposes of cartography. However, we should not forget that these plots are suited to variables with a quantitative level of measurement – a particular point which leads to serious constraints in their representation.

[6] In fact, the median value is used more than the arithmetic mean because it is considered as a reliable central tendency parameter or a "resistant" value to use the terminology of G. M. Robinson [ROB 98]. Unlike the mean, the median is not influenced by the extreme and exceptional values in a statistical set.