Color Coding for Data Visualization

Simone Bianco
Università degli Studi di Milano-Bicocca, Italy

Francesca Gasparini
Università degli Studi di Milano-Bicocca, Italy

Raimondo Schettini
Università degli Studi di Milano-Bicocca, Italy

INTRODUCTION

Visualization is a process of mapping data onto visual dimensions to create a pictorial representation. A successful visualization provides a representation which allows the user to gain insight into the structure of the data or to communicate aspects of this structure effectively, The use of color for encoding information can greatly improve the observer's understanding of the information depicted by image and his/her capacity for remembering it. However, many aspects of color itself and of its use are unknown to both users and system designers. Users are often allowed to choose colors that can not be reproduced, that can not be distinguished by the human eye, or are easily misinterpreted. The broad range of variables involved, as well as the interactions and the trade-offs among them, pose difficult problems in selecting specific colors, in predicting their appearance when they are seen in relation with others, and in predicting the observers' interpretation and reaction to them. These problems can not be solved without a working knowledge about color reference systems, color reproduction technologies, mechanism of color perception, users' cultural and emotional reactions. The goal of this article is to critically discuss how a color scale should be designed to effectively represent both qualitative and quantitative information. Color coding requirements, with regard to the task at hand, the characteristics of the media, and the characteristics of data to be coded are therefore analyzed providing examples and references to related works.

BACKGROUND

A color coding scheme can be characterized by a color mapping function $f: D \rightarrow C$ that maps data values D to colors from the color palette C. In case of univariate data, each value is mapped to a single color, with multivariate data, each combination of values is mapped to a single color.

In the following we make a distinction only between qualitative (nominal) and quantitative (ordinal, interval and ratio) data values.

Associating a set of colors with a set of items to express the significance of each is called "nominal color coding." Examples of nominal data values are: water, vegetation, and urban. There is no logical ordering in this sequence.

Color can be also used in a quantitative fashion, i.e. to convey information about ordered data set. We can distinguish here among "ordinal, interval and ratio color coding." In ordinal coding the data values are in some way ordered, i.e., the data values can be put into a sequence but no distance is defined between data values. Examples of ordinal data values are: very bad, bad, average, good, very good. In interval coding data one can define a distance between two data values, but the zero point is arbitrary. The hue by itself can be seen as interval data values, we can say that the distance between red and yellow is 90° but we can not says that yellow is bigger than red. Periodic functions typically produce interval data values. Data sets with both positive and negative values can have a zero point representing no change, average, or expected value. In

DOI: 10.4018/978-1-4666-5888-2.ch161

such data, deviation from zero is what is interesting. In ratio coding a zero point is therefore defined. Data values that represent the temperature are ratio data, in this case the zero point is defined at 0 Celsius degrees and we can state that a temperature of 20 Celsius degrees is twice a temperature of 10 Celsius degrees.

For nominal and ordinal data types discrete palettes must be used. For interval and ratio data one can use both discrete and continuous palettes, according to the data structure and to the visualization aims. If a discrete palette is used to code interval and ratio data, those values must be a-priori quantized into a finite number of ranges.

MAIN FOCUS OF THE ARTICLE

The goal of this article is to critically discuss how a color scale should be designed to be effectively adopted to obtain a successful visualization. For instance, adding a color which does not add additional insight to the visualization can sometimes cause confusion as users try to understand its meaning and should, therefore, be avoided. So, it is particularly important to perform the right choice in order to build visualizations which depict the desired information in a clear way. The color scales and methods used for encoding information depend on several factors. They certainly depend on the data type involved in terms of information carried (qualitative and quantitative) and dimensionality of the representations (univariate or multivariate). The goal of a specific visualization is also an important issue when choosing a color scale. Different tasks and goals require different color coding schemes. Certain goals may require specific subsets of data to be highlighted. This can be achieved for instance by using bright, warm, and fully saturated colors. Within this context, Bergman et al. (1995) have proposed to select the appropriate color scales according to four specific tasks, while Tominski et al. (2008) have presented a task-driven color coding. Different color perception of the viewers must also be considered. An estimated nine to twelve percent of the male population and less than 1% of females suffer from some form of color vision deficiency. The degree to which a person may have an abnormal color vision ranges from slight difficulty in recognizing color shades to a complete loss of color vision. The most prevalent type is a deficiency in perceiving red/green differences (approximately eight to ten percent of the male population), while one to two percent of men are blue/yellow color blind. VisCheck (http://www.vischeck.com/vischeck) is a web tool base on based on SCIELAB which shows how an image is seen by a user with some kind of color blindness. Apart from its aesthetic appeal, we can use image processing techniques to try to make information in images available to color blind people (http://www.vischeck.com/daltonize/) without distorting the color balance to an unacceptable degree. Color schemes that accommodate red or green-blind dichromats will accommodate most other forms of color deficiency.

Many colors have certain qualities associated with them because of their natural occurrence, cultural usage or technical norms. However color also depends on, and is constrained by, the cultural traditions or technical experience of the users. We commonly associate green with vegetation, for an USA audience the color green is associated with the color of money.

Thus in this article the color coding requirements, with regard to the task at hand, the characteristics of the media, and the characteristics of data to be coded are therefore analyzed providing examples and references to related works.

Color Space Selection

Several approaches to color description, whatever based on device-dependent data, colorimetry, or some other empirical organizations give rise to various color spaces. The use of RGB color spaces seems unrealistic for man-machine communications since its metrics does not represent color differences in a uniform scale and the colors are not organized in a manner intuitive for human observers. Color spaces, such as RGB and derived color spaces, are device-dependent and therefore do not univocally describe colors. If colors were identified only by their device-dependent coordinates it might be difficult to reproduce the same image on another device (display, printer, projector….) or worst still, an image that differs from the one intended is generated.

Device-independent color description, perceptual addressability, and the capability of representing color differences in a uniform scale can be achieved by specifying colors in the CIELUV or CIELAB color spaces. In our opinion CIELAB space represents the best that