DATA VISUALISATION IN OFFICIAL STATISTICS

Enara Galbete Ahechu

EUSKAL ESTATISTIKA ERAKUNDEA
BASQUE STATISTICS INSTITUTE

Donostia - San Sebastián, 1
01010 VITORIA - GASTEIZ
Tel.: 945 01 75 00
Fax: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Presentation

One of Eustat's aims is to disseminate statistical information in a clear and understandable way. In order to do this, it is making an effort researching into the field of the visualization of statistical data, so that they can be analyzed and managed dynamically and interactively by the users.

Eustat organized the 23rd International Statistics Seminar in 2010, on the subject of "Examining Data with Dynamic Interactive Graphics" and it made a call for a 2-year training and research grant in the field of visualization of statistical data.

This publication shows the research conducted under this grant and provides helpful tools for users who are interested in learning and using different types of data visualization.

This document is divided into three sections. The first section deals with certain concepts and criteria that should be accounted before making a graph. In the second section there is a catalogue of static and dynamic graphs. Finally, the third section explains the procedures and applications used to make the graphs selected for publication on the Eustat website and the ones chosen for production processes.

Vitoria-Gasteiz, January 2013.

Josu Iradi Arrieta

Director General of EUSTAT

Contents

Chapter
1

Introduction

The data in this Technical Handbook are the outcome of the work on data visualization that was carried out under the grant awarded in 2010 by the Basque Statistics Institute for training and research in the field of statistical and mathematical methodologies.

A 2-year training and research grant was announced in the field of statistical and mathematical methods of producing official statistics, with special emphasis on the visualization of statistical data.

This document is divided into the following sections:

The first chapter gives an introduction and describes the goals that led to this technical notebook.

The second chapter defines the criteria and concepts that should be taken into consideration before making a graph based on a dataset.

The third chapter presents a classification of graphs designed to demonstrate different types of data.

The fourth chapter is about the graphs that are automated for use by EUSTAT. Some are static graphs for production; others are dynamic and interactive graphs designed to be disseminated.

The last chapter gives a series of conclusions related to data visualization methods and techniques.

I would like to acknowledge the work accomplished by the members of Eustat's Department of Innovation in Methodologies and R&D, and by all the members of the visualization group. I would also like to thank Yosu Yurramendi Mendizabal, of the Department of Computational Sciences and Artificial Intelligence of the University of the Basque County, for his advice and assistance, and, in general, the kindness of all of Eustat's employees.

KEYWORDS: static graphs, dynamic graphs, interactive graphs

Chapter
2

Methodology

EUSTAT has a multitude of data from many different sources. There are many ways of analysing and presenting data. In turn, graphs can be used in different ways. Therefore, the main objective of data visualization is to use graphs of varying levels of difficulty and types to analyse, summarise and represent data at first glance.

A type of visualization is allocated to each type of data, providing an opportunity to extract information graphically and intuitively, depending on the level of difficulty of the data and kind of graph concerned.

Types of data

EUSTAT has two types of data: The data on the website, which are available in PC-AXIS format; and microdata files.

Before they can be represented graphically, the data must be classified according to two criteria:

1. **NATURE OF THE DATA:**

Depending on their nature, that is, on the number of variables they have, data can be classified into five groups:

i. **Univariate data.**

ii. **Bivariate data.**

iii. **Multivariate data.** Data with three or more variables

iv. **Time series.** These are sequences of values, observations or data, measured at specific points in time, placed in chronological order, and distributed uniformly, as a rule.

v. **Spatial data.** Variables related to specific spatial locations.

2. **ORIGIN OF THE DATA:**

Data can be divided into two types, depending on their origin:

i. **Frequency tables.** A frequency is assigned to each datum.

ii. **Microdata.** Each individual has associated a list of variables.

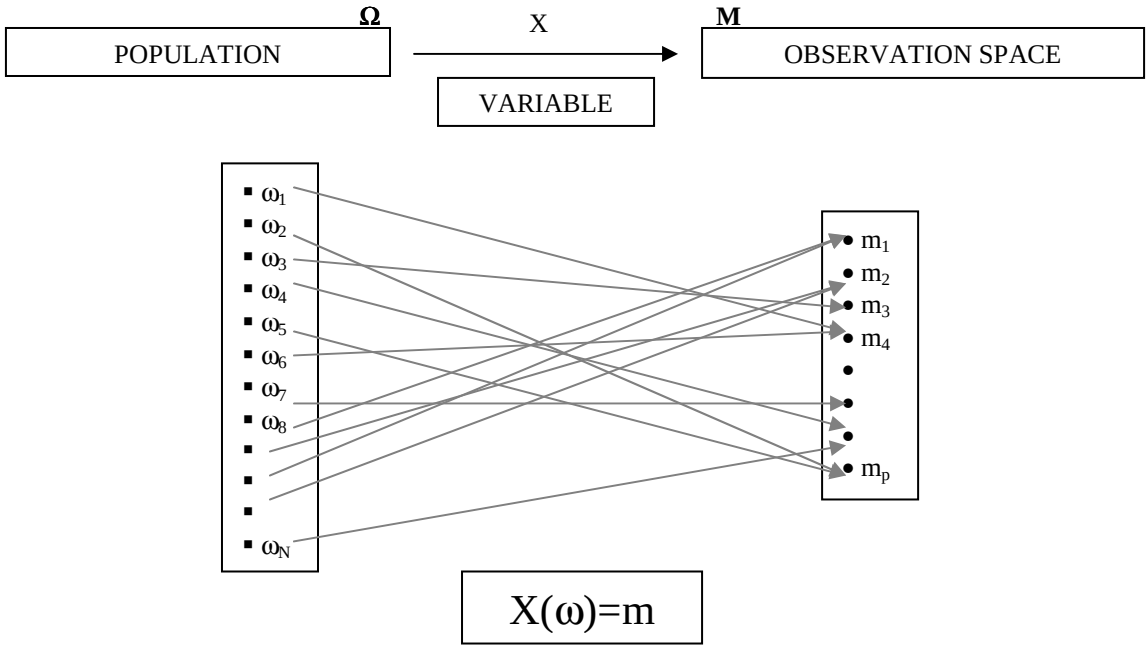
Conceptualization

There are three basic concepts: population, observation space, and variable.

Population: The set of elements that are to be analysed statistically.

Observation space: The set potential outcomes of a randomly conducted experiment.

Variable: The application existing between the population and the observation space.



The population and the observation space may belong to different types. This aspect should be accounted for when constructing a graph, because the nature of the variables also depends on the types.

Population

Different types of **population** exist, depending on their **nature, size and structure.**

1. Classification according to **nature**:

   i. **Municipal census.** They mean the number of individuals that make up a statistical population. Basically, they consist in using different methods of numbering or administrative information to measure the total number of individuals in a statistical population.

   ii. **Sample.** A sample is a population subset, obtained using sampling techniques, that represents the population as a whole.

2. Classification according to **size**:

    i. **Discrete.** A population that has a finite or infinite countable number of elements.

        - **Finite.** A population that has a finite number of elements.
        - **Infinite.** A population that has an infinite countable number of elements.

    ii. **Continuous.** A population whose random variable measures a continuous characteristic.

3. Classification according to **structure**:

    The elements of the population, the units of the set, can be related or not. According to the type of the relation the population can have different structures.

    i. **No structure.** When the elements of the population are not related. This is the most common case in the classical statistics (For example, *when doing a sample it is assumed that the elements, i.e. the people, are not related.*).

    ii. **Sequential structure.** When there is defined a total order among the elements of the population. When a particular element of a process is measured and when those measures cause something to it. (For example, *the step by step therapy of a patient, the steps of the study method of a student…*).

    iii. **Lineal structure.** The most common case is a time series, the units are one after another, and besides, in this case a unit is defined.

    iv. **Map.** In two dimensions (or more) a graph or map defines a structure (the vertex is the population, and the edges define the relationship).

    v. **Plane.** In two dimensions, and when a unit is defined, the population unit is situated in the plane R2. (For example, the pixels of a picture form a graph and besides there is defined a measurement among the pixels, R2).

Observation space:

Different types of **observation spaces** exist, depending on their nature, size and structure.

1. Classification according to **nature**:

    i. **Quantitative.** An observation space that consists solely of quantitative basic events.

    ii. **Qualitative.** An observation space that consists solely of qualitative basic events.

2. Classification according to **size**:

    i. **Discrete.** Said of an observation space that counts on a finite or infinite countable number of basic events.

        - **Finite.** An observation space that counts on a finite number of observations.

        - **Infinite.** An observation space that counts on an infinite countable number of observations.

    ii. **Continuous.** An observation space that counts on an infinite uncountable number of basic events.

3. Classification according to **structure**:

    i. **Nominal.** An observation space made up of observations with no order in relation to each other and that do not obey a hierarchy.

    ii. **Ordinal.** An observation space made up of observations that can take different ordered values, originating from a predetermined natural progression, sequence or scale.

        - **Total.** When the observation space is a set that is completely ordered; that is, when any pair of elements in the set can be compared with each other.

        - **Partial.** When the observation space is a set with a partial order; that is when, in several pairs of elements of a set, one of the elements of the pair follows the other. However, not all the pairs of elements in the set should be related to each other.

    iii. **Numeral.** An observation space made up of numeral observations.

Nature **Structure**	Qualitative	Quantitative
Nominal	√	
Partial order	√	
Total order	√	√
Numeral	*	√

Finite **size**

* $\{m_1, m_2\} \leftrightarrow \{0,1\}$

Variables

The nature of a variable is based on the nature, size and structure of its **observation space**.

The following types of variables can be distinguished, depending on their observation space:

1. **Qualitative.** Variables that reflect different qualities, characteristics or modalities. Each modality is called an attribute or category, and the measurement consists in classifying said attributes.

    i. **Nominal.** Variables that take values that have no order or hierarchy with regards to each other (e.g. *place of birth*).

    ii. **Ordinal.** Variables that can take different ordered values of a natural progression, sequence or preset scale (e.g. *a variable with the following modalities: small, medium, large*).

2. **Quantitative** Those that can be expressed by numbers (e.g. *age, size of the family unit, floor space of the dwelling*, etc.). The results of the experiments can be quantified.

    i. **Discrete.** The variables only take whole values (e.g. *age, size of the family unit*, etc.).

    ii. **Continuous.** They can take any value within an interval, and therefore they can have decimals (e.g. *floor space of the dwelling*).

Criteria to take into consideration before representing a graph

Certain criteria should be accounted for in order to make a "good" graph and obtain as much information as possible from it. W. S. Cleveland, for instance, wrote in his book, The Elements of Graphing Data (1985):

- Unobstructed view or perspective:

    Data: Leave aside anything that is superfluous.

    Use clear, evident graphic elements to show data.

    Use the right pair of scales for each variable. The data space will be the inside of a right angle formed by the right scales. Place the tick marks outside of the data space.

    Do not clutter up the data space.

    Do not put too many tick marks on the scales.

Use a proper reference line when a value must be seen across the entire graph, but do not let the line interfere with the data.

Do not mix the data labels in the space for quantitative data.

Avoid putting notes, keys and tick marks in the data space. Put keys and tick marks outside of the data space and notes in the legend or text.

Drawn signs should be clearly visible when they are superimposed.

Datasets should be perceivable at first glance when they are superimposed.

The clarity of a display should not be lost when a graph is made smaller or animated.

- Clear understanding

  Show the main conclusions graphically. Make information legends understandable.

  Give clear explanations of the meaning of error bars that show the variability of data (standard deviation, standard error, confidence intervals).

  Scale labels should coincide with tick marks when drawing a variable's logarithms.

  Correct mistakes in the graphs.

  Keep graphs uncluttered.

- Scales

  Choose the space between tick marks so they will include the interval of all or almost all the data.

  Maintain scale reductions and select the scale so it comprises the entire dataset.

  Sometimes it may be useful to use two correct straight lines for the same variable to show two different scales.

  Select the appropriate scales when two graphs need to be compared to each other.

  Do not insist on showing zero on scales that represent a magnitude.

  Use a logarithmic scale when it is important to understand the percentage variance in the multiplier factor or correct bias in a large range of values.

  Showing data on a logarithmic scale may make it easier to differentiate the data.

  Do not cut scales unnecessarily. If a cut is unavoidable, use a total scale cut-off. Do not join numeric values placed at either end of the cut.

- Overall strategy:

  A large number of quantitative information can be entered in a small space.

  Drawing graphs can become an iterative or experimental process.

  When necessary, make a graph of the data more than once.

  Many useful graphs require an exhaustive and responsible analysis.

Chapter
3

Graph classification

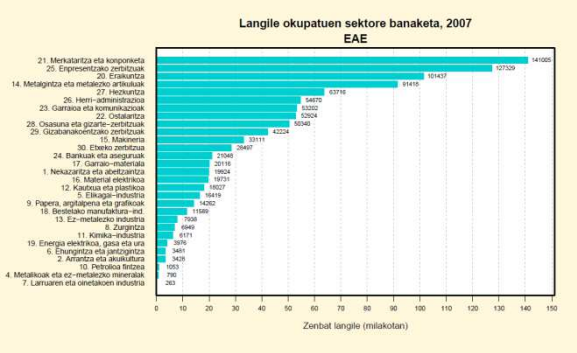
The following classification is based on the nature and origin of the data:

Univariate data

- **Tables:**

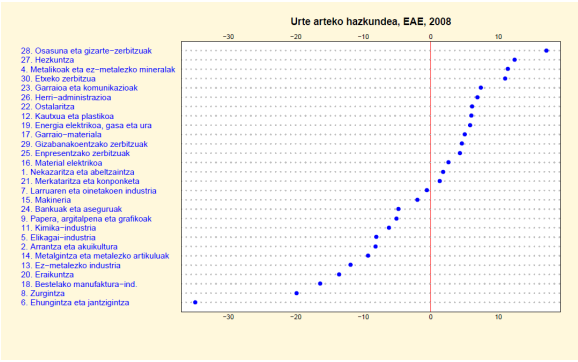
i. Bar charts

a. Simple bar charts



These are used to show the frequency of each modality of a variable. The frequencies are represented by parallel bars, one on the other and ordered from larger to smaller in size. We add on the right side of each bar the frequency of the modality that it represents and several vertical right lines to make the graph easier to interpret. It is also possible to change the colours of the bars and the scales used.

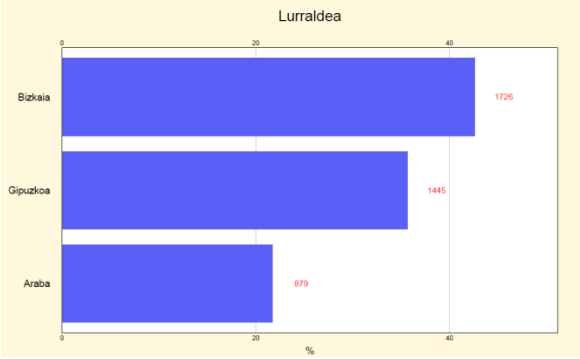
ii. Dot plots



These graphs are used to show the variations that take place in each modality of a single variable. Depending on their variation, modalities are ordered from larger to smaller and their values are represented by dots (a dot is used, instead of drawing a bar). The red line drawn in the 0 shows at a glance if the variation is positive or negative. Furthermore, the vertical right lines make the graph easier to interpret.

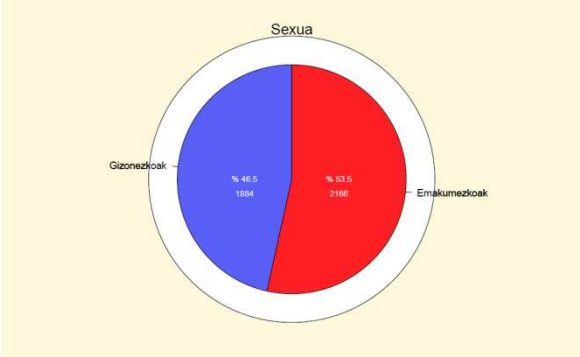
- **Microdata:**

i. Simple bar charts



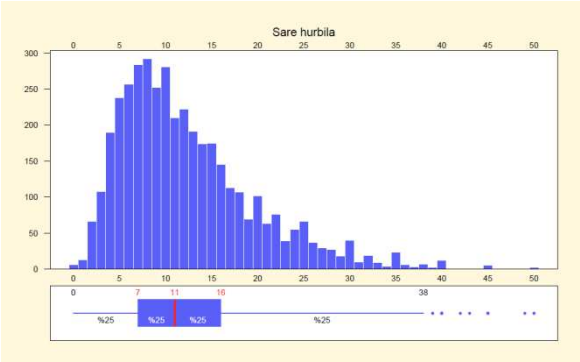
This is similar to the bar chart used in tables. In this case, when the modalities of a variable have an own order it is maintained. If there is no order, they are arranged from larger to smaller. When the variable is continuous, the space between bars is removed.

ii. Pie charts



These are used to represent discrete variables that comprise only a few modalities. A colour is assigned to each part of the pie chart (and, therefore, to each modality), specifying its corresponding frequency and percentage.

iii. Histogram and box plot



These are used to represent continuous quantitative variables. A horizontal box chart is drawn, in which a thick red line represents the median. The quartiles are also represented with red colour, and information on the meaning of the box and whiskers is also given to facilitate the interpretation of the data. This is appropriate for visualising the presence of unusual and atypical values and to analyse the distribution and symmetry of the variables.

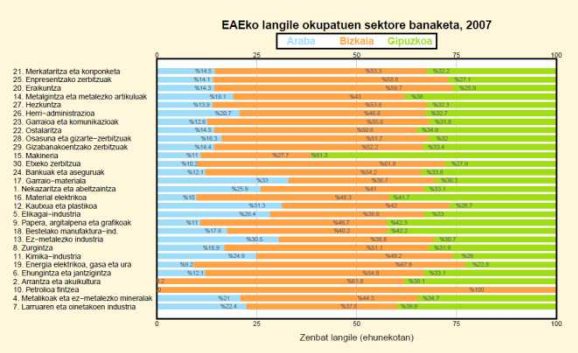
A histogram of the data is drawn on the bar chart to give an idea on the distribution of the variable being analysed.

Bivariate data

- **Tables:**

i. Bar charts

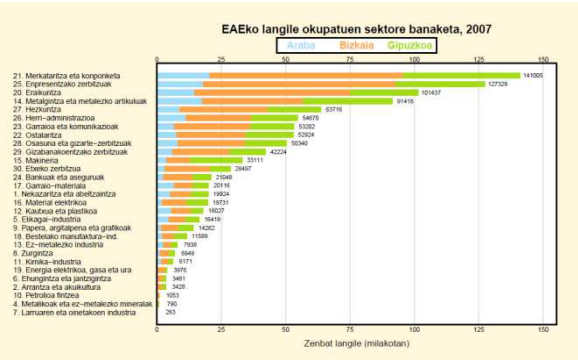
a. Percentage bar charts or relative frequency bar charts



These are used to represent the frequencies of the modalities of a quantitative variable in relation to the three provinces of the Basque Country. First, the modalities are arranged from greater to lesser frequency, depending on their frequency of the Basque Country. The percentages of the three provinces are calculated in the same order and drawn as contiguous bars of different colours, which are completed with a bar that represents 100% of the total frequency. These charts also add straight vertical bars to facilitate their interpretation, in the same manner as the simple bar charts.

It is possible to use, instead of the three provinces, any qualitative variable which has four modalities if one of them can be itemised in the others.

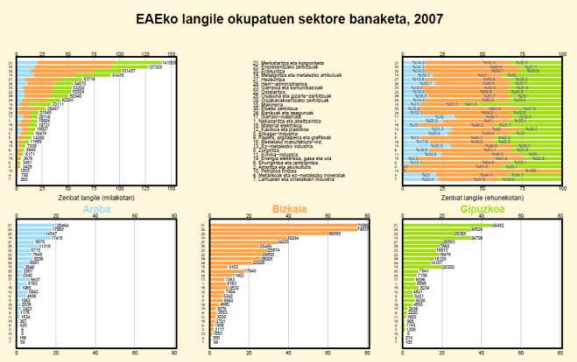
b. Categorised bar charts or absolute frequency bar charts



These are also used to represent the frequencies of the modalities of a quantitative variable in relation to the three provinces of the Basque Country. However, in this case bar size is proportional to the frequency of the modality corresponding to the Basque Country. Similarly to the preceding case, the data are ordered according to the frequency of each modality in the Basque Country, creating a simple bar chart. Next, each bar is divided into segments proportional to the frequency of the three provinces and the corresponding colour is applied to each one. Similarly to the above graphs, vertical lines are added to facilitate the interpretation.

Like in the previous graph, it is possible to use, instead of the three provinces, any qualitative variable which has four modalities if one of them can be itemised in the others.

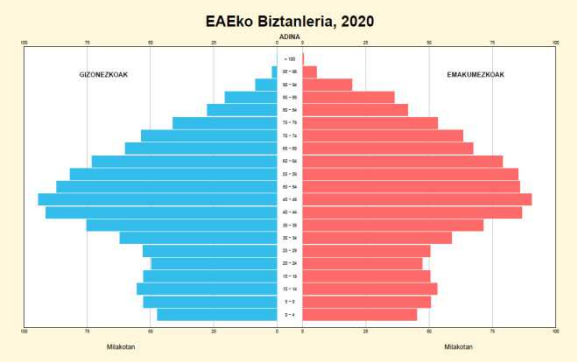
c. Chart composed of bar charts



This is a graphic representation comprising a categorised bar chart and a percentage bar chart of the Basque Country, and a simple bar chart of each province.

ii. Pyramid

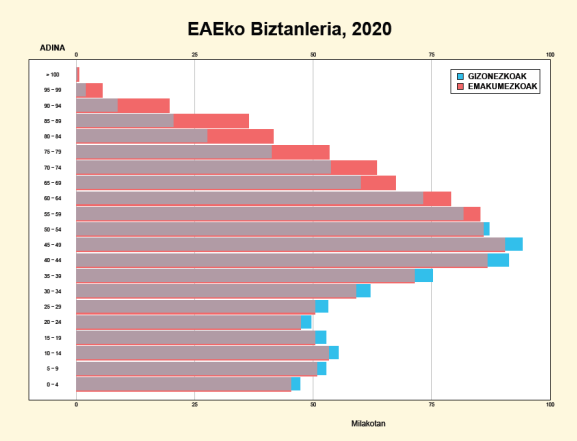
a. Simple pyramid



It is about a double frequency histogram. The bars of the double histogram are drawn horizontally, i.e. up to the abscises line, and the age groups are indicated among the two histograms. Thus, the data for males is shown in blue in the left side of the double histogram and the data for females is shown in red in the right side. It is possible to change the scales and colours.

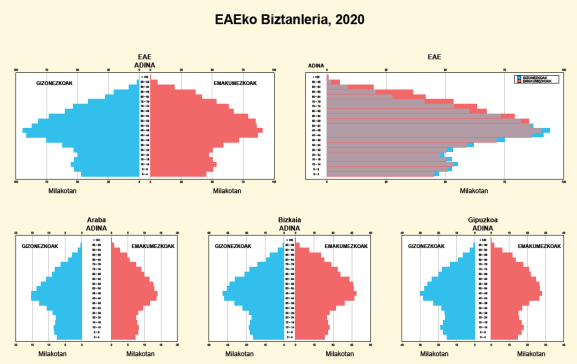
Although it is a graph traditionally used for population topics, it can also be used for economic topics, i.e. to represent commercial balances of importations and exportations.

b. Alternate pyramid



It is similar to the simple pyramid but in this graph the bars corresponding to both sexes are placed on top of each other to make it easier to compare them. Like in the previous graph, there are vertical lines to make it easier to interpret and it is also possible to use to economic topics.

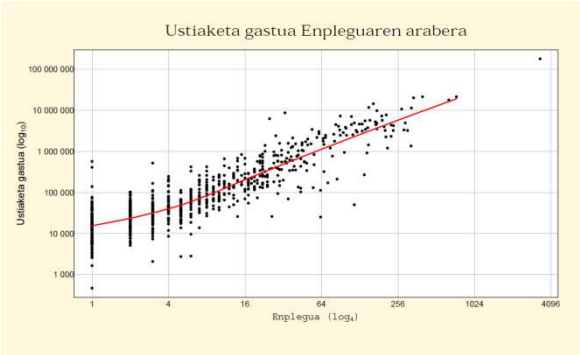
c. Graph composed of pyramids



This is a graphic representation comprising a simple pyramid and an alternative pyramid of the Basque Country and simple pyramids of each province.

- **Microdata:**

i. Stepwise linear regression



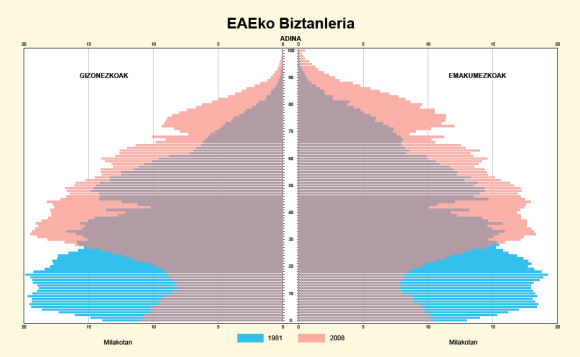
It is a graph used to analyse whether the relation between two quantitative variables is linear. Each data is shown as a point on the graph, in such a way that a stepwise linear regression produces a stepwise linear regression line (marked in red).
If the data are highly asymmetrical, they need to be modified beforehand (using logarithms, for instance).

Multivariate data

- **Tables**

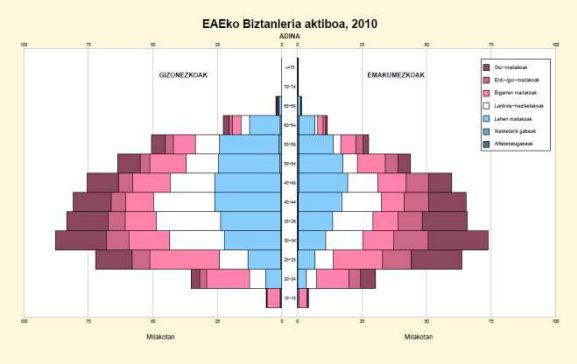
i. Pyramids

a. Bipyramid



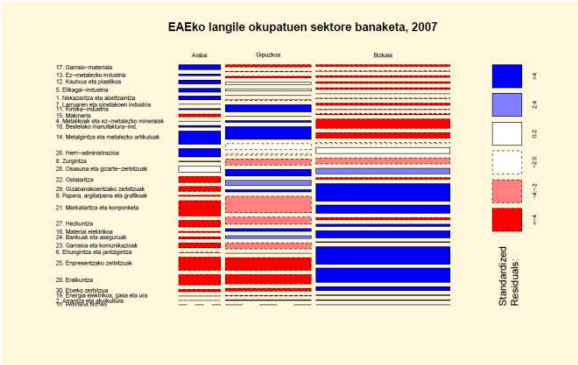
This graph represents, using two pyramids, the populations corresponding to two years, per age and sex. The two pyramids are drawn on top of each other, one in red and the other one in blue. Representing one pyramid over the other and the transparency of the colours make it easier to compare the populations of two different years, for each group of age and for each sex.

b. Categorised pyramid



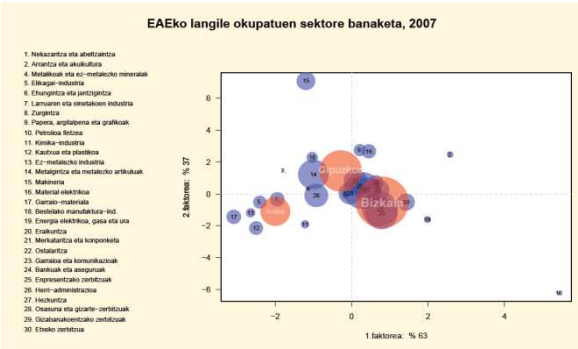
It is similar to a simple pyramid, for multivariate data. In this case, it is represented the population, for each age group and each sex, but according to the modalities of a qualitative variable. Precisely per each group of age and each sex it is drawn the bar that represents the size of the population, which is divided into segments proportional to each modality of the qualitative variable. Each segment is painted a different colour. Similarly to the aforementioned graphs, vertical lines are added to facilitate the interpretation.

ii. Two-category mosaic plots



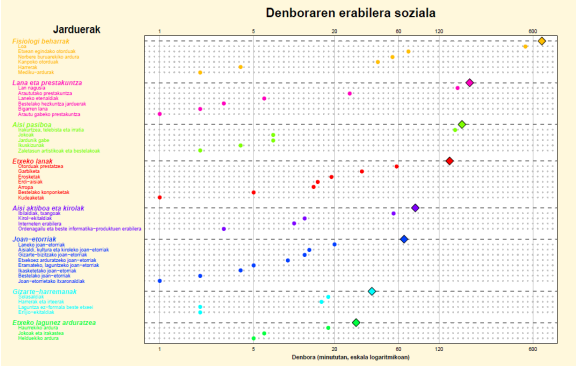
These are used to analyse the relation between two qualitative variables. The area of each rectangle shows the modality frequency of the variable it represents. Height, for instance, shows the frequency of the other variable's modality. The colours of the rectangles, however, show the value and sign of the rest. In other words, colour indicates whether cell frequency is above or below the frequencies estimated according to the model provided.

iii. Graphic representation of correspondence analysis



Correspondence analysis is a graphic method used to analyse the relation between the categories of variables in a contingency table. The closeness (or distance) of a point in relation to one or both axes shows the (positive or negative) similarity or distance with regards to the mean of the modalities defined in the axes. Thus, points tend to be closer to (or further from) the modalities which they resemble to a greater (or lesser) degree. The size of the points, however, indicates frequency. Two colours are used, one for the modality of rows and another for columns.

iv. Cleveland's dot plot

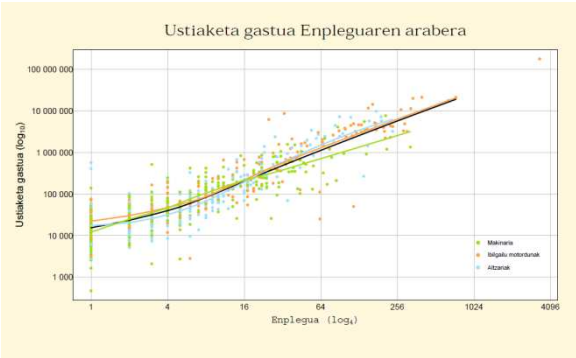


Originally, it is a dot plot with multiple variables that shows the partial and total order of them. It is used when the qualitative variable represented comprises several modalities and sub-modalities. Firstly, the dot plot for the modalities is completed; and then, a dot plot for the sub-modalities of each modality is made.

When the frequency of one modality is much higher than the rest, a logarithmic scale may be needed to see the differences with the other modalities more clearly.

- **Microdata**

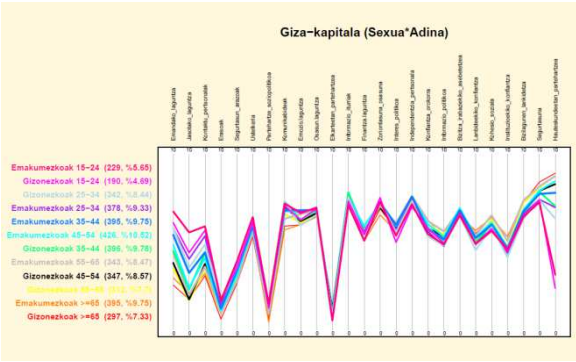
i. Stepwise linear regression by sectors



This graph is used to analyze if the relation between two qualitative variables is stepwise linear, according to the modalities of a third qualitative variable. Each observation is represented with a point and subsequently a general stepwise regression line (in black) and a stepwise regression line is adjusted for each modality of the qualitative variable (each in its own colour).

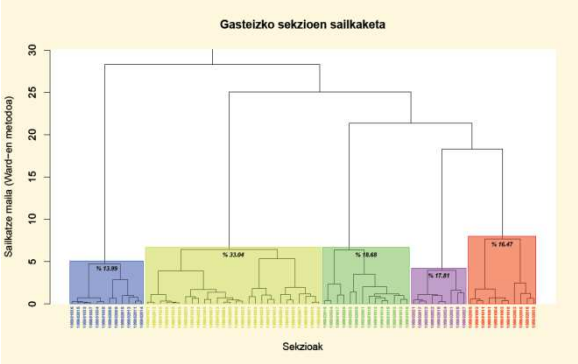
Sometimes, when the variables are highly asymmetrical, for instance, it is convenient to transform them beforehand (by taking logarithms, for example).

ii. Parallel coordinates graph



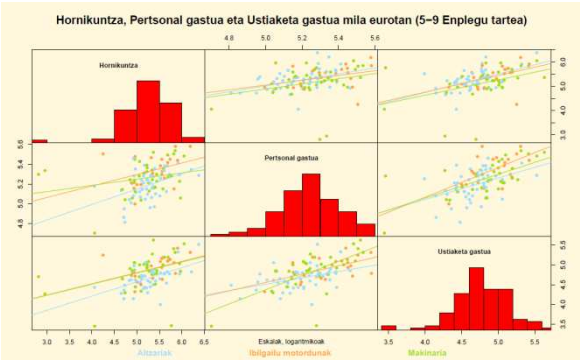
The purpose of parallel coordinates is to represent data sets that have several variables. Each variable is shown as a vertical scale parallel to the others. Observations are represented by points on each axis, using a different colour for each modality. Then a straight line is drawn to join all the points of the same colour. To represent a point on the axis, the value that was assigned to that variable in the observation is taken into account.

iii. Dendrogram



A dendrogram is a graphic representation or chart in the shape of a tree (dendro = tree) that classifies data into sub-categories. These, in turn, are subdivided until the desired level of specificity is reached. The aim is to classify the data into groups, so that the data in the same cluster or group are as similar to each other as possible and as different as possible to the data in the other clusters. This can be convenient for showing the structures and connections between data, which are aspects that can be useful even when they were unknown at the outset. To make the dendrogram easier to understand, each cluster is given a different colour and the percentage of individuals it has is put to one side.

iv. *Scatter plot Matrix*

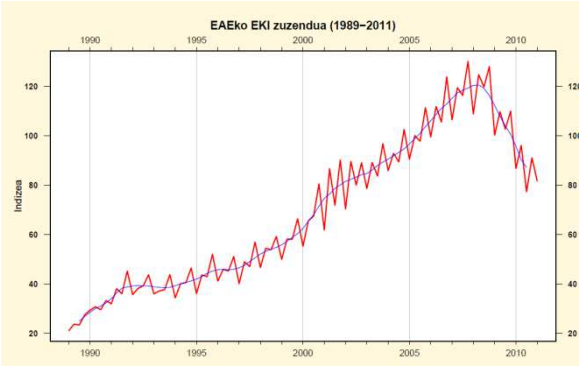


It is based on a matrix of nxn dimensions (where 3≤n≤5) matrix formed by bivariate charts. Histograms are represented in the diagonals and stepwise linear regressions in the other cells. It is possible to analyse variable distribution by looking at the diagonals, and to see the partial regression lines by crossing different variables in the other cells.

Time series

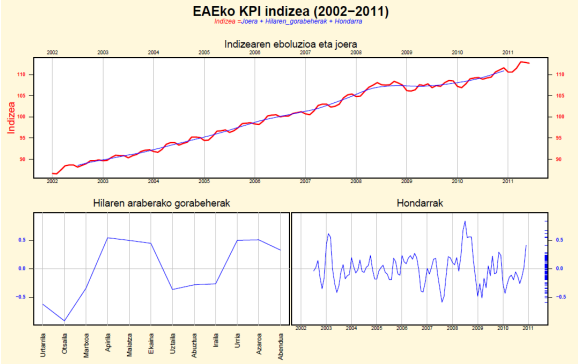
- **Tables**

i. Index graphs



These are graphs that compare an index (in red) to a deseasonalised index from which the calendar effect (in blue) has been removed. By comparing the two indexes, you can see the growths and decreasings in the index under study at different periods in the year.

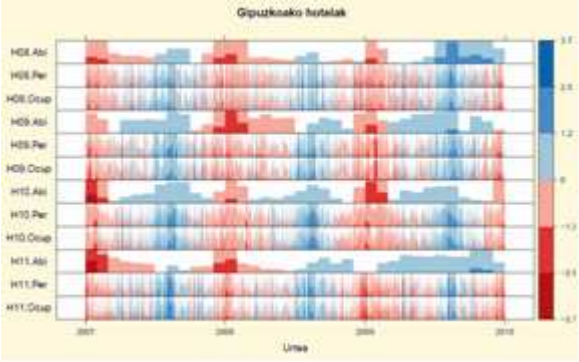
ii. Index graphs Cycles and trends



This graph also compares the index and deseasonalised index, although in this case the difference resides in the fact that when the data are given by months, a monthly graph and a residual graph are added.

- **Microdata**

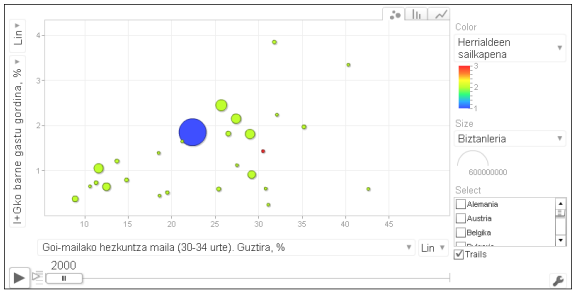
i. Horizon graphs



Horizon graphs allow you to visualise a multitude of time lines at the same time and in a limited space. In short, they are mainly based on the representation of a multitude of line graphs, using colours to save space and visualise all the series at the same time. Specifically, red is used to show negative values and blue is used for positive values. Horizon graphs present a series of benefits, since, among other things, they allow you to:

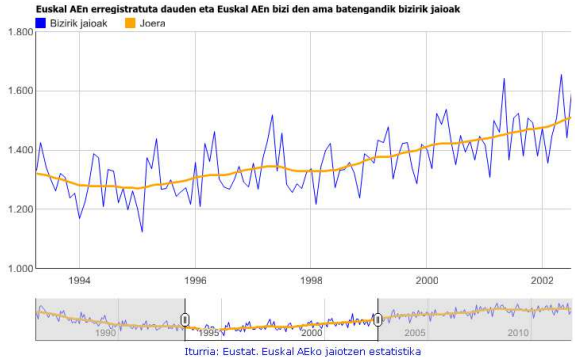
- Observe special behaviours and major trends.
- Analyse the series separately from each other.
- Compare series.
- Observe changes accurately enough to realise whether they are worth an in-depth analysis.

ii. Gapminder Motion Chart



The so-called Motion Chart is a dynamic graph that allows the analysis of data in five different dimensions. You can choose which of five variables you intend to analyse. For example, choose the variables of the X and Y axes; the variable that will measure time (which comes in the initial file); and the variables that will determine bubble size and colour. By analysing the data over time, the graph allows you to identify trends and models that cannot be detected in conventional graphs. It also makes it easier to interpret and offers several tools you can use to play with the graph.

iii. Time Line graphs

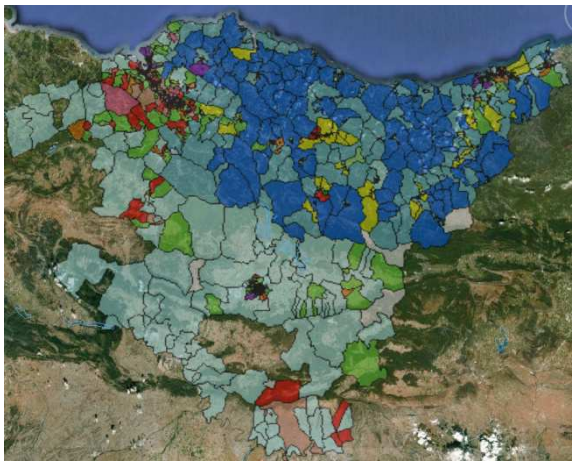


These are based on an interactive graph for visualising time series. When the cursor is placed on each of the series displayed, a bubble pops up, giving the details of the value and the date for that point in the series. The structure displayed under the line chart allows you to focus on a specific time space, automatically adjusting the Y axis scale to the space selected.

Spatial data

- **Tables**
- **Microdata**

i. Visualization using Google Earth



After analysing the spatial data and creating a kml file, Google Earth provides a useful, user-friendly tool to visualise the layer we have made. In this case, it was made an analysis of the census regions of the Basque Country, creating 12 types, and then the kml file was created. The municipal census sections in the Basque Country were coloured according to those 12 types. The transparency of the colours allows you, as you come closer to an address, to see the streets, buildings, parks and so on that are underneath the layer. On the contrary, if you click on the map a panel with information on the area pops up (name of the town; territory; census section; type; percentages of individuals under age 15, of individuals over 65, of Basque speakers and of people with higher education), as well as a link that takes you to the EUSTAT website.

Chapter
4

Applications

Having presented the catalogue of graphs, the most adequate and useful ones are selected. Some are automated and will be displayed on the website. Others are used as tools in Eustat.

These graphs have been implemented using the R programming language.

Graphs for production

The R Scripts of certain graphs have been automated and the inputs standardized so Eustat's Economic Statistics Group can use them as tools for their work. To use the graphs, you only need to load and execute the appropriate functions. The automated graphs are:

1. **Dot plot.**
2. **Chart composed by bar charts.**
3. **Graphic representation of correspondence analysis.**
4. **Stepwise linear regression, by sectors.**
5. **Scatterplot Matrix.**

1. Dot plot

FUNCTION FOR CREATING IN R DOT PLOTS: 'PuntuDiagrama()'

DESCRIPTION

The 'PuntuDiagrama()' function is an adaptation of the 'dotchart()' function of the 'graphics' library of R.

It is appropriate for visualising changes in variable modalities. The chart puts the modalities in alphabetical order or according to their variation, and uses dots to represent them. It draws a straight, vertical red line at the 0 value so the sign of the variation can be seen clearly.

USE

PuntuDiagrama(datuak, main=FALSE, ordenalfabet=FALSE, cex=0.8)
…

ARGUMENTS

datuak:	Numeric values and their labels.
main:	It will be FALSE by default; i.e., do not write the title of the graph. On the contrary, to write the title, indicate it in main="desired title, between inverted commas".
ordenalfabet:	It will be FALSE by default; i.e., the labels defined by dots will not be in alphabetical order. If it indicates TRUE, the labels will appear in alphabetical order.
cex:	Parameter to specify the dot and letter size. The default value is 0.8.

DETAILS

Dot plots are the perfect substitutes for bar charts, especially when analysing variations (either positive or negative).

OBSERVATIONS

Before using this function, save the function and the rda file you intend to represent in the same folder and then indicate the file's path in R. There are two ways to do this: by clicking the icons “File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas").

Next, load the file you intend to represent by entering load("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas").

INPUT FILE STRUCTURE

The input file must consist of two columns:

- The first column is for labels and categories (they can be names, numbers or names and numbers).
- Use the second column to indicate values (increments or drops).

EXAMPLE

setwd(‘C:/Produkziorako grafikoak/Puntu_diagrama’)
source(‘PuntuDiagrama’)
load('RGrafico1_1.rda')
PuntuDiagrama(datuak=RGrafico1_1, main=”HAZKUNDEA”)

2. Chart composed of bar charts

FUNCTION FOR CREATING IN R A GRAPH COMPOSED BY BAR CHARTS: 'MarraDiagramak()'

DESCRIPTION

This graph comprises several bar charts: Categorised bar charts, percentage bar charts and simple bar charts. The 'MarraDiagramak()' function includes three other functions: 'marradiagrama()', 'marradiagrama100()', and 'MarraDiagramak.gr()' (see the notes for further information).

USE

MarraDiagramak(datuak, titulua="")
...

ARGUMENTS

datuak:	Numeric values and their labels.
titulua:	By default, the graph's title is not written. To assign a title to it, indicate main="desired title, between inverted commas".

DETAILS

This visualization of several bar charts allows you to compare the data from each of the three provinces. On the one hand, you can compare the distribution of raw data, and on the other, the distribution of absolute and relative frequencies.

This can be useful for representing the frequencies of the modalities of a qualitative variable, according to the modalities of another, providing the second variable has a few modalities and one of them can be itemised in the others.

OBSERVATIONS

Before using this function, save the function and the rda file you intend to represent in the same folder and then indicate the file's path in R. There are two ways to do this: by clicking the icons "File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas").

Next, load the file you intend to represent by indicating load("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas ").

In turn, this function uses three other functions, which are:

- The 'marradiagrama()' function, that creates simple bar charts and absolute frequency bar charts. It is a special way of using the 'barplot()' function in R.

- The 'marradiagrama100()' function creates percentage bar charts and relative frequency bar charts. Therefore, this function is a special way of using the 'barplot()' function in R.

- The 'MarraDiagramak.gr()' function is a special composition of the 'marradiagrama()' and 'marradiagrama 100()' functions.

The 'MarraDiagramak()' function uses the 'marradiagrama()' and 'marradiagrama 100()' functions to create simple bar charts, as well as absolute and relative frequency bar charts. It also modifies the input data matrix to use the 'MarraDiagramak.gr()' function.

INPUT FILE STRUCTURE:

The input file must consist of three columns:

- The first column is for labels and categories (they can be names, numbers or names and numbers).
- Use the second column to indicate the provinces (you can enter the name of a province or its area code).
- Indicate values in the third column (increments and drops).

EXAMPLE

setwd('C:/Produkziorako grafikoak/Barra_diagramak')
source('MarraDiagramak')
load('RGrafico2_1.rda')
datuak <- RGrafico2_1[-c(1,1+nrow(datuak)/3,1+2*nrow(datuak)/3),]
MarraDiagramak(datuak, titulua="Langile-banaketa / Personnel distribution")

3. Graphic representation of correspondence analysis

FUNCTION IN R FOR REPRESENTING THE CORRESPONDENCE ANALYSIS 'korrespondentziak()'

DESCRIPTION

Use this function to represent a simple correspondence analysis between two qualitative variables.

It consists in modifying and using the 'ca()' function in the library or 'ca' package. (See the notes for further information).

USE

Korrespondentziak(datuak, main="izenburua", zeinux=0, zeinuy=0)
...

ARGUMENTS

datuak:	The file you intend to represent.
main:	Title. It is NULL by default. To assign a title to the graph, indicate main="desired title, between inverted commas".

zeinux:	The parameter that controls the direction of the x-axes. It is 0 by default, i.e. with no turn in the X-axis. Set the value to 1 to turn 180º.
zeinuy:	The parameter that controls the direction of the y-axes. It is 0 by default; i.e. with no turn in the y-axis. Set the value to 1 to turn 180º.

DETAILS

The correspondence analysis is the method used to make a graphic representation of the categorical variables of a contingency table. A correspondence analysis is the generalisation of a known graphic representation; i.e. of a scatter plot.

OBSERVATIONS

Before applying this function, save the function and the rda file you intend to represent in the same folder and then indicate the file's path in R. There are two ways to do this: by clicking the icons “File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas").

Next, load the file you intend to represent by indicating load("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas ").

This function, in turn, uses the following functions:

- The 'ca()' function modifies the 'ca()' function in the 'ca' package that makes a simple correspondence analysis. It gives the option to control the direction of the x and y axes, i.e. if we want to turn any of them 180º or not.

- The 'korresp.gr()' function uses the results of the 'ca()' function, showing the correspondences in two dimensions.

INPUT FILE STRUCTURE:

The input file must consist of three columns:

- The first column is for labels and categories (they can be names, numbers or names and numbers).

- Use the second column to indicate the provinces (you can enter the name of a province or its area code).

- Use the third column to indicate values.

EXAMPLE

setwd(‘C:/Produkziorako grafikoak/Korrespondentzia’)
source(‘korrespondentziak’)
load(‘RGrafico3_1.rda’)
datuak<-RGrafico3_1[-c(1,nrow(datuak)/3+1,2*nrow(datuak)/3+1),]
korrespondentziak(datuak, main=”Korrespondentziak”)

4. Stepwise linear regression, by sectors

FUNCTION IN R FOR MAKING AND REPRESENTING A STEPWISE LINEAR REGRESSION BY SECTORS: 'Erregresioa()'

DESCRIPTION

Use this function to analyse if the relation between two qualitative variables is stepwise linear, based on the modalities of another qualitative variable.

The 'Regression()' function includes three other functions: 'bitxiazein()', 'transf()', and 'LoessErregresio()' (See the notes for further information).

USE

Erregresioa(datuak, titulua=NULL, azpititulua=NULL, logbektorea=c(0,0))
...

ARGUMENTS

datuak:	The file you intend to represent.
titulua:	Title. It is NULL by default. To assign a title to the graph, indicate titulua="desired title, between inverted commas".
azpititulua:	Subtitle. This is also NULL by default. To assign a subtitle, indicate azpititulua="desired subtitle, between inverted commas".
logbektorea:	A vector used to make logarithmic computations when a logarithmic scale is to be used. Unless indicated otherwise, no logarithmic computations are made (it will be c(0.0) by default). 1 corresponds to the logarithmic scale and 0 to the existing scale.

DETAILS

LOESS or LOWESS (locally weighted scatterplot smoothing) is a new method for building based on 'classical' methods, such as linear and nonlinear regressions. These new methods are designed to address cases in which the classical procedures are inadequate or insufficient.

The purpose of LOESS is to create a function that describes the deterministic part of the variation in the data, point by point. It does this by fitting simple models to localised subsets, combining linear least squares regression with nonlinear regression.

OBSERVATIONS

Before applying this function, save the function and the rda file you intend to represent in the same folder and then indicate the file's path in R. There are two ways to do this: by clicking the icons "File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas").

Next, load the file you intend to represent by indicating load ("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas").

This function, in turn, uses three other functions:

- The 'bitxiazein()' function is used to inform on unusual or atypical values. To do this, it establishes a criterion for considering values as atypical or unusual. In fact, values lower than the minimum of the values that are higher than (q1-zenbath*(q3-q1)) are considered atypical, and so are those higher than the maximum of the values that are lower than (q3+zenbath*(q3-q1)). zenbath is a parameter whose default value is 2.75; q1 and q3 refer to the first and third quartiles, respectively.

- The 'transf()' function allows you to search for the best data transformation. To do this, it calculates the Shapiro-Wilk normality test statistic. Thus, the one that has the highest value is the one that provides the best transformation.

- The 'LoessErregresio()' function serves to evaluate whether or not you need to perform a transfer and use a logarithmic scale. Subsequently, if there are points that have the same value, they are moved to prevent them from overlapping. Next, draw the dots, each one in the right colour, and adjust the local regressions, totals and partials (discarding the atypical values). Lastly, it identifies and labels the dots that are away from the main model. This function, in turn, uses the 'loess()' function in the R graphics library. This latter function serves to calculate local regression.

INPUT FILE STRUCTURE:

The input file must consist of four columns:

- Place the dot identifiers in the first column (i.e., the ones you will use to identify atypical values, unusual values, etc.)

- The second column is for labels and categories (these can be names, numbers or names and numbers).

- Put the values for the variable that will go on the x-axis in the third column.

- Put the values for the variable that will go on the y-axis in the fourth column.

EXAMPLE

Setwd("C:/Produkziorako grafikoak/ Zatikako erregresio lineala, sektoreka")
Source("Erregresioa")
load("RGrafico4_1.rda")
datuak <- RGrafico4_2
Erregresioa(datuak, titulua="Balioak Enpleguaren arabera", Azpititulua="Eskala logaritmikoak, logbektorea=c(1,1))

5. Scatterplot Matrix

FUNCTION IN R FOR REPRESENTING A SCATTERPLOT MATRIX WITH HISTOGRAMS IN THE DIAGONAL: 'binakakoa()'

DESCRIPTION

You can use this graph to represent n quantitative variables (3≤n≤5).

In turn, the 'binakakoa()' function uses four other functions: 'transf()', 'hist.puntuak()', 'panel.puntuak()', y 'binakako.pairs()' (See the notes for further information).

INSTRUCTIONS

binakakoa(datuak, titulua=NULL, azpititulua=NULL, logbektorea=rep(0,ncol(datuak)-2)
...

ARGUMENTS

datuak:	The file you intend to represent.
titulua:	Title. It will be NULL by default, i.e. with no title. To assign a title to the graph, indicate titulua="desired title, between inverted commas".
azpititulua:	Subtitle. This is also NULL by default. To assign a subtitle, indicate azpititulua="desired subtitle, between inverted commas".
logbektorea:	A vector that makes logarithmic computations. Unless indicated otherwise, no logarithmic computations are made (it will be rep(0,ncol(datos)-2 by default). 1 corresponds to the logarithmic scale and 0 to the variable's scale.

DETAILS

The Scatterplot explains the relation between two variables. In other words, it can show the shape and the trends of the distribution, and whether the relation between the variables is strong or weak. It can also show clusters and models, as well as any atypical or unusual values.

Conversely, the histogram indicates the relative density of the sample distribution. Bars are used to indicate the amount of data in each interval of equal widths that are the result of splitting the original dataset. Bar height is proportional to bar frequency.

The plot matrix is made up of cells that house charts. Each chart cell in the matrix has a chart that shows the relation between the variables in the rows and columns of cells. Therefore, a plot matrix shows the relations between all the potential pairs of variables in the dataset.

All the graphs in plot matrix cells are not necessarily the same. In this case, we have used Scatterplot matrixes, except in the diagonals, where histograms are shown to give an idea of the distribution of the variables.

OBSERVATIONS

Before applying this function, save the function and the rda file you intend to represent in the same folder and then indicate the file's path in R. There are two ways to do this: by clicking the icons “File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas").

Next, load the file you intend to represent by indicating load("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas").

This function, in turn, uses three other functions:

- The 'transf()' function, which informs of unusual elements.

- The 'panel.hist()' function, which represents the histograms located in the graph's diagonal.

- The 'panel.puntuak()' function, which represents the scatterplots and regressions in the other cells (except for the ones in the diagonal).

- The 'binakako.pairs()' function, which is an adaptation of the 'pairs()' function in the R graphics library. This function, in turn, represents histograms in the diagonal of the matrix, as well as the scatterplots and regressions in the other cells.

INPUT FILE STRUCTURE:

The input file must have four (when n=3) to six (when n=5) columns, depending on the nature of the variables to be analysed:

- The first column is for labels and categories (they can be names, numbers or names and numbers).

- From the second column onwards, specify the variables to be analysed (between 3 and 5).

EXAMPLE

```
setwd("C:/Produkziorako grafikoak/ Scatterplot matrix grafikoa")
source("binakakoa")
load("RGrafico5_1.rda")
datuak <- RGrafico5_1
binakakoa(datuak, titulua="PB_sf, VAB_cf, C_P, eta EN_E (adin-geruzka)", azpititulua="PB_sf, VAB_cf, C_P, eta EN_E logaritmo-eskalan", logbektorea=c(1,1,1,1))
```

Graphs for dissemination

Two dynamic-interactive graphs and a visualization using Google Earth have been created on the Eustat's website. To use the graphs, you only need to load and execute the appropriate functions. The automated graphs are:

1. **Motion Chart Graph.**

2. **Time Line type graph.**

3. **Visualization using Google Earth.**

TOOL TO VISUALIZE MOTION CHARTS AND TIME LINE CHARTS. GOOGLE VISUALIZATION API

An application programming interface (API) is a set of functions and procedures (or methods, when programming targets objects) that offers a specific library to be used by other software.

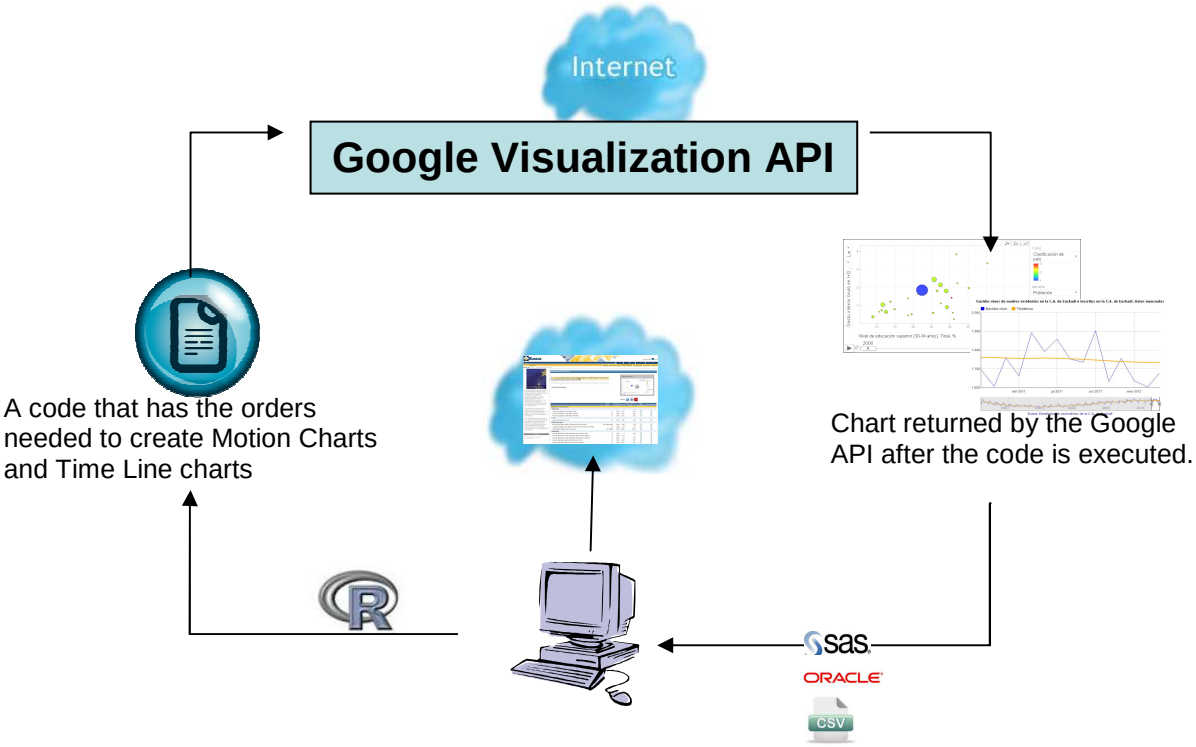
One of the main purposes of the API is to provide a set of general use functions that can be used to draw windows or desktop icons, for example. Programmers use the advantages of API to avoid having to program from scratch.

Google Visualization API is a JavaScript API designed to make advanced visualization technology more user-friendly. It can also be integrated with new visualizations quickly and easily. In fact, it is the interface commonly used to enter data on the web and obtain data for visualizations.

It is based on two key elements: user-friendliness and ubiquity. To create a chart, select one of the charts shown in the Google Visualization API chart gallery and program the code that creates it (naturally, providing you have the conditions and structure required to use the tool) and call the API. It will be executed online and the selected chart will be sent back to you.

You will need to have Flash Player installed on your computer to visualise the chart.

We will now outline the procedure to follow to create the selected chart (or any other chart):



TOOL FOR VISUALISING THE ANALYSIS OF CENSUS SECTIONS TYPOLOGIES.GOOGLE EARTH

Google Earth is a computer program that is similar to a Geographic Information System (GIS). You can use it to visualise images of the Earth by combining satellite images taken in the last three years with maps and a very good data base.

The Google Earth search engine can be used for many purposes, such as, for example, to:

- Search for a specific country, city or address of a city.
- Use the coordinates to search for any territory on Earth.
- Mark existing roads between two cities in the same country or between addresses in the same city.
- Mark the best route between two cities or different countries.

It also has several layers of information that can be enabled or disabled separately to visualise different types of geographic data.

Thanks to the layers, users can:

- See the names and streets of cities.
- Locate schools, hospitals, hotels, restaurants, parks, interesting places and so on.
- Visualise boundaries, roads and railway lines.
- See volcanoes, earthquake epicentres, lakes, lagoons and rivers and so on
- Identify the location of major historical and cultural sites.
- Visualise in 3D hills, mountains and highlands.

1. Motion Chart Graph

1. FUNCTION IN R THAT PREPARES FILE WITH THE RIGHT STRUCTURE FOR CREATING MOTION CHARTS BASED ON DATA FROM EUSTAT INDICATORS: 'Elkartu_gapminder()'

DESCRIPTION

This function takes all the indicators' csv files (one table per indicator) and returns the file needed to use an adaptation of the 'gvisMotionChart()' function in R's 'googleVis' package.

USE

Elkartu_gapminder(lista1, fitxategia_izena, euskara=TRUE, populazioa)

ARGUMENTS

lista1:	This is a txt file, which you can use to specify the names of the csv files for the indicators you want to use, without extensions and separated by spaces. They should be in the following order: first the x-axis indicator; then the y-axis indicator; next, the population indicator (if any); and last, the remaining indicators. To finish, specify the number of indicators or files used (number).
fitxategia_izena:	Name of the output txt file.
gaztelania, euskara, ingelesa:	Language of the input and output files. You will always have to select one and only one. Assign TRUE to the language selected.
populazioa:	It will be FALSE by default. If there is a population indicator, assign TRUE to it.

DETAILS

R's googleVis package provides a useful, user-friendly tool for creating Google Visualization API Motion Chart style graphs. To create this style it was created the 'Motion_chart()' function, which is based on the 'gvisMotionchart()' and 'gvis()' functions from the googleVis package. To execute this function as shown in the next section, the input file must have a special structure; specifically, the output file structure that is obtained precisely by executing the function 'Elkartu_gapminder()'

The 'Motion_chart()' function's input file (and the 'Elkartu_gapminder()' function's output file) must have the following structure:

- First column: A character column with the names of the entities (of the countries, in this case).
- Second column: This column gives an ordered list of the years subject to analysis.
- Third column: The variable to be represented in the x-axis.
- Fourth column: The variable to be represented in the y-axis.
- Fifth column: The variable that is going to determine the colour of the bubbles.
- Sixth column: The variable that is going to determine the size of the bubbles.
- From the seventh column onwards: The remaining variables to be analysed.

NOTE: In the event that values are missing in the sixth column, the next column that has no missing values will set the size of the bubbles. Therefore, files that contain populations should not have missing values.

The 'Elkartu_gapminder()' function gives a txt file with the aforementioned structure. When you create the file, an R editor pops up. Click on the variables to change the names. Any changes you make are saved automatically.

OBSERVATIONS

First, save the files of the indicators you want to use in the same folder. If the population file is not among the indicator files, the folder should also contain the files that have the population for the EU 27 (downloaded from the Eurostat website) and the Basque Country (downloaded from the Eustat website). All the files must be in csv format.

Before you execute the function, indicate in R which directory has the indicator files. There are two ways to do this: by clicking the icons "File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas"). The file that is created after executing the function is saved in the same location.

INPUT FILE STRUCTURE:

The files for each indicator must have the following structure (except for the population files, providing they are not an indicator file):

- In the first column, the names of all the countries (do not leave the first column blank; give it a title). Take into account that the names of countries must be written in the same way in all the files (with the same accents they have now, with no parentheses, etc.).

- The years will appear in the other columns. The titles of those columns will be the years.

In addition, the same symbol used to express missing values (the symbol ":") and decimal values (the symbol ",") must be used in all the start files.

EXAMPLE

Elkartu_gapminder(lista1="indicadores.txt", fitxategia_izena="Euskadi_en_EU27", gaztelania=FALSE, euskara=FALSE, ingelesa=FALSE, populazioa=TRUE)

2. R FUNCTION FOR GENERATING A CODE TO CREATE A MOTION CHART: 'Motion_chart()'

DESCRIPTION

The Motion Chart is a dynamic graph that allows the analysis of data in five different dimensions. This code, given as text, refers to the Google Visualization API and can be entered on a website. Use the file obtained after executing the "Elkartu_gapminder()" function as the input file.

The Motion_chart() has four other functions inside it: 'my.gvis()', 'my.gvisMotionChart()', 'my.gvis_web()' and 'my.gvisMotionChart_web()' (See the notes for further information).

USE

Motion_chart(lista1, fitxategia_izena, izenb_ind, euskara=TRUE, irteerako_fitxategia, populazioa=FALSE, chartid, grafikoa=FALSE, kodea=FALSE)
...

ARGUMENTS

lista1:	This is a txt file, which you can use to specify the names of the csv files for the indicators you want to use, without extensions and separated by spaces. It should be the same txt you used in the 'lista1' argument in the 'Elkartu_gapminder()' function.
fitxategia_izena:	Name of the file to be visualised. It should be the same name you used in the 'fitxategia_izena' argument in the 'Elkartu_gapminder()' function.
izenb_ind:	If the argument is NULL, it will give the indicator variables the same names that are on 'lista1'. Otherwise, you can specify a character vector with the names you want to give the indicators.
gaztelania, euskara, ingelesa:	Language of the input and output files. You will always have to select one and only one. Assign TRUE to the language selected.
irteera_fitxategia:	Name taken by the output file.
populazioa:	It will be FALSE by default. If there is a population indicator, assign TRUE to it.
chartid:	A name to identify the chart.
grafikoa, kodea:	You will always have to select one and only one. To see the chart online: chart=TRUE. To create the code only: code=TRUE.

DETAILS

It can be considerated as a result of the extension of a scatterplot for multivariate data. In fact, once the x-axis and y-axis variables have been selected, you can select the bubble size and colour, for instance, or show how the bubbles have evolved over time. These five variables are for the user to select, so they is not only a dynamic graph but also interactive.

Specifically, the aspects you can select are:

- The x-axis variable.
- The y-axis variable.
- The variable that will measure time. Truthfully speaking, the latter cannot be selected because it coincides with the variable that measures time in the input file, which will normally be the year.
- The variable that is going to determine the colour of the bubbles.
- The variable that is going to determine the size of the bubbles.

Thanks to the option for analysing data over time, this graph allows you to identify trends and models that cannot be seen in conventional graphs.

In addition to allowing you to select data, it allows you to see the data on three different charts:

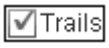
 A Motion Chart (mentioned above).

 A dynamic bar chart. It adds to a conventional bar chart a dynamic bar that measures time. This allows you to analyse the evolution of the bars over time.

 A static line chart. It indicates the evolution of each bubble over time statically, via a line.

It also provides other useful tools, such as, for example:

Change scale. It allows the choice between a linear and a logarithmic scale. Sometimes, when one variable is much more frequent than the others, the differences between variables are not easy to distinguish. In such cases, a logarithmic scale can be the solution.



Trails. Using the trails option allows us to observe the trends in the variables over time.

 *Speed.* You can choose the speed of the chart's animation.

Zoom: You can close in on an image to focus on a specific part of a chart.

 *Control panel.* Via the moveable control, you can adjust the transparency of the bubbles or select the bars that aren't selected.

Below there is an image that gives a graphic explanation of the chart's structure and how it works:

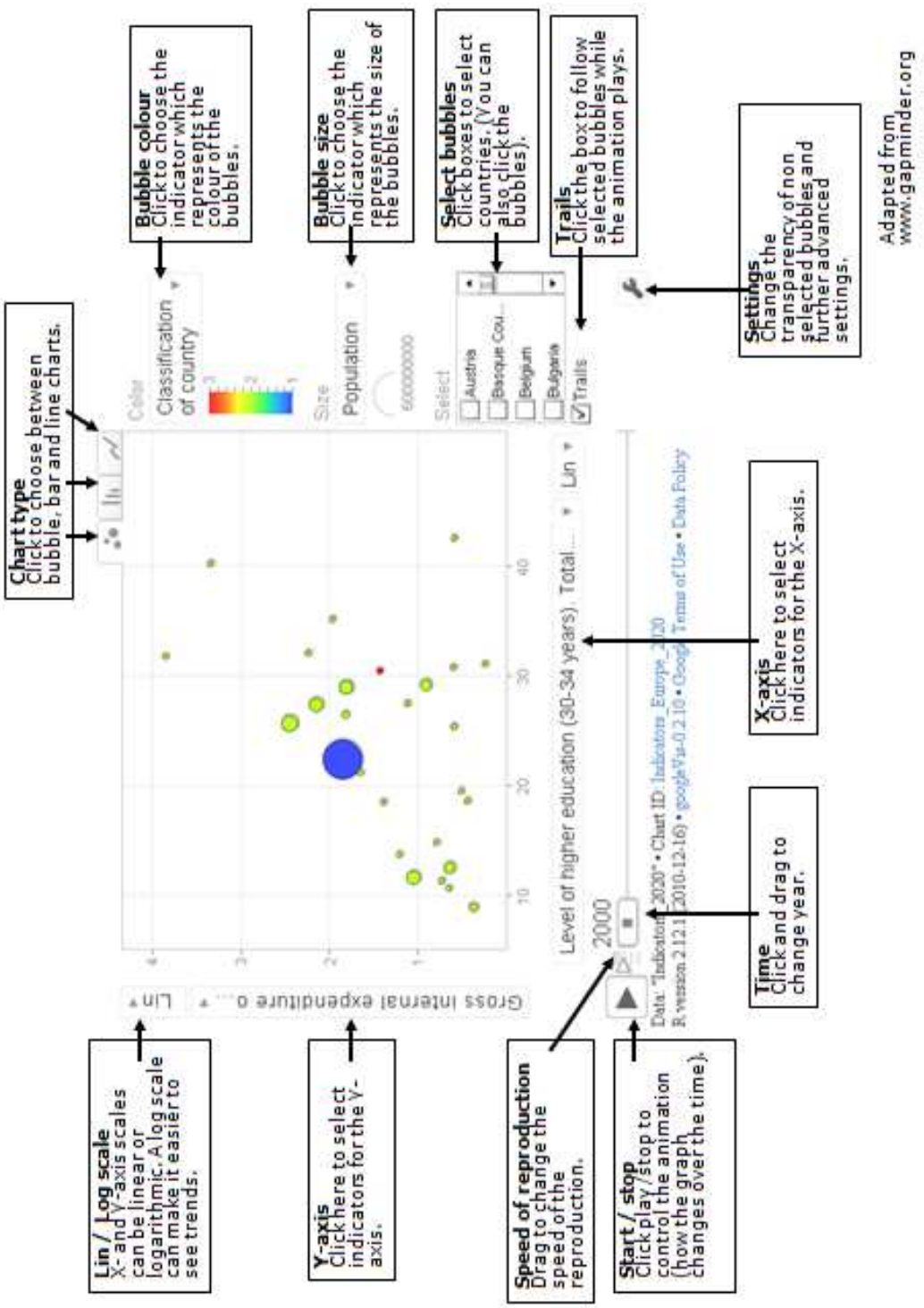


Chart type
Click to choose between bubble, bar and line charts.
Bubble colour
Click to choose the indicator which represents the colour of the bubbles.
Bubble size
Click to choose the indicator which represents the size of the bubbles.
Select bubbles
Click boxes to select countries. (You can also click the bubbles).
Trails
Click the box to follow selected bubbles while the animation plays.
Settings
Change the transparency of non selected bubbles and further advanced settings.
Adapted from www.gapminder.org
Lin / Log scale
X- and Y-axis scales can be linear or logarithmic. A log scale can make it easier to see trends.
Y-axis
Click here to select indicators for the Y-axis.
Speed of reproduction
Drag to change the speed of the reproduction.
Start / stop
Click play/stop to control the animation (how the graph changes over the time).
X-axis
Click here to select indicators for the X-axis.
Time
Click and drag to change year.
Color
Classification of country
Size
Population
Select
Austria
Basque Cou...
Belgium
Bulgaria
Trails
Lin
Level of higher education (30-34 years), Total...
Gross internal expenditure o...
2000

OBSERVATIONS

To implement this function you will need to have the 'googleVis' library installed in R. If it is not installed, follow these steps: Packages → Install package(s) → "select a CRAN mirror" → "select 'googleVis' package". Once it is installed, you will not have to repeat this procedure again.

First, execute the 'Elkartu_gapminder()' function. Then assign the same name that appears in the 'Elkartu_gapminder()' function to the 'lista1' and 'fitxategia_izena' arguments.

Here, too, the files you want to use and all the R programs must be in the same file. In addition, like in the previous function, R must point to this directory.

INPUT FILE STRUCTURE:

The file's structure must be the same as the file obtained by executing the 'Elkartu_gapminder()' function.

EXAMPLE

Motion_chart(lista1="indicadores.txt", fitxategia_izena="Euskadi_en_EU27", izenb_ind=nombres_indicadores_castellano, euskara=TRUE, irteerako_fitxategia="codigo_web", populazioa=TRUE, chartid="Euskadi_en_la_UE_27", kodea=TRUE)

2. Time Line graph

R FUNCTION FOR GENERATING A CODE TO CREATE A TIME LINE CHART 'TimeLine()'

DESCRIPTION

This function takes a time series file and using charts and control structures of the Google Visualization API generates a code that creates a Time Line chart.

The graph is made up of a line chart and a control structure. The movable control allows you to focus on a specific time range.

USE

TimeLine(fitxategia_izena, chart.type, chartid, controlid, chart.options, legend, series, control.options, control.state, euskara=TRUE)

ARGUMENTS

fitxategia_izena:	Name of the csv input file.
chart_type:	Type of chart you want to represent. In this case, it is a LineChart, but the same program would also work for an AreaChart, ComboChart or ScatterChart from the Google Visualization API.
chartid:	A name to identify the chart.

controlid:	A name to identify the movable control structure.
chart.options:	All the options that the chart can have. Both the LineChart and the rest provide multiple configuration options. Among others, you can choose the colour and width of the series; the size and appearance of the X and Y axes; text type and size; and even the chart's height and width. We have selected some of these aspects when we defined the function. Others are the tool's default settings.
control.options:	All the options can have the movable control structure. On this occasion, some options have been selected from all the existing options; the rest are the tool's default settings.
control.state:	This is a time range in which we want to focus on the control structure, initially, and, therefore, on the chart. The vector will have two parts: One to establish the start date and another to establish the end date. The date should be between commas, as follows: "yyyy-mm-dd".
azpititulua:	Subtitle of the graph.
gaztelania, euskara, ingelesa:	Language of the input and output files. You will always have to select one and only one. Assign TRUE to the language selected.

DETAILS

The Time Line graph is an interactive chart that can represent more than one time series at a time. If you get close to the mouse over the line that represents one of the time series, a dialog box pops up with information on the value that the series has on the selected date.

Underneath the line chart there is a tool that can be used to focus on a specific time range. Furthermore, when you focus on a specific period, the chart rescales automatically.

Below there is an image that gives a graphic explanation of how the chart's structure and how it works:

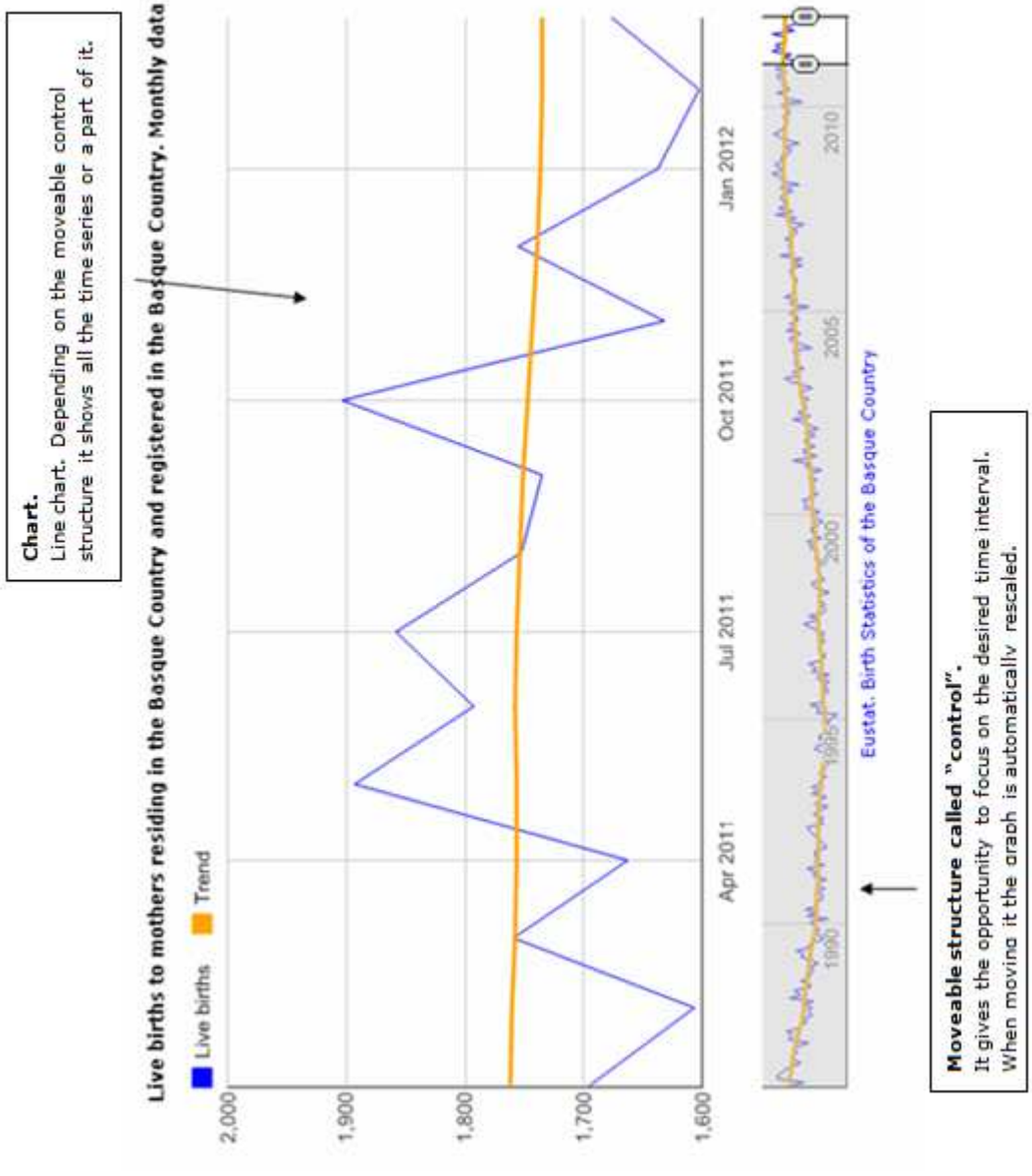


Chart.
Line chart. Depending on the moveable control structure it shows all the time series or a part of it.
Live births to mothers residing in the Basque Country and registered in the Basque Country. Monthly data
Live births
Trend
2,000
1,900
1,800
1,700
1,600
Apr 2011
Jul 2011
Oct 2011
Jan 2012
1980
1985
2000
2005
2010
Eustat. Birth Statistics of the Basque Country
Moveable structure called "control".
It gives the opportunity to focus on the desired time interval.
When moving it the graph is automatically rescaled.

OBSERVATIONS

Before you execute the function, indicate in R which directory has the csv file that you are going to use. There are two ways to do this: by clicking the icons “File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas"). The file that is created after executing the function is saved in the same location.

Next, load the file you intend to use by indicating load("file name and extension, between inverted commas") and the function you intend to use, with the order source("name of the function, between inverted commas").

INPUT FILE STRUCTURE:

The file structure must be:

- Dates in the first column, written "dd/mm/yyyy" (do not leave the first cell blank; give it a title).
- In the second column, enter the data of the series you want to colour blue. In this case it is the raw data of the birth series (do not leave the first cell blank; give it a title).
- In the third column, enter the data of the series you want to colour orange. In this case it is the birth series trend (do not leave the first cell blank; give it a title).

EXAMPLE

TimeLine(fitxategia_izena="IPI_euskara", chart.type=”LineChart”, chartid= "chart_id", controlid= "control_id", chart.options= list(chartArea=list(height="80%", width="90%"), legend=list(position="top"), series=list(seriea1=list(color='blue', lineWidth=1), seriea2=list(color='orange', lineWidth=1))), control.options= list(filterColumnIndex=0, ui=list(chartType="LineChart", chartOptions=list(chartArea=list(width="90%", height="50%"), series=list(seriea1=list(color="blue"), seriea2=list(color="orange"))), minRangeSize=2592000000)), control.state=list("2011-01-01", "2011-12-01"), euskara=TRUE)

3. Visualization using Google Earth

The census sections for the Basque Country have been classified by types in 2012, the same as in 1991, 1996 and 2001. The purpose of this analysis is to classify the census sections according to shared characteristics, by joining the sections that share common characteristics. Thus, sections of the same type will be as similar to each other as possible, and as different as possible to the others.

This type of analysis is very useful in several areas, such as designing stratified samplings, analyses related to urban planning, and in-depth studies of the social and economic situation of the wider community.

ANALYSIS OF CENSUS SECTIONS METHODOLOGY

Initially, the main statistics on family dwellings and the people who inhabit them were selected from the social and demographic statistics data base (BSD). Next, we selected the study variables and divided the majority (with the exception of income) into modalities. There is no maximum number of divisions, since the Multiple Factor Analysis used does not restrict the number of modalities.

Most of the variables selected show the main characteristics of the individuals, such as gender, age, level of education, income, relation to the labour market, sociolinguistic status, and native autonomous region. Further data on the dwelling (e.g. size and year of construction) and the family are added at the end.

After the variables to be analysed have been selected, the frequency of each modality is assigned to each census section, in relative values, with 100 being the result of the sum of all the modalities for each variable (with the exception of income). Although it does not enter in the analysis, family income has also been classified by modalities, because it is useful for describing the typos.

A Multiple Factor Analysis was performed, based on the data matrix created according to the above premises (discarding income and maternal language). The first seven factors were saved, since they represent 70.56% of the initial variance (of the preliminary information). Subsequently, a hierarchical classification was made, using seven factors. This gave 15 typos, and a statistical stability process was implemented. It was thought that there were too many typos, so some of them were joined according to criteria of proximity and distance, until 12 typos were left. Lastly, the nuclei of the 12 typos were taken and used as a seed to create a new partition via k-means clustering or *nuées dynamiques*.

THE FUNCTION IN R FOR VISUALISING THE TYPOLOGY ANALYSIS USING GOOGLE EARTH: 'KmlSortu()'

DESCRIPTION

This function starts with an shp file and adapts the '*kmlPolygon()'* function in the R 'maptools' library to create a file that can be visualised with Google Earth software.

Google Earth is a free, well-known, user-friendly tool that can be used to visualise and analyse spatial data at first glance, quickly and easily.

USE

KmlSortu(n_tipol, fitxategia, gaztelania=TRUE, alpha)

ARGUMENTS

n_tipol:	The number of types obtained after analysing the typologies.
fitxategia:	An shp file that contains all the necessary information
gaztelania, euskara, ingelesa:	Language of the visualization that will be created You will always have to select one and only one. Assign TRUE to the language selected.

alpha:	This parameter controls the transparency of the colours used to colour the Google Earth sections. It will be 120 by default.
legenda:	It will be FALSE by default, and the legend is not shown. When it is TRUE, the legend will be added, along with a description of the typologies on the left side of the display.

DETAILS

When using the coloured map that classifies the Basque Country census sections into several typos, you can use all the options provided by the Google Earth tool.

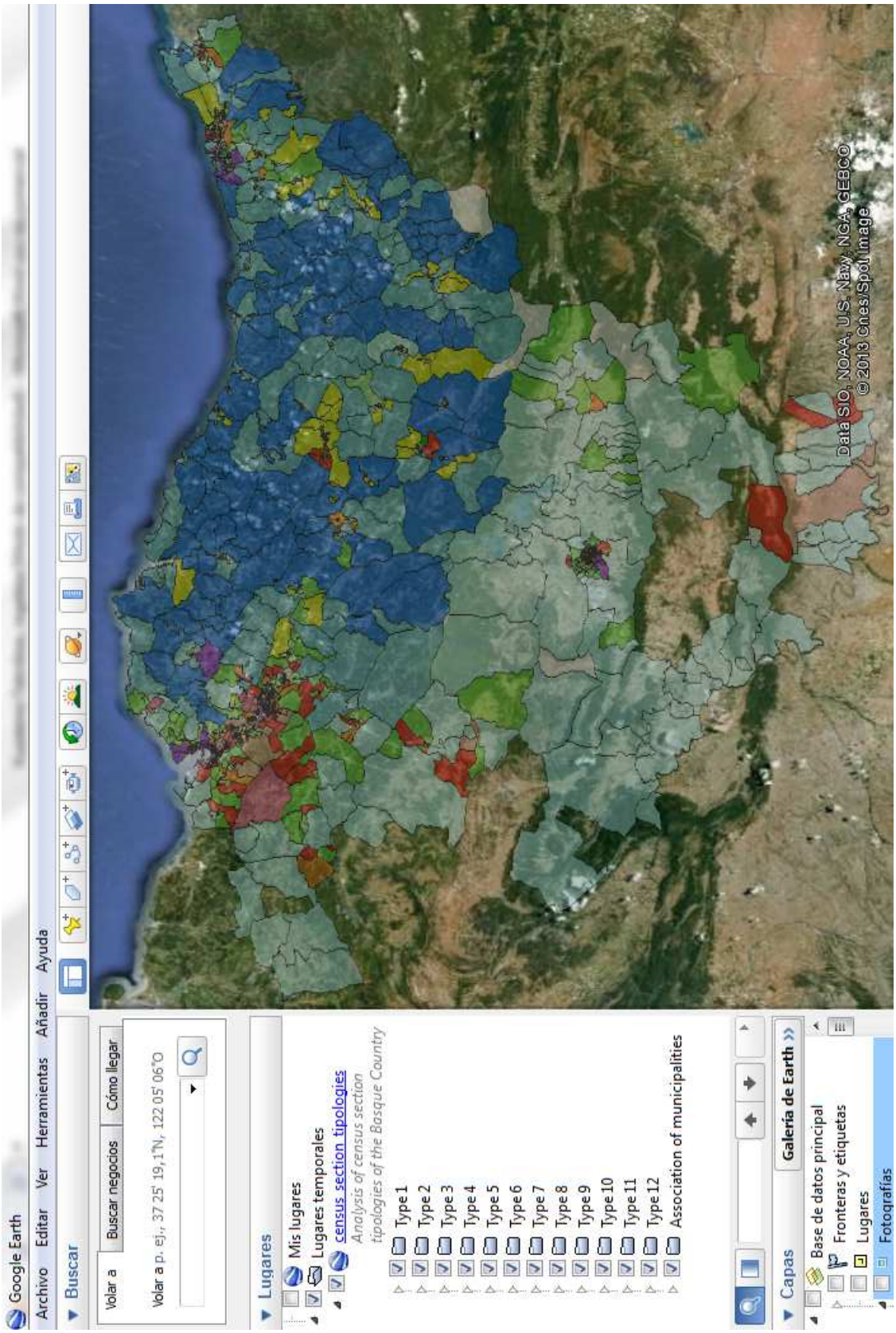
The kml file of typos that have been created also has other options you can use.

For example, by clicking on the dropdown menu on the left, you can select only the typos that are of interest to you. Only the census tracts for the selected typos will be coloured, and when you click on the typos in the dropdown menu, a list of the sections classified in them pops up.

As you come closer to an address, the layer will gradually come into focus until you can see the streets, buildings and parks underneath it.

Click on the map to make a dialog box pop up with information on the census section (name of the municipality; municipal district; section; typo; percentage of inhabitants under age 16 and older than 65, Basque speakers and people with higher education; link to the documentation of the analysis of the typologies of the Basque Country on the Eustat's website).

Below you can see a screenshot of what the visualization of the typology analysis for the census tracts in the Basque Country looks like in Google Earth.


Google Earth
Archivo Editar Ver Herramientas Añadir Ayuda
Buscar
Volar a
Buscar negocios
Cómo llegar
Volar a p. ej., 37 25' 19,1"N, 122 05' 06"O
Lugares
Mis lugares
Lugares temporales
census section tipologies
Analysis of census section tipologies of the Basque Country
Type 1
Type 2
Type 3
Type 4
Type 5
Type 6
Type 7
Type 8
Type 9
Type 10
Type 11
Type 12
Association of municipalities
Capas
Galería de Earth
Base de datos principal
Fronteras y etiquetas
Lugares
Fotografías
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2013 Cnes/Spot Image

OBSERVATIONS

In this case, there is no need to indicate the directory where the shp file is saved before loading the function. You can indicate it later, before executing the function.

There are two ways to indicate the directory's location: by clicking on the icons “File → Change path… → (appropriate file) in R; or by entering the command setwd("full path to the file's location, between inverted commas"). The kml file that is created after executing the function is saved in the same location.

INPUT FILE STRUCTURE:

To create a file that colours the Basque Country's census sections by types, you need an shp file called shapefile.

Shapefile is a digital vector storage format for storing the location of geometric elements and related attribute information. It consists of several digital files. To be specific, the complementary shp used here are:

- .shp. A file that contains the geometrical features of objects.
- .shx. A file that contains the index of the geometrical features.
- .dbf. A database that contains information on the attributes for objects, in dBASE format. The file structure should be:
    - Section code, aggregated and disaggregated (province, municipality, territory and section).
    - Total number of individuals in the section.
    - Typo to which belongs each section.
    - Name of the municipality where the section is located.
    - The variables listed in the dialog box: percentage of inhabitants over age 16, older than 64, Basque speakers and people with higher education.

NOTE: The shp file with the 2009 tracts has been used, associations of municipalities included.

EXAMPLE

KmlSortu(n_tipol=12, fitxategia="Seccion_regionT.shp", gaztelania=TRUE, alpha=150)

Chapter
5

Conclusions

Data visualization makes it possible to observe models, data structures and imperceptible trends by simply looking tables or files. Charts enable us to appreciate the complex models that adjust to the data and judge their usefulness. Data visualization and synthesis highlights its usefulness, and turns statistical data into knowledge.

To take full advantage of data visualization you need to use certain methodology and take into account some basic principles, specified in the second chapter.

Normally, when making a chart, the quantitative and categorical information are coded on the basis of position, size, signs and colours. To analyse a chart, the information is decoded at first glance, so if the decoding is effective, the method used to make the chart is considered good.

A visualization method that leads to an ineffective decoding process at first glance may prevent the detection of main data characteristics or distort the way the data is perceived.

Occasionally, using visualization tools to analyse data avoids the surprises and misinterpretations that can arise in the course of gathering data.

The purpose of the work was to assign the charts that could be adjusted the best to the type of data available to Eustat. At times it was difficult to make a general, configured R-script because it was very close to the characteristics of the file data.

Sometimes changes need to be made in the file before running the scripts. The R software is a versatile tool for making such changes.

It is intended to let available the code that creates the graphs from the Google Visualization API, to have the chance to attach them to their blogs and web-pages.

Current technology enables us to do what would have been unthinkable until recently. Even so, the official statistics are still in the early stages of the possibilities that this techniques offers within the area of the analysis and diffusion of data.

The work done to date in the field of visualization, and what remains to be done, will have consequences and a direct impact on the use made of official statistics and on those who use them.

Chapter
6

Bibliography

[1] YURRAMENDI, Y. (2011)

Estatistika ofizialak sintetizatzeko eta ikustarazteko metodo grafikoen eta teknologia berrien azterketa. Konputazio Zientziak eta Adimen Artifiziala saila.

[2] CLEVELAND, W.S (1985)

The Elements of Graphing Data. Monterey, CA: Wadsworth.

[3] YOUNG, F.W.; VALERO-MORA, P.M. AND FRIENDLY, M. (2006)

Visual Statistics. Seeing Data with Dynamic Interactive Graphics. Wiley, New Jersey.

[4] GALBETE, E.; ADIN, A.; ARAMENDI, J.; IZTUETA, A. AND YURRAMENDI, J. (2012)

Análisis de tipologías de las secciones censales de la CAE. Visualización en Google Earth. IX Congreso Vasco de Sociología y Ciencia Política.

[5] ESCOFIER B. AND PAGÈS, J. (1992)

Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación. Euskal Herriko Unibertsitatea, Argitalpen zerbitzua.

[6] LÊ, S., JOSSE, J. AND HUSSON, F.. (March 2008)

FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, Volume 25, Issue 1.

[7] EUSTAT (1996)

Análisis de tipologías de las secciones censales de la C.A. de Euskadi. Vitoria-Gasteiz: Instituto Vasco de Estadística.

[8] EUSTAT (2001)

Análisis de tipologías de las secciones censales de la C.A. de Euskadi. Vitoria-Gasteiz: Instituto Vasco de Estadística.

[9] GREENACRE (2008)

La práctica de análisis de correspondencias. Vitoria-Gasteiz: Instituto Vasco de Estadística. Fundación BBVA, Bilbao. Rubes Editorial.

[10] **R Graph Gallery:** http://gallery.r-enthusiasts.com/

[11] **R-Project:** http://www.r-project.org/

[12] **FactoMineR:** http://cran.r-project.org/web/packages/FactoMineR/index.html

[13] **Google Visualization API:** https://developers.google.com/chart/

[14] **API. Wikipedia:**
http://es.wikipedia.org/wiki/Interfaz_de_programaci%C3%B3n_de_aplicaciones

[15] **Google Earth. Eduteca:** http://www.eduteka.org/GoogleEarth.php