CHAPTER 1

GRAPHICAL RHETORIC

> Academic science is a social institution devoted to the construction of a rational consensus of opinion over the widest possible field.
>
> JOHN ZIMAN *An Introduction to Science Studies* (1984)

> They [*sc.* the Royal Society] have exacted from all their members, a close, naked, natural way of speaking; positive expressions; clear senses; a native easiness; bringing all things as near the Mathematical plainness, as they can: and preferring the language of Artizans, Countrymen, and Merchants, before that, of Wits, or Scholars.
>
> THOMAS SPRAT *The History of the Royal-Society of London, for the Improving of Natural Knowledge* (1667)

> The advantages proposed by this [*sc.* graphical] mode of representation, are to facilitate the attainment of information, and aid the memory in retaining it: which two points form the principal business in what we call learning, or the acquisition of knowledge. Of all the senses, the eye gives the liveliest and most accurate idea of whatever is susceptible of being represented to it.
>
> WILLIAM PLAYFAIR *The Statistical Breviary; Shewing, on a Principle Entirely New, The Resources of Every State and Kingdom in Europe* (1801)

1.1 Introduction

Science, as stated in the first quotation, is fundamentally a social enterprise;[1] so being a scientist is not only getting new results, but also communicating

[1] There is now a large literature on "Social Studies of Science"; Ziman's book is a good introduction.

them to other researchers. As with all human communication we can break this into four steps:

1. Something – an idea, result, or whatever – is a thought you have.
2. Based on this thought, you produce some kind of signal: speech, or shapes on paper or on a screen.
3. Somebody else receives these signals, through their ears or their eyes.
4. They interpret these signals to produce thoughts in their own head.

In scientific communication, as in scholarly communicaton in general, the goal is for the thoughts at both ends to be the same; so your aim, in Step 2, should be to produce signals that will cause Steps 3 and 4 to make someone else think about things as you do.

Scientists use three kinds of signals, some little used in other contexts:

1. Language, spoken or more usually written.
2. Equations and other mathematics.
3. Pictures, maps, diagrams, and graphs.

Because we acquire and use spoken language so easily, it is deceptively easy to believe that what you write is as clear to others as it is to you; "deceptively" because written prose can easily be tedious or obscure – sometimes so obscure that few readers can understand it. You can reduce the chance of this happening by learning techniques for clear writing,[2] and by revising your writings carefully: no first draft should survive unscathed. You should aim for a style that matches Sprat's description above.

Presenting mathematics can be easier, since mathematical notation is much more structured and much less ambiguous than natural language. But there are still choices to be made when writing mathematics. Your first choice is how detailed you need to be – remembering that what is perfectly comprehensible to some (especially you) will be too condensed for other readers. Notations also require choices. The more symbols the reader has to assimilate,

[2] I recommend Joseph Williams (1990) *Style: Toward Clarity and Grace* (Univ. of Chicago Press, Chicago), as a practical guide to writing clear academic prose. G. D. Gopen and J. A Swan (1990) The science of scientific writing, *Amer. Scient.*, **78**, 550-558 provide a summary, which shows how to improve a seismological example. This paper is available online at the *American Scientist* website,

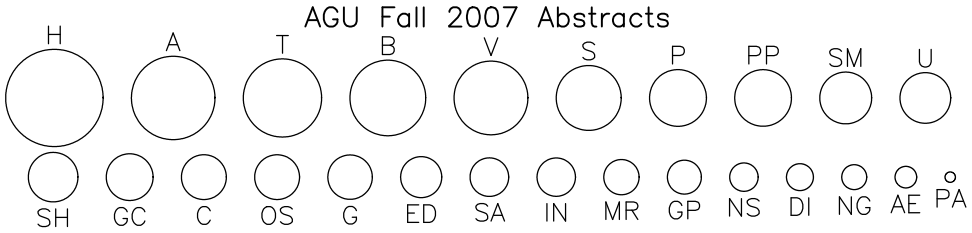

Figure 1.1:

the harder the mathematics is to read. Little is gained and something lost by a plethora of subscripts and superscripts; for example, once you have learned vector notation, it is easier to read $bgradcrossbr$ than $eijk\partial sjr_k$. Similarly, it is usually a bad idea to use names for variables or subscripts, however convenient this may be in writing programs. And subscripting can make the notation easier to remember: displacements in spherical coordinates are easier if written as u_r, u_θ, and u_ϕ rather than with arbitrary labels such as u, v and w.

But the most difficult element of communication is pictures, particularly graphical representations of data, which were the last class of communication methods to be developed.[3] One reason for this late development was a prejudice (still present in some areas of scholarship) that pictures are inferior to words;[4] another was technological: pictures were more expensive to produce, and reproduce, than written language. Now that everything is done with bits, this cost discrepancy (color printing aside) is now gone, and there is software to make complicated graphics easy to produce. This software does not solve the problem that graphics can communicate your message well, or badly – and many popular software packages make the bad choices the easiest ones.

My aim is to set out some principles to help you produce clear and informative graphs. Many of these principles are taken from the books by Tufte and by Cleveland, but adapted to geophysically relevant examples.

[3] Statistical graphics were largely invented by William Playfair, in two books published in 1786 and 1801, and only gradually came into use through the nineteenth century. A good online resource for this history is M. Friendly and D. J. Denis, "Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization"

[4] Rudwick, M. J. S. (1976). The emergence of a visual language for geological science: 1760-1840, *Hist. Sci.*, **14**, 149-195.

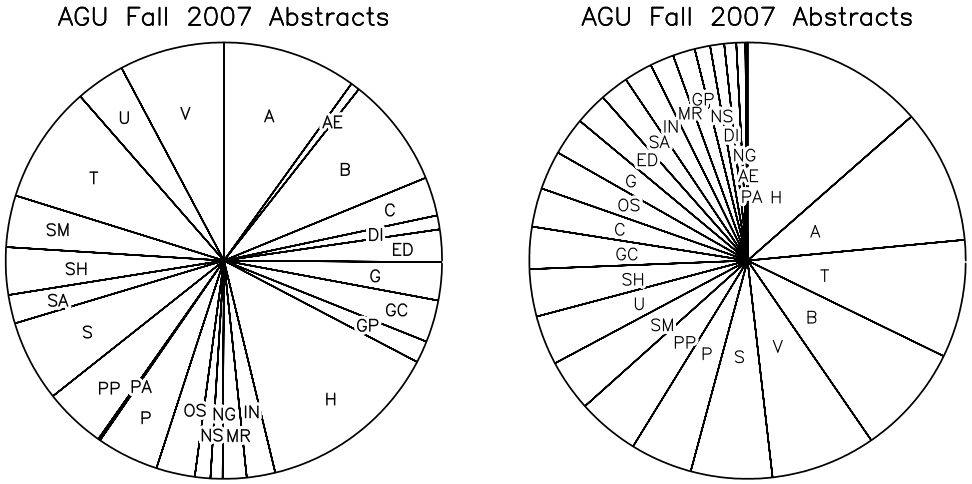

Figure 1.2:

1.1.1 Encode Your Data Intelligently

We start with one of Playfair's inventionsm which is to represent the size of something (such as population) by circles with proportional areas. For our example, we look at the size of various geophysical populations, as given by the number of abstracts submitted for each section of the Fall 2007 AGU meeting. Figure 1.1 shows the data plotted using circles with areas proportional to the number of abstracts, and two-letter codes as identifiers. This certainly gives a more immediate impression of relative sizes than a table would, but it is difficult to compare areas when these are nearly the same, unless the circles are next to each other. Also, representing numbers by areas is a possible source of ambiguity; we might wonder if diameter is the variable being used instead, though it shouldn't be.[5]

A more compact approach uses another of Playfair's inventions, the pie chart: a staple of business and newspaper graphics, with a popularity far greater than its usefulness. The left plot in Figure shows the pie slices arranged alphabetically; on the right, we follow Playfair (many piemakers do not) in showing the segments sorted from largest to smallest. The second version is better because it puts similarly sized segments close to each other. The pie chart is more compact than the circles, but we still cannot easily compare

[5] Scaling two- or three-dimensional images by the relative sizes of the data, is a well-known form of graphical deception; the viewer can assume that area or volume, not length, is meant. This practice is rare in science, but you should be wary of it in popular graphics.

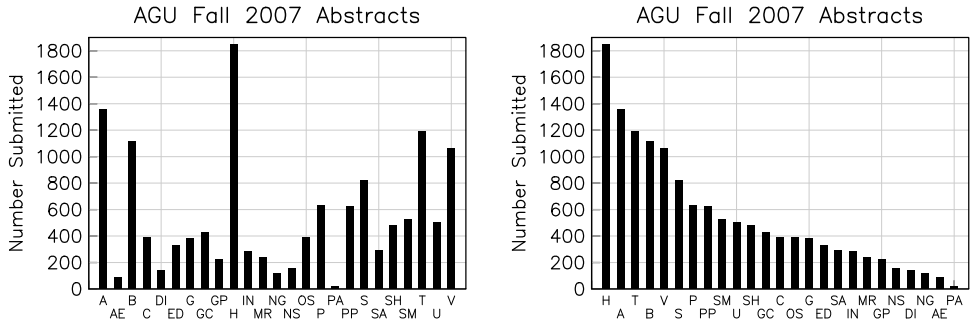

Figure 1.3:

relative sizes: for example, it is difficult to estimate the ratio of GC to OS.

To understand the problem, and its cure, we should realize that what we we try do with this and other graphics is to to **encode** something – in this case a quantity – into visual form. The question to ask of any graphical design is, what visual element does it use to mark relative sizes? Figure 1.1 used area, and the pie charts of Figure 1.1.1 can be viewed as using either area or angle. But our visual system is in fact not particularly good at judging if two areas, or angles, are equal, or estimating their relative sizes. To decide if a graphic is good, or not, we should ask what is the most effective encoding for the what we are trying to show: in this case, relative sizes.[6]

Because we are not good at comparing angles, they are a poor method of encoding – a good, simple graphical rule is never to use a pie chart.[7] Our eyes and brain are better at judging length, either relatively or (even better) in comparing small differences. Encoding by length leads us to another venerable graphic, the bar chart, shown in two arrangements (alphabetical and numerically sorted) in Figure 1.3. Again, the numerically sorted one is better: by putting similar bars near to each other it makes comparisons between them more accurate.

Our visual system does even better in judging relative position along a line, rather than length from a base. This encoding leads to a design (invented by Cleveland) called a dot chart, in which we plot the values along a set of lines with a common scale. Since text is read horizontally, stacking these lines vertically gives space for better labels. Figure 1.4 shows a dot chart of the AGU data. Once again, plotting the data in numerically sorted order

[6] Cleveland, W. S., and R. McGill (1984). Graphical perception: theory, experimentation, and application to the development of graphical methods, *J. Amer. Stat. Ass.*, **79**, 531-554.

[7] Except for introducing the concept of fractions in elementary school.

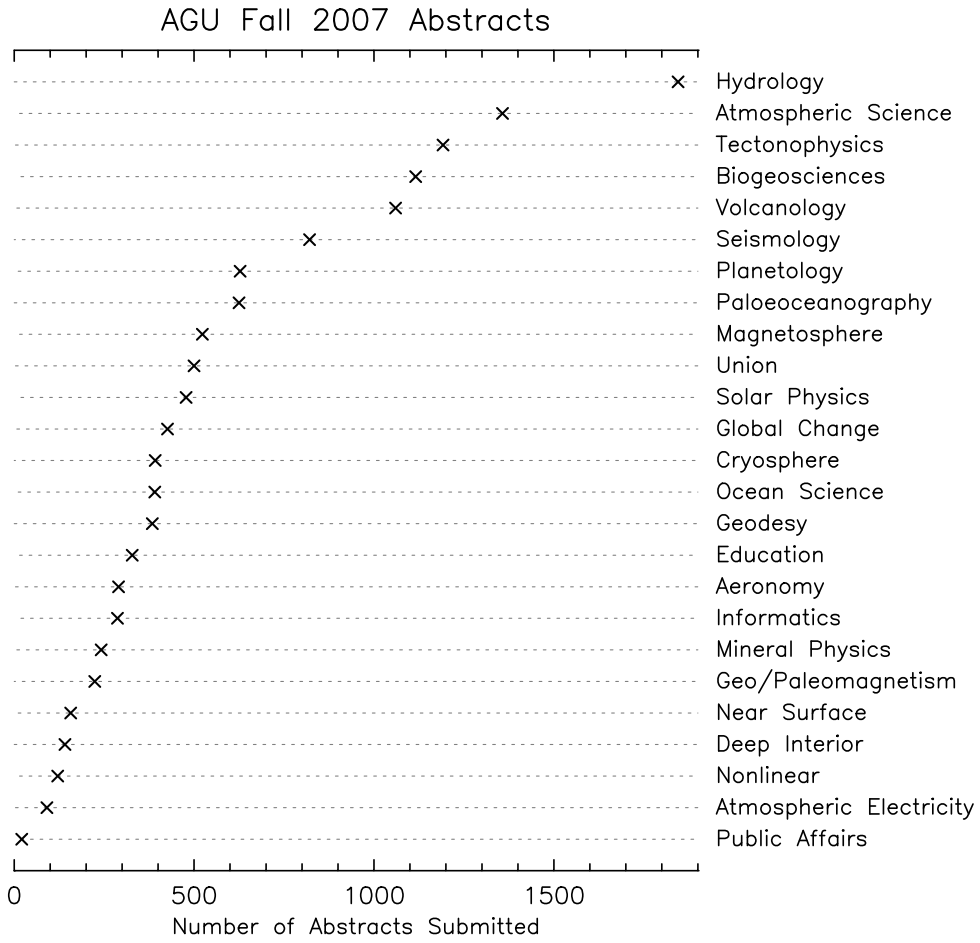

Figure 1.4:

makes relative comparisons much easier than an alphabetic ordering would; and after all the alphabetic order is completely arbitrary.[8] Unlike the bar chart, the dot chart can use a log scale; and with different symbols it can show more than one type of data; though you cannot then sort them all. And it is not too difficult to recover the numbers from the chart; indeed, if you don't need to provide more than two decimal places, a dot chart makes a table unnecessary.

Table 1 lists the different visual encodings, ranked in the order they can be used to discriminate values. What is meant by "position along common

[8] In tables as well as figures, it is worthwhile to think about how to order your data. Never enslave yourself to names or to the alphabet.

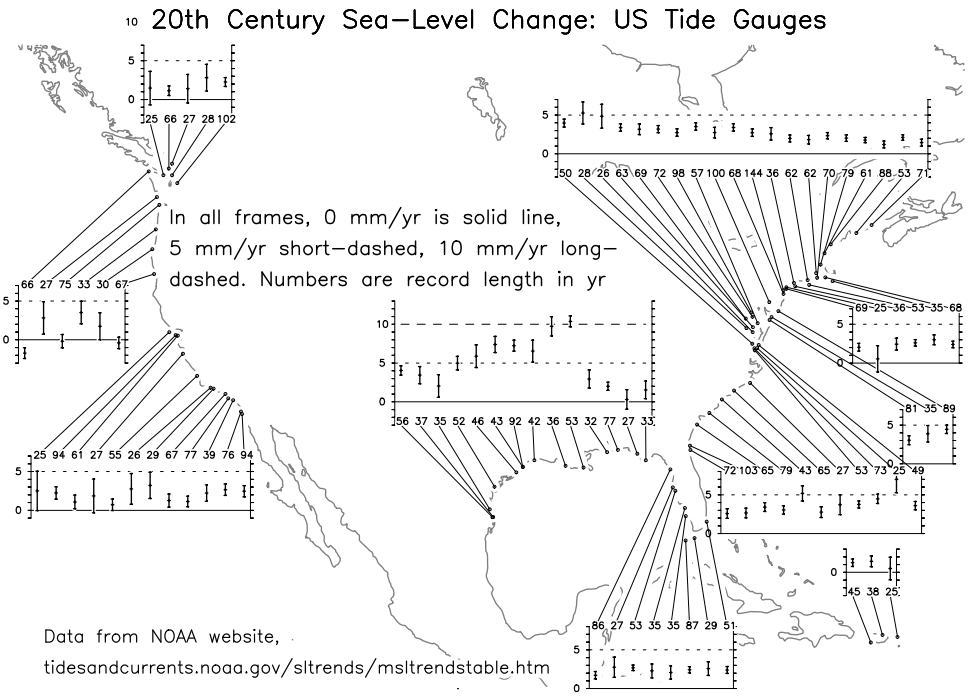

Figure 1.5:

but nonaligned scales" is shown in Figure 1.5, which illustrates (in a modified form) another of Cleveland's inventions, the framed-rectangle plot. This plot shows scalar quantities defined at different locations. This can be done with color or a gray scale, but these are the poorest encodings of all. Bars at each point, which encode the data as length, are an improvement; but even better is to include an identical scale with each bar, which makes it easier to judge differences for regions not close to each other. Figure 1.5 shows the rate of change of sea level at tide gauges in the lower 48 United States, plus nearby US-controlled locations; groups of values are arranged, as much as possible, to lie along common scales, with lines pointing back to the locations. As much as possible a common scale is used for to each group. Two signals are evident: going south along the east coast there is an increase in the rate of sea-level rise, from the collapse of the bulge caused by glacial loading during the Ice Age. And high rates in Louisiana are from loading by the Mississippi delta . Even though the west coast is tectonically active, there is little variation in rates of sea-level change, except for a few sites close to the Cascadia subduction zone and the Mendocino triple junction.

1.1.2 Show What You Want To Show

> These [*sc.* rules] require that the author shall: 12. *Say* what he is proposing to say, not merely come near it. 13. Use the right word, not its second cousin.
>
> MARK TWAIN Fenimore Cooper's literary offenses (1895)

That a plot should show what you want it to might seem obvious, but surprisingly it is not. Perhaps because we know what we are trying to show, we do not easily realize that while something in a plot is clear to us, it may not be to someone else. So you always need to ask, "What is the message in this plot?" and "Have I made it obvious?".

Making the message obvious is especially important for plots shown in a talk: obscurity is disastrous when the plot may be visible for less than a minute. All too often, the speaker who shows such a plot‘ says something like, "If you imagine doing [X] to what is shown on the plot, you would see that...", a remark almost as fatuous as "I know you can't read the table, but if you could you would see that ...".[9]

The three plots in Figure 1.6 may seem to be a silly example, but they are wholly realistic, in that the bad versions are very common. These plots all show time series of one component of displacement for two nearby GPS stations. In the first plot (upper left), about all we can see is that both show a steady, and similar, decrease with time. This might be acceptable if that was the only message, but such a message would be equally well provided by two numbers giving the two rates of change.

It is easy to convey more information, say about fluctuations in the series.[10] To do this, we simply remove the long-term trends, which can still be included on the plot as numbers, and show the residuals; this is done in the plot on the upper right. This improved plot mostly shows something not obvious before, namely that there are motions shared by both sites. All too often we are invited to look at such plots and see that the series are similar, though not entirely, with a phrase such as "some part of the variation is common to both signals" – to which an obvious response is "Yes, that is clear, but how much?".

[9] Perhaps because of the low resolution of projected images (section NN.5), a character size that is easily readable on the page will be illegible when projected, except maybe to the people in the first few rows. Make text, whether in tables, or axis labels and numbers in graphics, twice as high as it would need to be on the page.

[10] I have seen plots of just this type used to show amount of variation, usually, again, with the words "I know you can't see this, but ...".

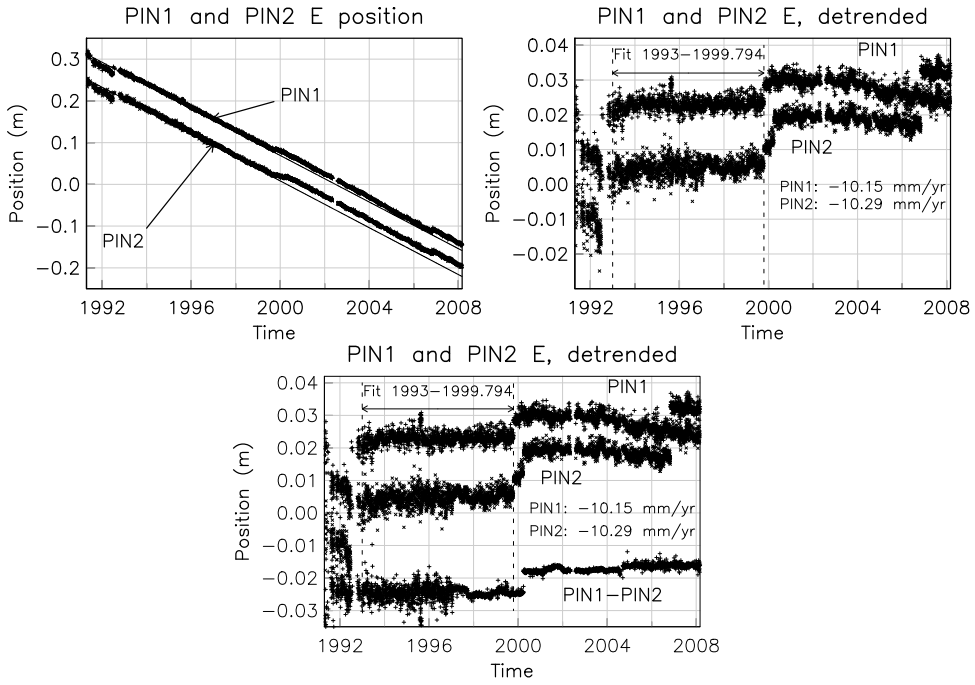

Figure 1.6:

Again there is a simple cure, which is to show the difference between the two, as in the bottom plot. Since the time series and the trend are also included, this plot has all the information, while allowing us to see much more about the behavior of the two series.

This is just one example of a general rule: show whatever you are trying to draw a conclusion from. If you want to demonstrate how well a model fits some data, show the difference (the residuals), not just the model overlain on the data. If you want to show differences between data sets, or ratios of values, show these differences or ratios – *not* the two data sets individually. This approach may mean that plots take up more space, if (for example) you want to show both data sets and their differences. Such plots may seem redundant: why show (say) the ratio as well as the data? But it is better to be redundant than obscure.

Another example of showing what you want occurs when showing several time series whose sum is also important. It is then tempting, and all too common, to stack them on top of each other, adding as you go – but then the only variations that are clearly visible are those in the bottom series and those for the total; the variations in all but the bottom one become obscured.

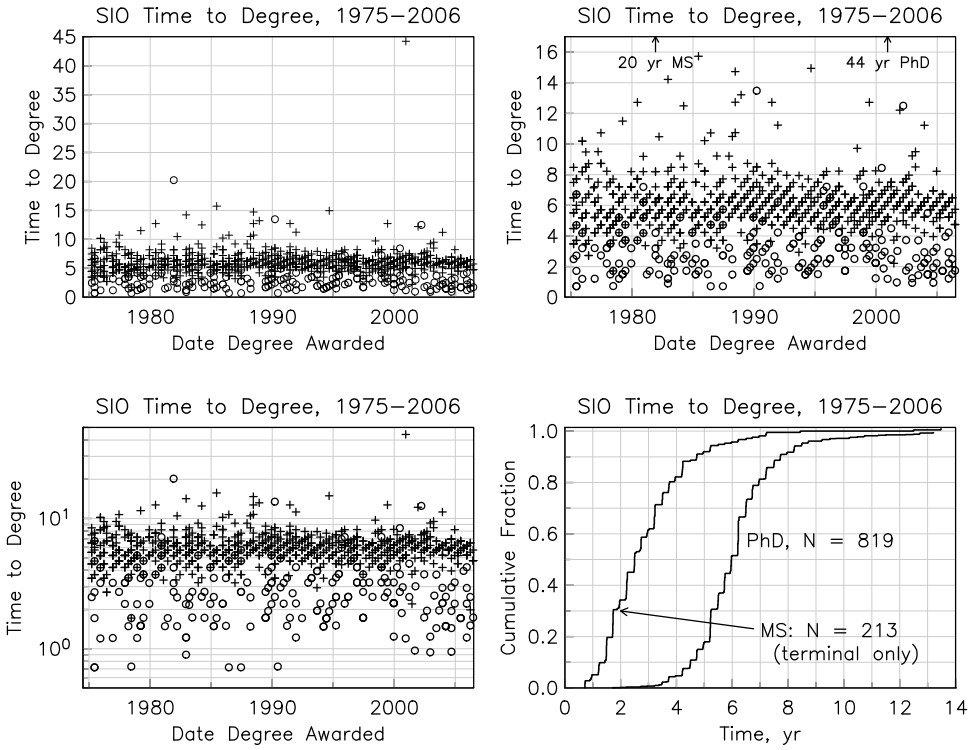

Figure 1.7:

Instead, show all the series, including the sum – and if that takes several frames, so be it.

1.1.3 If There is a Lot of White Space, Revise

Since the purpose of a plot is to provide information about something, having large areas of of white space, provides little information and often indicates bad design, Figure 1.7 shows a common reason for this: an unintelligent choice of plot limits. Many people learn that plots should always include zero, since a plot that does not can exaggerate the size of small fractional changes. Most scientists can spot and allow for this effect, and it is foolish to produce a lot of blank space just to have zero in the plot. In Figure 1.7, the left frame shows what can happen: a scale that includes two outlying points squeezes most of the data into a small area. The cure is simple: use narrower plot lim-

its and indicate the extreme values separately.[11] Another possibility would be to use a log scale, though in this case even a log scale (bottom left) still leaves a lot of white space.

Three of the frames in Figure 1.7 are designed to provide information on how the time to degree has varied over the years: on average, not a lot though the more recent times vary less. In keeping with the advice given the previous section, it is worthwhile to also make a different plot to answer a different question, namely "How are the times distributed?". This can be guessed at from the time series, but is better shown directly, either as a histogram or (as here) using a cumulative distribution – which is especially appropriate because it shows the answer to the question most students would have, namely, what fraction of people take more than X years. The time series in Figure 1.7 show that the distribution of times is roughly the same from 1975 through 2006, so we can aggregate this part of the data to construct cumulative curves.

1.1.4 Do Not Show What You Do Not Need To

> So geographers, in Afric maps,
> With savage pictures fill their gaps,
> And o'er unhabitable downs
> Place elephants for want of towns.
>
> Jonathan Swift (1733) *On Poetry: a Rhapsody*

> 14. Eschew surplusage.
>
> Mark Twain, *op. cit.*

The viewer has to look at and evaluate everything on the plot, so you make her life easier by leaving out as much as you can, saving her the trouble of deciding that what she has seen is irrelevant. This might seem obvious, but experience shows that geophysicists, like geographers of old, often abhor graphical vacua – especially in maps. One reason is the widespread use of the GMT (Generic Mapping Tools) package, which marries many powerful capabilities to two major weaknesses: an arcane and cumbersome set of

[11] These are the PhD thesis of W. R. Gayman (entered 1956, PhD exam 1970, degree awarded 2000) ."Barrier formation in the Gulf of California" and the MS thesis of C. Jerde (entered 1961, MS 1981).

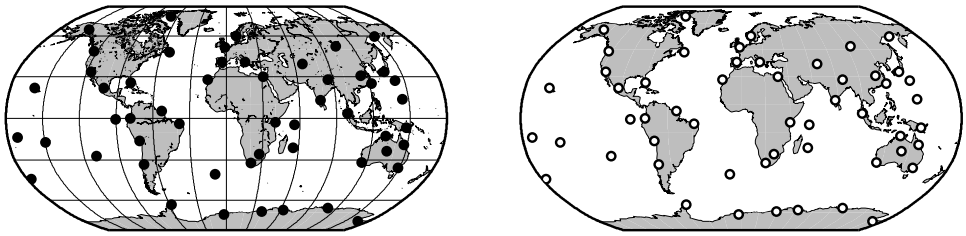

Figure 1.8:

commands, and the ease with which the user can overdecorate the result. A minor example of overdecoration is the pointless striped bar that surrounds most GMT maps; a more important one is the inclusion of unnecessary geographic detail that clutters the plot, often obscuring the data of interest.

Figure 1.8 is a world map, showing the locations of GPS monitoring sites. The left plot (modeled after a figure I have seen) is an example of excess clutter: the latitude-longitude graticule is not needed, and a lot of marks inside major landmasses come from various lakes and rivers. On the right, both of these have been dropped to leave just the outline of the continents, which is all that is needed to see where the sites are.* [12] The right-hand map also uses hollow circles for the symbols. Like the thin line along the coasts (something found in good maps) this relies on the strong sensitivity of the human visual system to edges. Because of this sensitivity, the outline of any shape of a uniform shade will be enhanced by an edge that contrasts with the shades on either side.

GMT is not the the only source of overdecoration, since many commercial graphics packages, driven by the aesthetic inclinations of graphic designers to make plots "interesting", allow (and sometimes demand) pointless elaboration: for example, showing bars on a bar chart as 3-D shaded columns. Even if a plotting package has a feature, (*especially* if that package is PowerPoint), it may well be one you should eschew.

1.1.5 On a Good Design, You Can See a Lot

You should not avoid complication, or a lot of information, if it really is needed; the human visual system can absorb a great deal, and you can pack a sur-

[12] For GMT users: to remove the graticule, use a large interval in the `-B` option of `pscoast`; to remove the lakes and rivers, and small islands, set `-A10000/0/1`. Use the `-Dc` option for any global map.

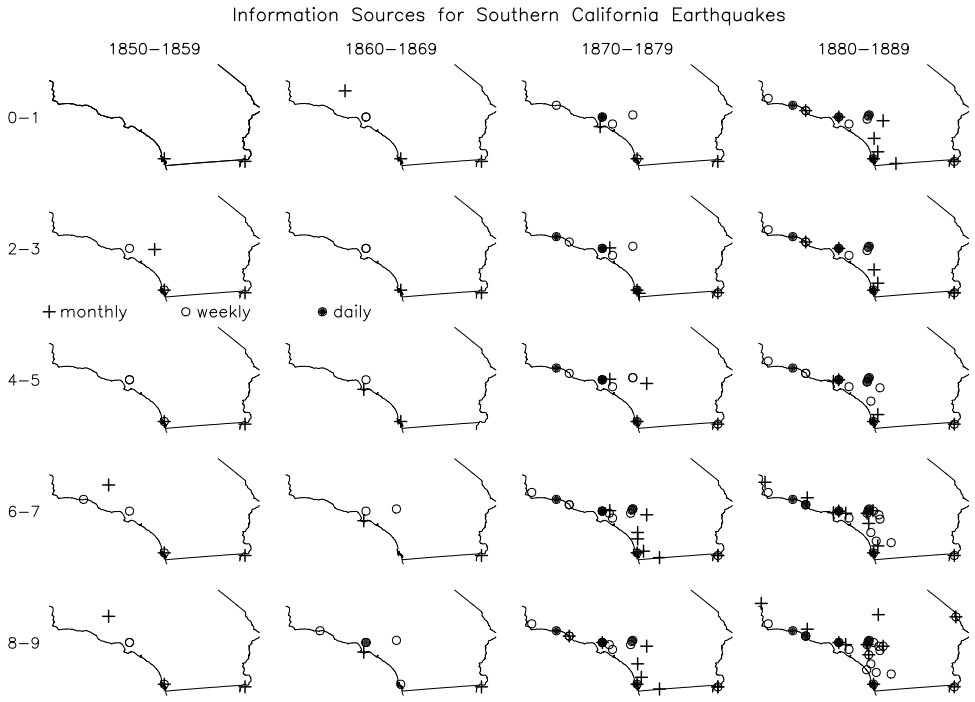

Figure 1.9:

prising amount into a well-designed plot. This is of course truer for plots in papers (or posters) than those used in talks, for two reasons. One is that the viewer has time to study the plot, while in a talk there may be only a minute. Also, printing technology can provide much higher density than all but high-resolution video screens. A typical (current) resolution for these is 640 by 480 pixels, so most figures will use about 300 by 300; this is equivalent to a printed figure about 3 inches on a side.

The second map in Figure 1.8 shows how much can be made legible in a small space. A method that makes use of this is what Tufte calls "small multiples": instead of trying to stuff all your information into one plot, you should show different parts of it on a series of smaller and simpler plots. Figure 1.9 is an example, showing the history of different sources of information on California earthquakes from 1850 to 1890, with different symbols for how often they reported; the more often, the better. The individual figures are simple, but together provide a detailed visual history that shows, among other things, how the first Southern California real estate boom (in the 1880's) caused many weekly newspapers to be started.

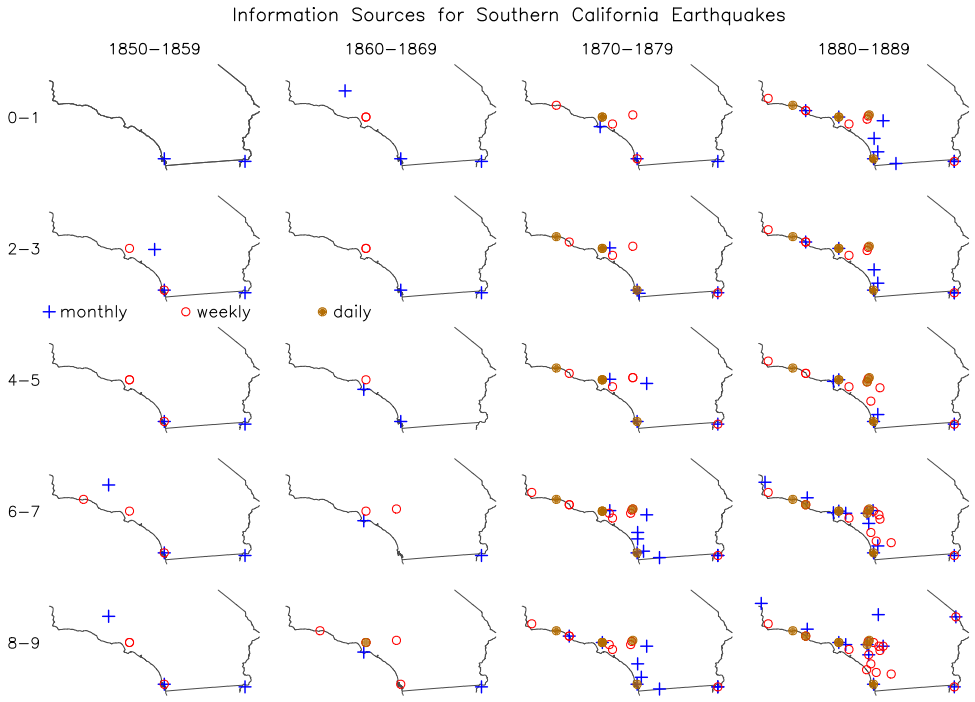

Figure 1.10:

1.1.6 Color: Essential, Useful, or Superfluous?

While Figure 1.9 would be perfectly adequate for a paper, it would not work well in a talk. The time needed to assimilate it can be reduced by differentiating the different information types not just by symbol shape, but also by color, as shown in Figure 1.10. Our visual system (like those of other species that eat fruits) is superb at quickly distinguishing one color from another. Until recently, displaying colors was technically challenging; now it is so easy, and the vividness of color is so tempting, that colors are often overused or abused.

There are several reasons to use color. From good to bad, they are:

1. To mimic the colors we would see if looking at an actual scene (a color photograph) or else to highlight features in a scene that we would not naturally see. Color photographs, or false-color ones, convey much more information than black and white.

2. To display densely pixelized information, such as InSAR data.

3. To distinguish one element of a plot from another, as in Figure 1.10. Color does not automatically cure bad design: while a jumble of black lines may be incomprehensible, the same lines if colored may still be a jumble.

4. To encode, through some ordering of colors, quantitative information. This is the second most common use, and I discuss it further below.

5. To provide a nominal sense of "reality"; for example, using aerial or space photographs as a map base.

6. To provide pointless visual excitement and make the plot more "interesting"; this is common in plots influenced by graphic designers.

As noted in Section 1.1.1, colors encode values very poorly. If you are thinking of using colors to encode value, ask yourself if you really have to. One case where you may need to is showing the variation of a scalar over a two-dimensional area: the "densely pixelized" case above. Colors easily show small-scale variation – not surprisingly, given that color variations are ubiquitous in natural scenes.

Any good map (I will venture to offer Figure 1.15, below) shows how coloring allows us to effortlessly separate different elements of a plot. But for simple plots, if you do not need color, do not use it, since it then becomes just another thing that the viewer has to decide doesn't really matter.[13]

But, if you have other things to show than just the value of some scalar, the colors are often *too* noticeable and obscure other data. Elaborately colored topography (easy to do in GMT) is the commonest example. You can do much better by keeping colors in the background: which is to say, pastels are good. If some large area of your plot will have the same hue, whitening it to a pastel shade will keep it from overwhelming the rest of the plot. A simple example is using a light gray grid: available but unobtrusive. The next two figures show the contrast between using bright and (some) pastel colors for a map of seismic intensity from a hypothetical earthquake; because of the variations from geology, this is the kind of spatially-varying quantity that would be difficult to contour. Figure 1.11 uses bright (saturated) colors throughout; Figure 1.1.6 restricts them to those which take up only a small part of the plot. Letting

[13] And, until journals abandon the creation of paper copies, there are good economic reasons to avoid color, which is still much more expensive to print many copies of. And the reader of your proposal may not have a color printer available.

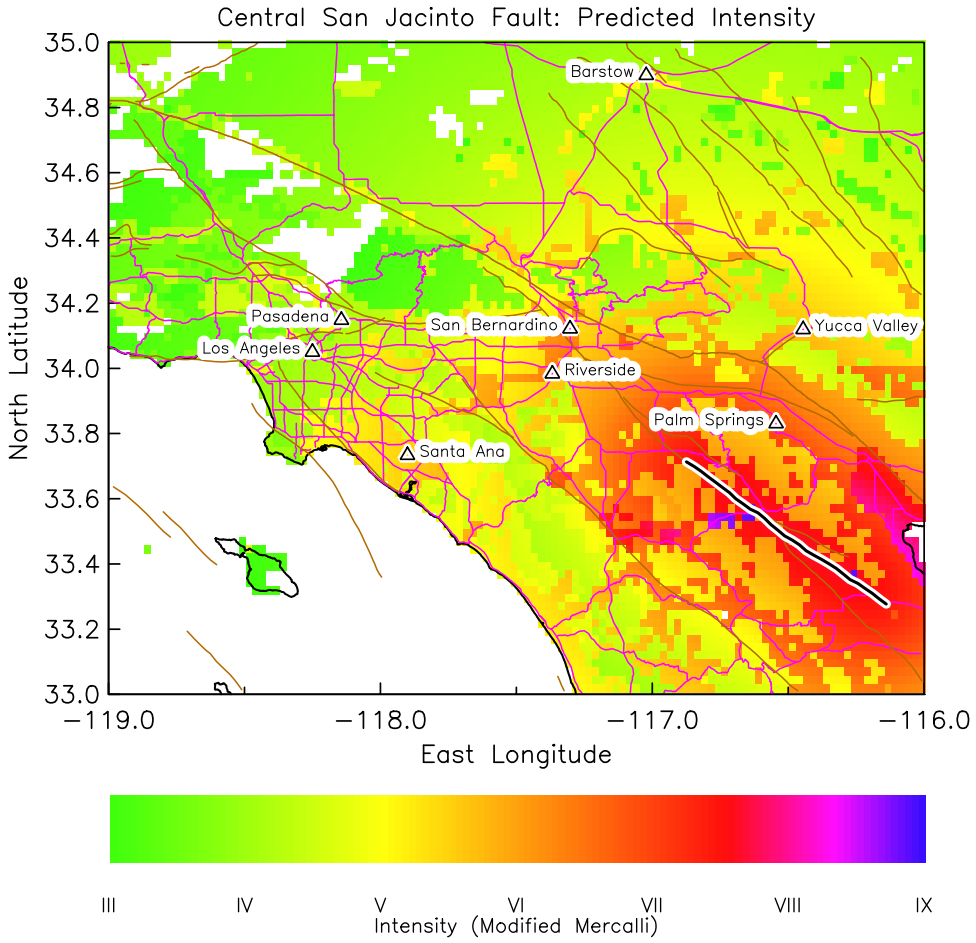

Figure 1.11:

the background approach whiteness makes it easier to see anything overlain on it: in this case, other faults and the state highway network.

A final point about color is that you should *never* assume colors will come out just as you thought they would – at least, not unless you are willing to learn something about how colors are specified.

You should also keep in mind is that a part (up to 7%) of your audience may have much less ability to distinguish colors than the rest do.[14] The commonest case is that red and green become indistinguishable. The best single

[14] This is the percentage for men; the rate for women is much lower. There is some evidence that many women might have additional types of color receptors, and may therefore distinguish colors better than men do.

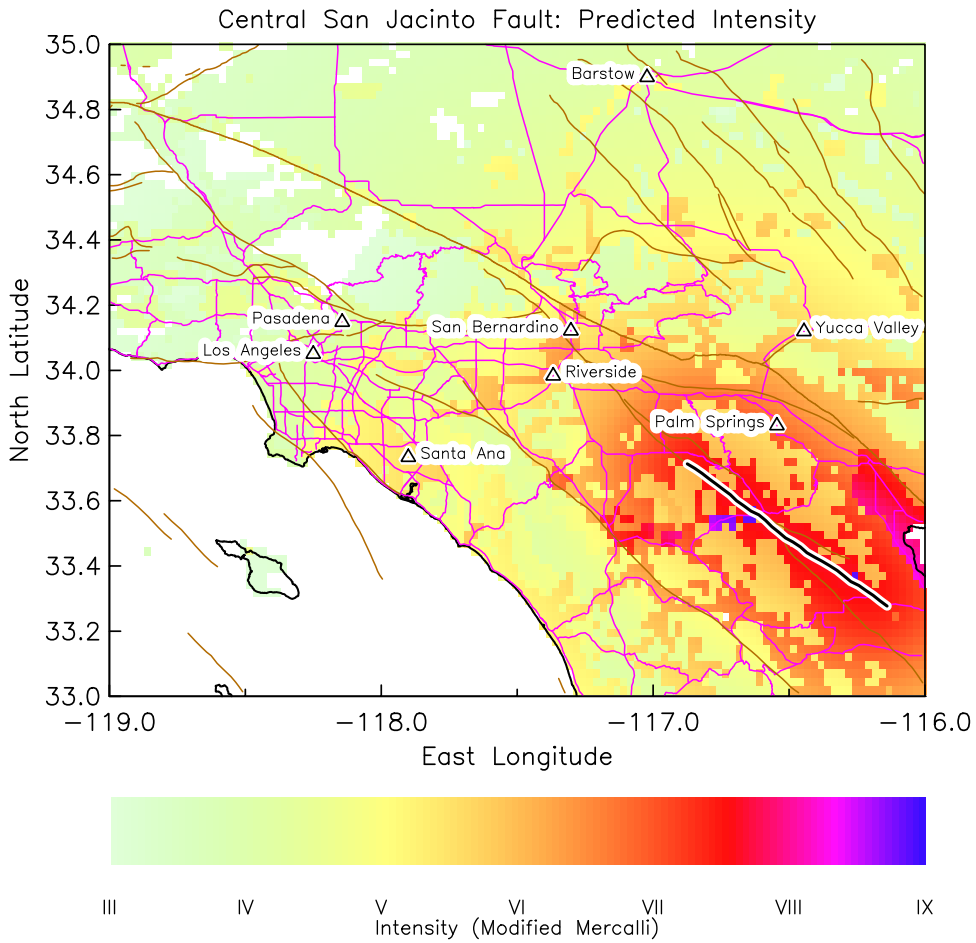

Figure 1.12:

colors to distinguish points or lines from each other are black, red, and blue.

Colors used to encode value have to be chosen carefully for the results to make sense to those with deficient color vision; in particular, the common "spectral sequence" does not work well – but then, it is almost *never* a good choice anyway.

1.1.7 Avoid Realism in Showing Surfaces

> Although the establishment of a curved surface of this kind seems to require the three dimensions of space, we possess a notation as simple as it is expressive, by means of which it is easy to replace the constructions that we have just indicated in space by others effected on a plane surface. ... This process is altogether the same as that which is employed for painting to the eye elevation of land on topographical plans, carefully raised and traced out.
>
> L. Lalanne (1843/45). "Graphical representation of laws with two variables"

Almost all graphics packages allow you to make perspective views of a surface, for example, one defined by a function of two variables $f(x_1, x_2)$; this is often referred to as "2.5D data". Perspective views produce an image that the visual system can easily interpret as three-dimensional, especially if the software included sophisticated rendering algorithms for simulating lighting and shading.

But though such images are easy to recognize, they cannot be interpreted quantitatively, and should be avoided in scientific use. A much better method has been available, for a long time: a contour plot. It takes practice to learn to interpret such a representation, but it is probably a safe assumption that an audience of scientists will know how to do this – especially an audience of earth scientists, who have to "read" contour maps of topography.

As an example, Figure 1.13 and Figure 1.14 show a perspective and contour representation of data that define a function of two variables, namely the mean temperature in Halle, Germany, as a function of time of day and time of year. [15] The original data were published in that paper, and I have used modern software to make the plots. The perspective plot immediately shows one of its defects: for any viewing position some portion of the surface will be invisible or badly foreshortened. Being able to change the viewpoint

[15] I pick this dataset because it was used in the first non-topographic contour plot ever made, by Leon Lalanne, as part of the paper from which I took the quotation at the head of this section. This is an Appendix to L. F. Kaemtz, C. Martin (trans.), L. Lalanne, and C. V. Walker (trans.) (1845). *A Complete Course of Meteorology, with Notes, and an Appendix, Containing the Graphic Representation of the Numerical Tables* (London: Hippolyte Baillière). The original appendix is L. Lalanne (1843). Appendice sur la representation graphique des tableaux météorologiques et des lois naturelles en général, in the French translation of Kaemtz' German original. For the history of contour plots and related developments, see Hankins (1999); Palsky (1996).

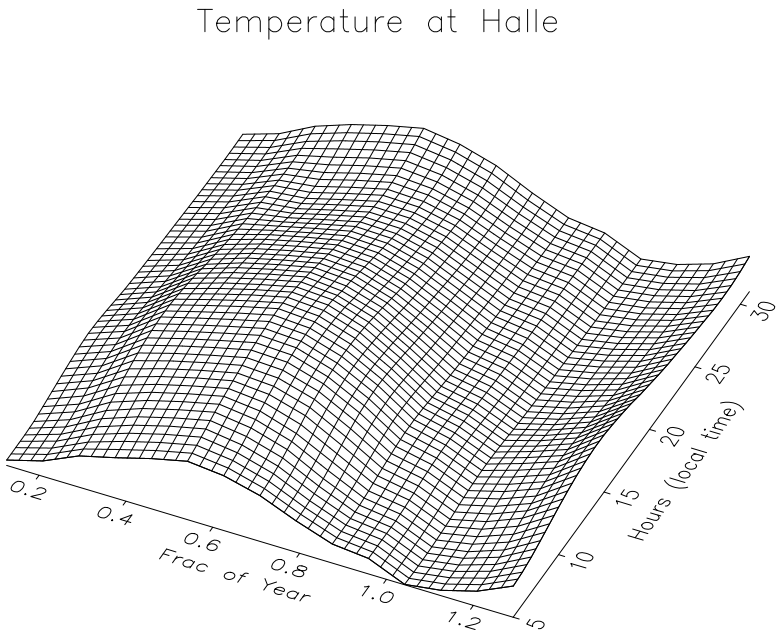

Figure 1.13:

would solve this, but that is an expensive patch to a poor initial design. And not even a dynamic display can show what value of temperature corresponds to any particular pair of times.

In contrast, the contour plot shows the entire surface, makes it possiboe to estimate the value for any pair of times, and allows us to get a sense of, for example, the daily variation at a particular season – though in keeping with Section 1.1.2 if that is what we want to show, we should use a separate plot to do so.

But, as we noted in Section **??**, a contour plot is hard to read if it shows a complicated surface, not that this makes 3D views any more useful. Since landforms can be very complicated, if you need to represent such a surface, you should investigate the techniques that cartographers use to represent surface topography (Imhof, 2007); since landforms can be very complicated.[16] But even topographic maps sometimes use a special symbol for "intricate topography". Layer coloring, of which Figure 1.1.6 is an example, is probably the only method that can be used for really complex surfaces. Even though it is easy to do (especially with GMT) it should not be used for simple situations

[16] Two websites with many useful ideas are `http://www.reliefshading.com` and `http://www.shadedrelief.com`.

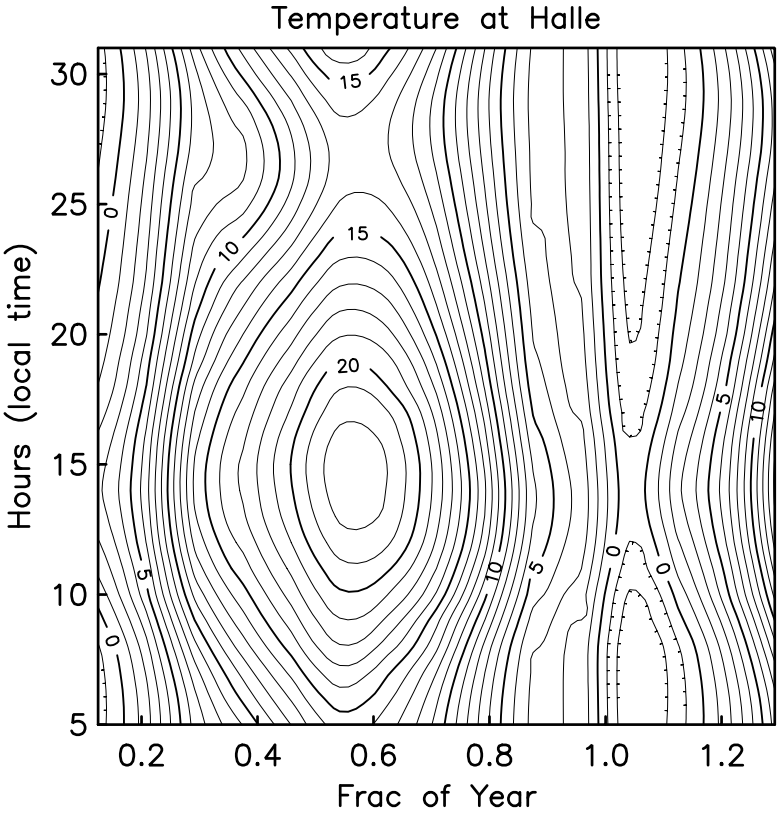

Figure 1.14:

such as Figure 1.14. You should first try to make contour lines more legible by thoughtful choice of contour intervals, by using varying line thicknesses, and in extreme cases using varying colors.

It can be also be useful to combine contours with another method from cartography: applying a very light shading to show local slopes. Figure 1.15 shows this applied to some actual topography, with yellow on the "sunlit" slopes (those facing the NW) and gray on the "shaded" ones (facing SE). This range of shaded is better than the more common white-to-gray shading because the darkest gray can be paler, and any flat area is left white.

1.2 Conclusion

I have tried to offer some precepts for what makes for effective graphics: "effective" meaning that the viewer will get whatever message you are trying to

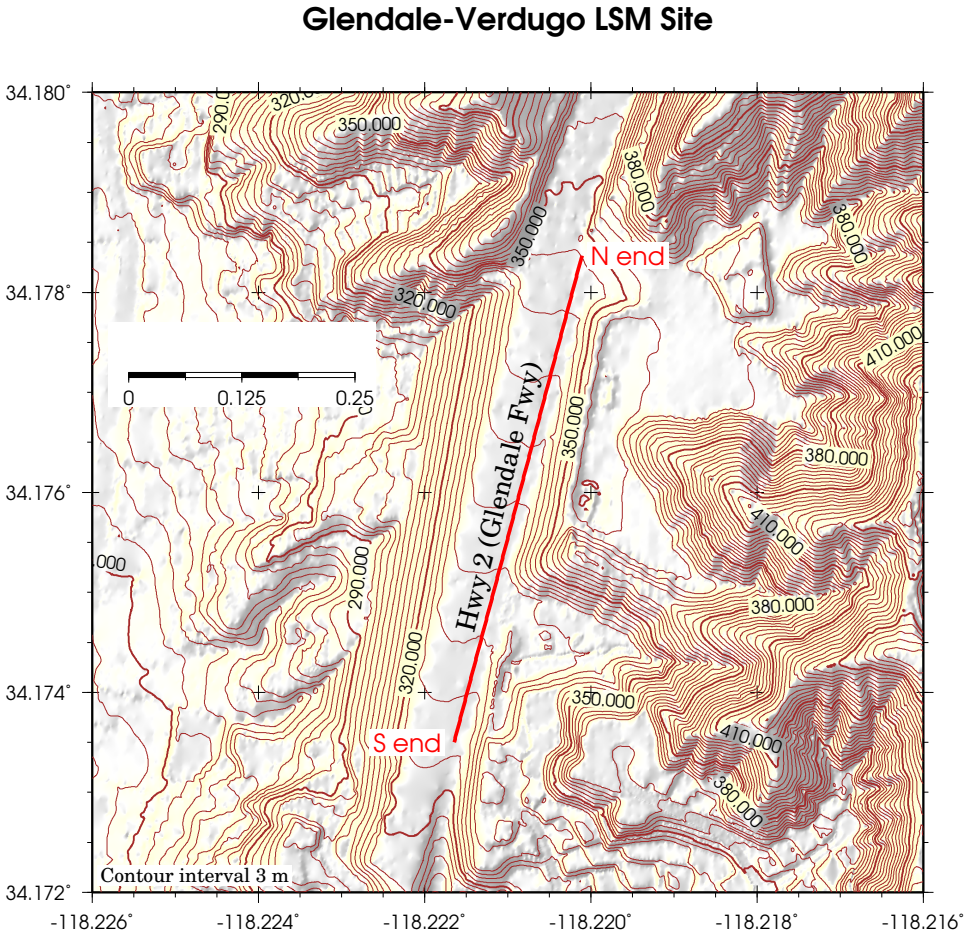

Figure 1.15:

convey. These suggestions are only useful if you start by asking, for each figure you create, what the message is, and indeed if a figure is the best way to convey this message. Often it is, but only if you are prepared to spend effort on designing it, and evaluating what visual elements to include and exclude. As with any other attempt at communication, having someone else look at it is the best way to find out how clearly a graphic makes your point. With luck, you may find that after careful design, much revision, and the use of the principles set forth here, the viewer will take one look and say “That’s obvious”; then, if your idea is as good as your graphic, you are well on your way.

1.3 Specifying Colors

The vast commercial importance of color, and the complexity of how we preceive it, means that there is a large literature with much complexity of its own. This section is my attempt, as a non-expert, to summarize a small part of this, paying particular attention to what is most important to know if you want to specify a color in quantitative and (somewhat) reproducible terms. This is easier given a mathematical background; these notes assume a little knowledge of basic physics, but also require that you know something about vector spaces.

The reason vector spaces are important is that they are the mathematical way of describing preceived colors: "perceived color" because the first thing to realize is that light itself is not colored: color is something we preceive. What this color is depends on the spectral distribution of the photons entering the eye; but these photons stimulate just three kinds of receptors, the cones.* [17] So only three signals flow to the brain, which (somehow) combines them to cause the sensation we call color. This includes different shades of gray from black to white,

This makes a three-dimensional vector space adequate to describe colors – with the caveat that any specification only applies to how colors are perceived in a standard situation: usually a disk of pure color, on a black background, with the disk having a particular angular size. The actual appearance of a color depends on the colors surrounding it: what we see as a dark yellow disk against a black background will become a brown one if the background is white. This complication goes well beyond what we can do with three dimensions.

Given a three-dimensional space, such that a color is given by a vector $\mathbf{x} = (x_1,\ x_2,\ x_3)$, we would like it to have the following properties:

1. Any color has a unique $\mathbf{x}$ associated with it: that is, all perceptible colors can be mapped into points in the space. Since any vector space is infinite, we should not expect the converse: the volume occupied by all the perceptible colors will not be all of the space, so some values of $\mathbf{x}$ will not correspond to any color.

2. The values of the coordinates are always non-negative, since we cannot subtract photons.

[17] A fourth receptor, the rods, only contributes in low light levels, when the cones are inactive, so our night vision is devoid of color.

3. It is intuitively obvious how colors map to values of $\mathbf{x}$, and vice-versa: the user can easily see what a given $\mathbf{x}$, or a sequence of colors specified by a function $\mathbf{x}(s)$, will look like.

4. The space is linear, such that if we have two photon distributions that produce $\mathbf{x}_a$ and $\mathbf{x}_b$ respectively, the color produced by the sum of these two distributions would be specified by $\mathbf{x}_a + \mathbf{x}_b$.

5. The Euclidean norm in this space (which is just what we call "distance" for three-dimensional space) corresponds to perceived differences in color; in particular, a just perceptible difference from a color $\mathbf{x}$ corresponds to the same distance from $\mathbf{x}$ in all directions and is the same for any $\mathbf{x}$.

Given that the our visual system evolved to find food, not fit a vector space, it would be a miracle if there was a vector space for which all of the above was true – and that miracle did not occur. In particular, D and E are not compatible, something which seems to have caused a lot of unneccessary trouble as people have tried to develop color spaces based on simple geometric solids in a Euclidean space. Unfortunately, it is also difficult to reconcile C and D: the spaces best suited to describe colors quantitatively are not particularly intuitive – which is why I have written these notes.

There can also be a fifth aim, born not of human perception but out of the need to reproduce colors, either through colored lights (as in video) or by coating a surface with substances which emit different photon distributions when light shines on them; examples are inks on paper, dyes on cloth, and paints on canvas) This aim is:

1. A value of $\mathbf{x}$ should specify how to actually *create* an appropriate spectral distribution of photons.[18]

This last requirement is of course specific to whatever technology is being used; hence the problem that colors that are "the same" according to one specification appear different when created using different methods. In terms of vector spaces the problem is that $\mathbf{x}$ in a space defined independently of any technology can map into different $\mathbf{x}$'s spaces tied to particular technology; I

[18] I say "an appropriate" rather than "the appropriate" because many different spectral distributions correspond to the same $\mathbf{x}$, and thus the same perceived color; these spectral distributions are said to be **metameric**.

will give some examples below. It is also the case that none of the spaces actually available for generating images cover all the colors that can be perceived; that is, they violate A above.

With all this as background, we can proceed to particular vector spaces used in the business and science of color. In keeping with the usage of the field, we designate these spaces (in part) by the letters used for the three variables, even though this leads to a plethora of names.

The most "fundamental" of these spaces uses coordinates called X, Y, and Z; this space is called the CIE 1931 XYZ space. The CIE stands for the Committe Internationl de l'Eclarage,* [19] a standards-setting body for illumination engineers; the 1931 is when this particular standard space was put forward. The aim in creating the XYZ space was to use coordinates that would roughly match the three signals sent from the eye to the brain; for this reason it is called a tristimulus space.

To start to explore this space, consider the transition from total dark to pure white, without any sense of a particular hue. The XYZ space attempts to make this transition coincide with the Y axis. Where to put the origin is easy: black, which is no stimulus at all, corresponds to $\mathbf{x} = 0$. "White" is more challenging to define, for two reasons. One is that we can have white for different amounts of energy entering the eye, up to a level that causes damage. So the CIE space defines the maximum value of Y, which is set to one, to correspond to a given energy flux. ??? The other problem is that, as noted above, different spectral distributions can produce the same color, which holds as much for white as for anything else. So ??? something about D65 ???.

Because Y corresponds to brightness, it is very common to work, not in XYZ space, but in what is called Yxz, formed by dividing the other two coordinates by Y. Now consider a series of "pure spectral colors"; that is, narrowband spectral distributions. As the wavelength of the light changes, the color seen will go from red to yellow to green to blue, and this will trace out a space curve in the XYZ space. If the energy is such that $Y = 1$, this space curve maps to a curve in in the xz plane. If we take varying combinations of lights with the two colors at the opposite end of the curve, we see purple, so this curve is said to be closed by the **line of purples**.

Figure xzcie shows this curve and line in the xz plane, in what is called a **chromaticity diagram**. The region inside the curve covers all colors with $Y = 1$, white included, For smaller values of Y we can draw a similar bounding

[19] International Committee on Lighting: `http://www.cie.???`.

curve in each xz plane; each such curve will outline a region of darker colors. But as Y decreases, the extent of this region shrinks, until at $Y = 0$ it is a point at the origin. All possible colors thus fall inside a volume in XYZ space that can be described as an irregular cone-shaped object with its tip at the bottom, and the chromaticity plane at the top. Figure chroxyz shows the contours of this solid, and Figure ch2roxyz two views of it from the bottom.

You should realize that Figure xzcie, and every other version of it you have ever seen, is a cheat. Because the XYZ space is linear, a color inside it can be formed by summing any three colors corresponding to points in the diagram (after suitable normalization for the increase in brightness); but this summation can only cover a triangular region in this plane. ??? explain summation Because the outer boundary is convex, there is no triplet of colors that can be combined to cover the whole of the color region; and no method of reproduction that will produce them all. (Different triplets of spectral colors can cover the whole region, though). So even though I have colored in the whole diagram, the colors you see in it only approximate the true ones.

So now we have a vector space, or color space, that satisfies A, B, and D. Many others have been suggested, either to more closely approach C or E, or to fit within the constraints imposed by a particular technology (F). We can only review a small number of these, and will take one example of each.

For those professionally concerned with color, requirement E ranks especially high, since in a commercial or artistic setting the closeness (or otherwise) of colors can be all-important. Distances in XYZ space do not map well to perceived differences in color; Figure ciexyz more than hints at this, since a huge part of the color region is green, and only a small part given to yellow.

Bibliography

Hankins, T. L. (1999), Blood, dirt, and nomograms: a particular history of graphs, *Isis*, **90**, 50–80.

Imhof, E. (2007), *Cartographic Relief Presentation*, vol. visualization, surfaces, ESRI Press, Redlands, Calif.

Palsky, G. (1996), *Des Chiffres et des Cartes: Naissance et Developpement de la Cartographie Quantitative Francaise au XIXe Siecle*, Ministere de l'enseignment superieur et de la recherche, Comite des travaux historiques et scientifiques, Paris.