Environment and Planning A 1998, volume 30, pages 1547–1561

Visualizing georeferenced data: representing reliability of health statistics

A M MacEachren, C A Brewer
Department of Geography, Penn State University, University Park, PA 16802, USA;
e-mail: alan@essc.psu.edu
L W Pickle
National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782, USA;
e-mail: lwp0@cdc.gov
Received 23 January 1997; in revised form 25 May 1997

Abstract. The power of human vision to synthesize information and recognize pattern is fundamental to the success of visualization as a scientific method. This same power can mislead investigators who use visualization to explore georeferenced data—if data reliability is not addressed directly in the visualization process. Here, we apply an integrated cognitive-semiotic approach to devise and test three methods for depicting reliability of georeferenced health data. The first method makes use of adjacent maps, one for data and one for reliability. This form of paired representation is compared to two methods in which data and reliability are spatially coincident (on a single map). A novel method for coincident visually separable depiction of data and data reliability on mortality maps (using a color fill to represent data and a texture overlay to represent reliability) is found to be effective in allowing map users to recognize unreliable data without interfering with their ability to notice clusters and characterize patterns in mortality rates. A coincident visually integral depiction (using color characteristics to represent both data and reliability) is found to inhibit perception of clusters that contain some enumeration units with unreliable data, and to make it difficult for users to consider data and reliability independently.

Introduction

How to deal with variation in data reliability is an unresolved question for data-driven scientific visualization in general and for visualization of georeferenced health statistics in particular. When data are georeferenced, visualization is intended to prompt hypotheses about patterns and clusters in geographic space, and thus about the role of location in human and environmental processes (MacEachren, 1995). For epidemiological research, mapping is a primary visualization tool used to identify geographic clusters (that is, 'hot spots') and other regular features of disease patterns (Mason et al, 1975; Pickle et al, 1987; 1990) and to compare these features with patterns of potential etiologic agents (Blot and Fraumeni, 1976; Blot et al, 1978; Croner et al, 1992; Mason et al, 1975; Pickle et al, 1987). For example, a map of oral cancer death rates among white females in the United States prompted a study of occupation and lifestyle in North Carolina that identified snuff dipping as a major risk factor for this tumor (Winn et al, 1981). Geographic variation in data reliability across a georeferenced data set can result in identification of false patterns or failure to notice real patterns; the equivalent of type I and II errors with statistical hypothesis testing (MacEachren and Ganter, 1990). A limiting factor in use of mapping as a visualization tool for epidemiological research, then, is that the aggregate georeferenced statistics depicted with health maps vary in reliability, reliability varies geographically, and the few methods that have been developed for adding reliability depictions to maps of enumerated data are untested.

Variation in reliability of health statistics is a direct function of what is known in epidemiology as the *small area problem* (Diehr, 1984) and in geography as the *small number problem* (Jones and Kirby, 1980). If the population at risk is small, derived measures such as rates and ratios are unstable and, therefore, unreliable estimates of the true values of these measures in the population. One solution to this problem is to aggregate data over time or geographic units, but this may mask important local-scale patterns. Another alternative is to blank out areas on the map which have unreliable statistics, but this has been shown to interfere with pattern recognition on the map (Lewandowsky et al, 1995). A third alternative is to depict data with population cartograms, for which areas with small values become visually insignificant (Dorling, 1996). Cartograms based on total population, however, do not necessarily cause small numbers for particular health statistics to result in small units represented on the map (this only occurs when there is a high correlation between total population and the particular health statistic). If cartograms are based on the particular health data of interest (for example, heart disease or lung cancer rates), rather than on total population, maps are difficult to compare (because enumeration units differ in size and shape from one map to another). A newer alternative is to model the underlying data to produce 'predicted' statistics which account for variation in the observed data because of small numbers. Unfortunately, these methods require specification of a parametric model for the data which is often unknown; furthermore, several recently proposed Bayesian methods have been shown to produce biased results and distorted map patterns (Pickle and White, 1995). Embedding the visual representation of reliability directly into the maps avoids the loss of resolution associated with data aggregation, the comparability problems of cartograms, the statistical bias of proposed Bayesian methods, and the need to specify a parametric model for the data.

Here we report on development and empirical assessment of methods for embedding data reliability representations into choropleth maps[1] of health data. Methods devised are grounded in an integrated cognitive–semiotic framework for cartographic representation (MacEachren, 1995). This research was the second stage of a coordinated research project, the first stage of which explored the effect of map color schemes on health data visualization (Brewer et al, 1997). This research was, in turn, part of a larger effort designed to inform decisionmaking (Pickle and Herrmann, 1995) related to the *Atlas of United States Mortality* (Pickle et al, 1997).

Background

Our overall research project had both general and applied goals. The general goal was to develop a conceptual framework for selecting representation methods to depict both data and data reliability on statistical maps. The applied goal was to enhance the potential of the *Atlas of United States Mortality* as an exploratory visualization tool (and as an information presentation device). The approach we take to both goals draws upon perspectives from cognitive linguistics, anthropology, and color theory (study of 'basic' colors) (Berlin and Kay, 1969; Boynton et al, 1989), vision research (dealing with color deficiency) (Regan et al, 1994; Travis, 1991), psychology (concerning color preference and harmony) (Granger, 1955; Whitfield and Wiltshire, 1990), statistics (measurement and representation of uncertainty) (for example, Carr et al, 1992), cartography (dealing with the impact of color vision deficiencies on map design) (Olson and Brewer, 1997), and the semiotics of map color schemes and methods for reliability representation (Brewer, 1994; MacEachren, 1995).

In this section we review our work on color choices for data representation (the initial stage of the overall research—reported in Brewer et al, 1997). Then we outline our framework for developing integrated data–reliability representations for georeferenced data.

[1] Choropleth maps depict aggregate values for contiguous enumeration units by using some form of area fill to indicate each of several data ranges into which the aggregate values are divided.

Color schemes for representing georeferenced health data

We began with a systematic analysis of color selection for maps of aggregate quantitative data. The goal of this analysis was to develop a logical framework for selecting sequential and diverging color schemes that avoid potential perceptual and conceptual problems.[2]

In relation to conceptual problems that might be encountered by map users, there were two specific objectives. The first was to implement principles of map semiotics that call for matching the ordinal qualities of health statistics to symbols having a systematic (ordered) progression of color lightness and color saturation (Brewer, 1994; MacEachren, 1995). The second objective was to identify color hue pairs for diverging symbolization schemes that avoid naming confusions (for example, purple and green are never confused regardless of lightness or saturation, whereas purple and red might both be called a kind of 'red', or purple and blue might be called a kind of 'blue').

Perceptually, there were also two specific objectives. The first was to accommodate color blindness in the selection of hue pairs for diverging schemes (to make sure that the hues used as opposite ends of the scheme appear to be different to people with various forms of color vision deficiency) (Olson and Brewer, 1997). The second perceptual objective, for both sequential and diverging schemes, was that color schemes should result in all data categories being visually discriminable.

Based on these conceptual and perceptual objectives, a robust set of principles was devised for selecting map color schemes to depict enumerated data. Eight color schemes meeting the principles were empirically assessed. All performed well, although a monochrome (gray scale) scheme (in which only color lightness is varied) resulted in significantly more map-reading errors and lower preference ratings than did any of the seven schemes that combine variation in color hue, saturation, and lightness (Brewer et al, 1997).

Three successful map color schemes identified in this first stage of research were carried over to the second stage (reported here). The first color scheme is a 'diverging' spectral scheme that ranges from dark blue for low values through light yellow for the nation's median to dark red for high values. The second color scheme, the most effective 'sequential' scheme from the first experiment, ranges from light yellow for low data values through medium orange to dark red for high data values. The third scheme, a 'diverging' scheme based on two hues (rather than hues across the full spectrum), ranges from dark green for low values through light gray for the nation's median to dark purple for high values. The challenge for the research stage reported here was to develop and test an approach to reliability representation that is compatible with these color schemes for data representation.

Reliability representation for georeferenced health data

Reliability representation for georeferenced data is a topic that has received increasing attention from researchers over the last several years (under a variety of labels, including data quality visualization and representation of uncertainty). Several authors have addressed the question of symbol methods for reliability representation at a conceptual level. Most of the research to date has focused on environmental data, using approaches ranging from cartographic semiotics (MacEachren, 1992), to statistical modeling (Goodchild et al, 1994), to exploratory data analysis (MacEachren et al, 1993).

Within a semiotic approach, MacEachren (1992), McGranaghan (1993), and van der Wel et al (1994) have all presented frameworks based on Bertin's (1983) graphic variable typology (with some extensions). MacEachren argued that two graphic variables missing from Bertin's original set, color saturation and clarity, are particularly well suited

[2] Sequential schemes are ordered from high to low, whereas diverging schemes are arranged with an order that increases in both directions from a midpoint (Olson and Brewer, 1997).

to depicting data reliability—because both have the potential to suggest uncertainty or lack of precise knowledge. A comprehensive syntactics for reliability representation is provided by van der Wel et al (1994).[3] This syntactics specifies appropriate links between graphic variables and kinds of reliability information.

When representing data reliability, we must be concerned, not only with a logical syntactics to match reliability measures with visual depictions, but with the way data and metadata are linked and accessed (that is, the 'interface' style). The fundamental choice here is whether data and metadata will be coincident (on one map) or adjacent (on two maps) (MacEachren, 1992). For spatially coincident data and data reliability representations, a further distinction can be made between representation forms expected to result in *visually separable* components and those expected to result in *visually integral* components.

Research dealing with perceptual organization distinguishes between 'separable' components of a visual depiction, for which viewers are able to attend to one element of a visual scene and ignore others (Pomerantz, 1985) and 'integral' components, in which viewers find it easier to attend to the components as a unit (Carswell and Wickens, 1990). Psychological research on separable versus integral components has focused on perception of visually isolated symbols that occupy a small portion of the visual field (within foveal vision). Related cartographic research has considered how principles derived extend to visual interpretation of patterns occurring across the visual scene (across a map) (Bertin, 1981; MacEachren, 1995).

In the research presented here, one method of separable display (achieved through symbol *overlay*) and one of integral display (achieved through symbol *merger*) were selected for comparison with adjacent displays. The specific overlay and merger symbolization methods developed and tested are detailed below.

Representation methods for mortality map reliability

Because one goal of our research was to inform design decisions for the planned US mortality atlas, there were a variety of initial constraints that we needed to consider in developing potential data–reliability representation schemes. Among the most important were that the atlas will be a paper product, printed in color, and will use choropleth symbolization to depict the death rates (one cause per map) for each of 798 Health Service Areas (HSAs) in the conterminous United States (Makuc et al, 1991). The component of reliability of interest is the reliability with which each rate value reported for an HSA represents death rates throughout the territory included in that HSA. Thus, a measure of variation around the single rate reported for each HSA is the appropriate reliability index to represent.

As noted above, emphasis in this project is on developing representation schemes that combine a choropleth map of the data (using the color schemes described above for data depiction) with reliability information by means of one of three methods: adjacent display, coincident visually separable display, and coincident visually integral display. The adjacent display method devised consists of map pairs in which a data map identical to that used in our previous experiment (Brewer et al, 1997) is paired with a smaller binary reliability map (a map with HSAs classed as either reliable or unreliable).

Coincident displays can be considered to be a special case of bivariate mapping in which two characteristics of one variable are mapped (for example, mean and variance for each enumeration unit) [see MacEachren (1995) for discussion of bivariate mapping as it relates to visually integral and visually separable symbolization]. A representation

[3] The term 'syntactics' as used here refers to a conceptual framework for relating attributes of and relations among sign vehicles (map symbols) to corresponding attributes of and relations among referents (the things signified by the map symbols).

in which the reliability depiction is coincident with, but visually separable from, the data depiction proved hard to achieve, particularly for use with both diverging and sequential color schemes for the data. After consideration of dozens of possible representation methods, the solution arrived at was to build a texture overlay from parallel lines split down their length, with one side black and the other white. When superimposed on a dark color, the white half stands out; when superimposed on a light color, the dark half stands out; and when superimposed on a color of medium lightness the combined contrast of the black and white line components stands out from the color on which they are superimposed.

A practical coincident integral display of data and reliability can be achieved by using one color scheme to depict several classes of reliable data and a variation of that scheme to depict less reliable data (colors of the same lightness are shifted in color space, either toward lower saturation, making them grayer, or toward a different hue). This representation approach implements the general semiotic principles derived from the literature discussed above. Specifically, hue shift was applied to the diverging nonspectral scheme and saturation change was applied to the spectral and sequential scheme.

Methodology

Problem

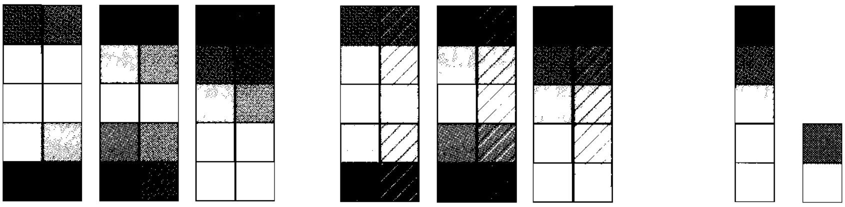
Once a set of potentially effective data – reliability representation schemes suited to the choropleth maps planned for the *Atlas of United States Mortality* had been developed, the next stage of the research involved an empirical assessment of these schemes. Adjacent, coincident integral, and coincident separable display methods for data – reliability representation were assessed for maps in which data categories were depicted with each of the three color schemes detailed above, applied to six mortality data sets representing two levels of spatial clustering, at two map-scale – data-grouping combinations (full-page 7-class and quarter-page 5-class maps).

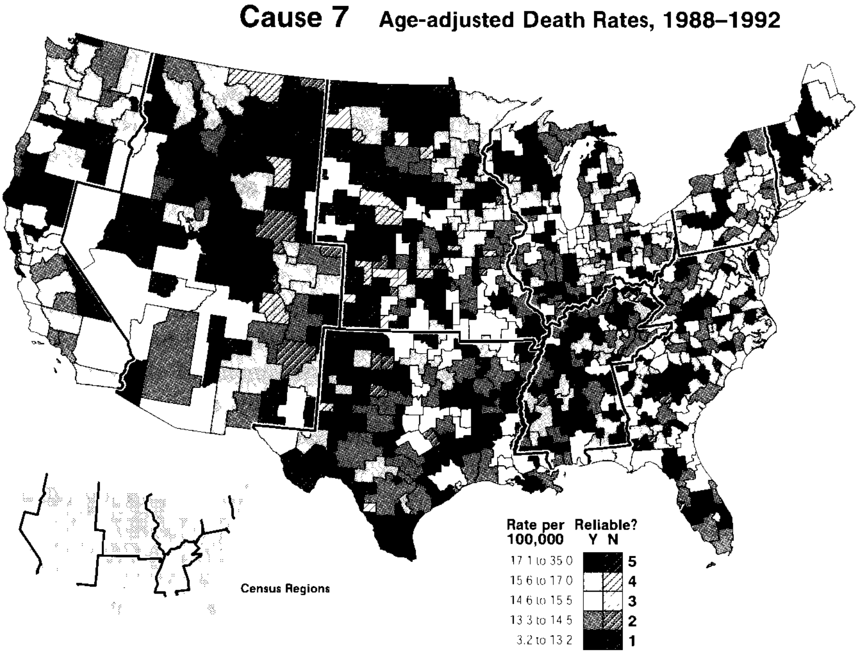
Test maps

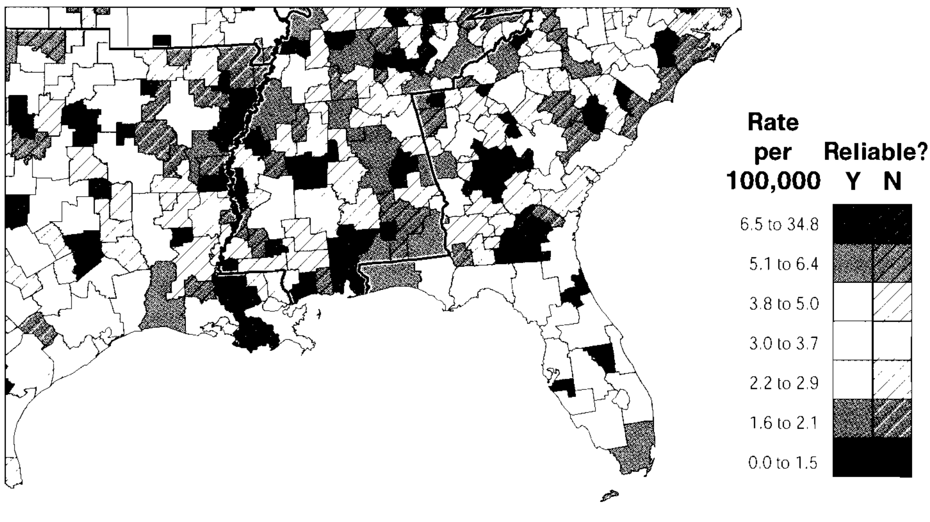
A base map of the conterminous United States, with nine census divisions and 798 HSAs delineated was common to all test maps. Age-adjusted death rates for six different causes were mapped by HSA. Data for the full-page maps were categorized into seven classes according to a variant of quantiles.[4] Straightforward quantiles group one fifth of the HSAs in each class. The lowest and highest fifth of the HSA rates were further divided into equal halves to produce seven classes and provide more detail in the classification of the extreme rates. Of a total of 798 HSAs, 79 with the highest rates fell in the highest class, the next 80 HSAs fell in class 2, 160 in class 3, 160 in class 4 (with the median rate in this middle class), 160 in class 5, 80 in class 6, and 79 HSAs with the lowest rates were in class 7. On the quarter-page maps, data were categorized into five classes with equal numbers of HSAs (that is, standard quantiles). Color plate 1 (see over) includes a set of map legends for the 5-class data – reliability combinations tested along with a sample 5-class test map and a section from a 7-class map.

Data were selected so that three of the resulting map distributions had generally dispersed patterns and three had generally clustered patterns (as reflected by a spatial autocorrelation measure—Moran's I-statistic). The clustered variables had I-statistics between 0.40 and 0.63 whereas the dispersed variables had I-statistics between 0.05 and 0.19. The six maps used here were also used (without reliability symbolization) in our

[4] Quantile data classification (or the modification of it described) is the standard classification method used for mapping at the National Center for Health Statistics. Thus, the method of classification described was used to meet the applied goals of our experiment.




Cause 7 Age-adjusted Death Rates, 1988–1992
Census Regions
Rate per 100,000
Reliable? Y N
17.1 to 35.0
15.6 to 17.0
14.6 to 15.5
13.3 to 14.5
3.2 to 13.2
5
4
3
2
1


Rate per 100,000
Reliable? Y N
6.5 to 34.8
5.1 to 6.4
3.8 to 5.0
3.0 to 3.7
2.2 to 2.9
1.6 to 2.1
0.0 to 1.5

previous experiment. As a result, subject responses can be compared for maps with and without reliability depiction.

Reliability of data was determined through use of the coefficient of variation (CV) calculated for death rates in each HSA. HSAs having a CV greater than 20% were classed as unreliable. At this threshold, reliability for the six causes of death varied from one cause having only three unreliable HSAs to another cause where about half of the HSAs were deemed unreliable. The six binary reliability maps for the sample data are shown in figure 1 (see over) (at the eighth-page size used to accompany the quarter-page adjacent data maps in the experiment).

As in the first stage of research, test maps were complete maps that included a title, legend heading, and legend showing class ranges. Maps were printed on a desktop color printer that approximated the colors specified in the final report from the first experiment.[5]

Subjects

Eighty-four subjects participated in the experiment. Individual subjects completed tasks using all three reliability representation methods with each applied to a clustered and an unclustered data set. Subjects were divided randomly into six groups (3 color schemes × 2 map scales—see table 1 (over) for details of experimental design). Within each group, three subgroups worked with maps in a different order so that each of the three reliability methods was encountered on the first map by one third of each group. Subjects (equal numbers of women and men) were solicited from the university community through advertising posted on campus and in the campus newspaper. Subjects were paid $5.00 each for approximately 30–40 minutes participation. Brief background information questionnaires were administered that allow characterization of subject ages, occupations (majors for students), and experience levels with geography. Subject ages, although skewed toward that of college students, had a mean of 25 (figure 1). The subject group contained eight geography students (<10%) and these subjects were dispersed among test groups.

Plate 1. Legends (top row) are depicted for the 5-class test maps (matching the three reliability representations to the three color schemes). A color shift (top row left) is used to create a visually integral form of data–reliability display (hue shift in conjunction with the purple–green diverging scheme and saturation shift in conjunction with the spectral and yellow–red schemes). To create visually separable data and reliability depictions, a color sequence is used to represent data categories and a texture overlay is used to represent data reliability (top row center). For adjacent display, each data map was paired with a smaller two-category monochrome reliability map (at 50% scale, or one quarter of the area). Shown here is a legend from one data map and from one reliability map (top row right). The sample map shown (middle) depicts one of the clustered causes of mortality with the spectral color scheme to represent data categories and the visually separable texture overlay to represent those Health Service Areas (HSAs) in which data are categorized as unreliable. The map shown is a reproduction of a test map used in the experiment (subject to inevitable changes that result from color separation and printing). For the 7-class maps tested, symbolization schemes are comparable to those of 5-class maps. Shown here is a section from one of the unclustered test maps which uses the purple–green scheme to depict data and texture overlay to indicate HSAs for which death rates are considered to be unreliable. Again, the map shown is a section from a complete map used in the study.

[5] Color reproduction on the computer printer used is fairly close to colors that can be expected in the atlas (for which maps will be printed with an offset lithographic press). In both cases, color fills are created with a dithered pattern of dots made up of four process colors (cyan, magenta, yellow, and black). The most important difference between the computer printer used to produce maps for the experiment and the offset printing to be used in the atlas is that the computer-printer maps have a coarser dithering pattern—which makes thin lines difficult to reproduce.

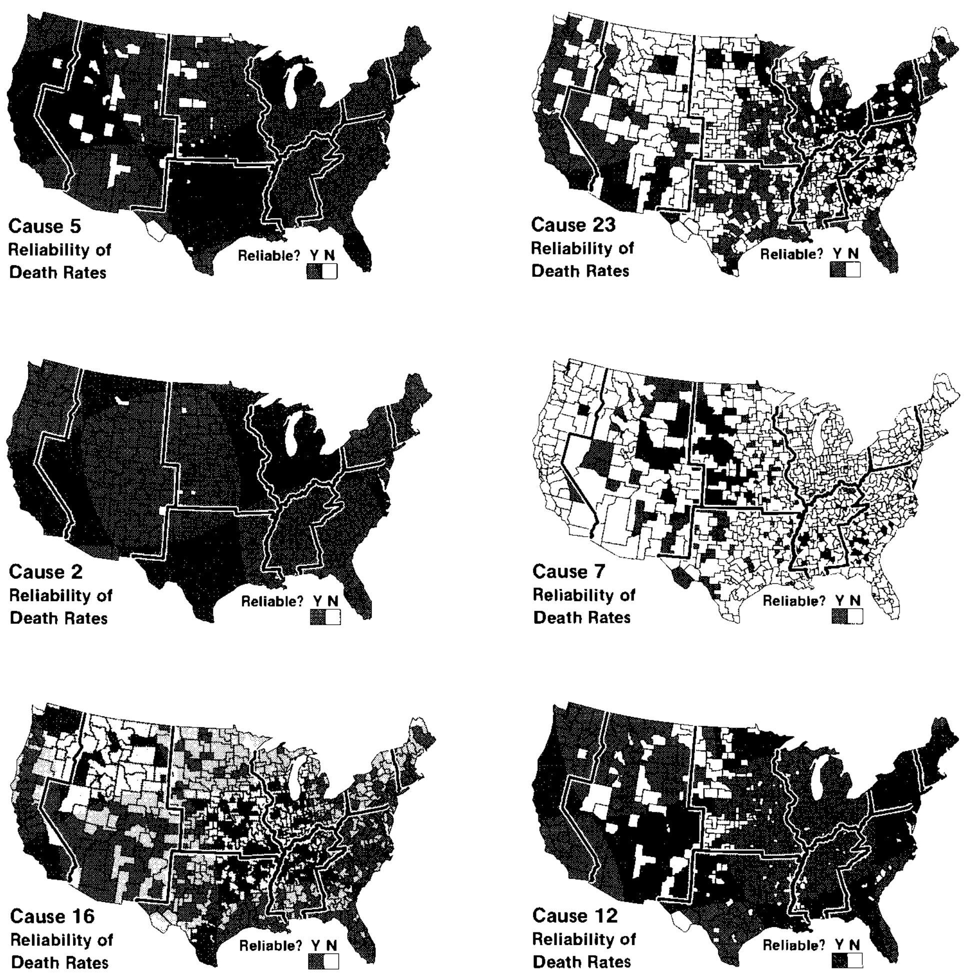


Figure 1. Binary maps representing reliability of Health Service Area (HSA) data for the six sample causes of death depicted on experimental maps. Dark gray represents those HSAs with a coefficient of variation less than 20% (reliable data) and light gray represents those with a coefficient of variation greater than or equal to 20% (unreliable data). These maps are printed at the size used to accompany the 5-class quarter-page data maps tested.

Map-use tasks

Data – reliability representation methods were assessed by comparing map interpretation performance on map-reading tasks typical of epidemiological analysis. Emphasis was on subject ability to make judgments about data under conditions of varying data reliability and on subject ability to assess visually spatial variations both in mortality statistics and in reliability of those statistics. In addition, preference and ease-of-use assessments were obtained. Tasks represented three levels of geographic detail (and of task complexity): (a) local—rate retrieval for individual map areas, (b) regional—judgments about, and comparisons among, regions within a map, and (c) national—overall map-pattern assessments.

Response accuracies and responses along rating scales were recorded in an experiment designed to suit within-subjects analysis of variance (ANOVA) of the continuous-scale response data. Chi-squared analysis was used for the categorical data. Between-subjects

Table 1. Experimental design—subjects each see maps at one scale and with one color scheme. Subjects are grouped to allow for analysis of variance: $3 \times 2 \times 2 \times 3$ factorial with repeated measures on two factors (reliability).

		C1	U1	C2	U2	C3	U3
Group 1							
pg/sF	G1	RC	RT	RP	RC	RT	RP
	G2	RT	RP	RC	RT	RP	RC
	G3	RP	RC	RT	RP	RC	RT
pg/sQ	G4	RC	RT	RP	RC	RT	RP
	G5	RT	RP	RC	RT	RP	RC
	G6	RP	RC	RT	RP	RC	RT
Group 2							
ry/sF	G7	RC	RT	RP	RC	RT	RP
	G8	RT	RP	RC	RT	RP	RC
	G9	RP	RC	RT	RP	RC	RT
ry/sQ	G10	RC	RT	RP	RC	RT	RP
	G11	RT	RP	RC	RT	RP	RC
	G12	RP	RC	RT	RP	RC	RT
Group 3							
sp/sF	G13	RC	RT	RP	RC	RT	RP
	G14	RT	RP	RC	RT	RP	RC
	G15	RP	RC	RT	RP	RC	RT
sp/sQ	G16	RC	RT	RP	RC	RT	RP
	G17	RT	RP	RC	RT	RP	RC
	G18	RP	RC	RT	RP	RC	RT

Definitions

Group	Three groups—group 1 has 36 subjects and groups 2 and 3 have 24 subjects each.
C-, U-	Level of clustering: two levels (U-unclustered, C-clustered) are tested within subjects.
-#	Variable mapped: three death-cause variables are nested within each level of clustering so subjects evaluate each scheme with a different mapped variable.
pg/, ry/, sp/	Color scheme: three levels (different schemes) are tested between subjects (pg is a diverging scheme—purple–green; ry is a sequential scheme—red–yellow; and sp is the spectral scheme).
/sF, /sQ	Scale and categorization of map: two levels (sF: full page—7-class; sQ: quarter page—5-class) are tested between subjects.
G	Subgroups: The sF subgroup of group 1 has 24 subjects; the remaining subgroups have 12 each. The four subjects in each subgroup (8 in pg/sF) evaluate the same color scheme (pg, ry, or sp), reliability representation (R), and map variable (V) combinations; the two subgroups are nested within each color scheme set (C).
R	Reliability scheme: three levels (different representation methods) are tested within subjects (RC—color change; RT—texture overlay; RP—map pairs).

ANOVA was used to compare results from this experiment (in which maps depicted reliability as well as data) and our prior experiment (where only data were depicted).

Testing procedure

Each test session was approximately 30–40 minutes in length. In the first 5 minutes, the purpose of the study was briefly explained, a background information questionnaire was completed, color vision was tested, and the general testing procedure was explained to subjects. Subjects then answered questions about six maps. At the end of the session, they were given the opportunity to ask questions and comment on the experiment and the maps.

Subjects examined each of six maps twice—as they answered two sets of questions for the maps. First, subjects reported the death rate and reliability legend categories of three HSAs on a map and chose (from a pair of regions) the region with the highest rate, the one with greatest clustering, and the one with highest overall reliability.[6] Subjects completed these six questions for each of the six maps. The second time they saw an individual map, subjects evaluated an individual region (census division) by rating the average death rate, clustering, number of clusters, and reliability on 7-step scales, with each scale defined by pairs of descriptive words or phrases (for example, 'low' or 'high' rate).[7] Subjects were then asked to determine which of the nine census divisions outlined on the map had the most distinct cluster of high values and which had the lowest proportion of reliable data. Next, subjects were asked to assess whole-map distributions with additional 7-step scales similar to those used for regions. Specifically, they rated the pattern of the entire map from 'clustered' to 'not clustered', from 'simple' to 'complex', and from 'reliable' to 'unreliable', and they rated the number of clusters on the map from 'none' to 'many'. Finally, subjects were asked to rate the reliability representation schemes along scales designed to elicit their impression of ease of use and map attractiveness.

The first set of questions for the maps were always completed before the second set of questions. Thus subjects saw the full range of clustering patterns and reliability on the maps before they were asked to scale clustering, pattern complexity, and map reliability and before they were asked to assess the effectiveness and attractiveness of the various reliability representation schemes. If they had been asked to perform all tasks with each map before working with subsequent maps, subjects would have had no basis from which to perform ratings of the entire maps and order effects might have overwhelmed any symbolization effects.

Results

Results presented here include comparison of performance with and preference for the three reliability representation schemes matched with the three color schemes for representing death rates. In addition, results for the current experiment are compared with those of our previous experiment in which identical color schemes were used to depict rates, but reliability was not represented. This second comparison makes it possible to assess the relative effectiveness of the three reliability representation methods as well as the impact of adding reliability information to maps. Results are organized by geographic scale of tasks completed by subjects.

Local-scale tasks

Subjects were very accurate at the local-rate retrieval task. For each test map examined, subjects made three death rate judgments for individual HSAs. In 93% of the cases all

[6] Regions used in this set of tasks were the nine census divisions of the conterminous United States. Region pairs for these questions were selected to provide a balanced number that are similar and different in their levels of clustering and average rates, as well as a balance of pairs that are geographically adjacent and nonadjacent. Regions used here were identical to those used in our previous experiment, thus allowing direct comparisons to be made.

[7] These scales, except for reliability, were identical to ones used in our previous experiment (Brewer et al, 1997). In that study, pilot testing was used to refine the specific terms used on each scale. Between-subject differences in interpretation of the rating scales (even if they exist) are not relevant to our comparison of reliability representation methods, because this comparison is based on within-subject analysis.

three judgments were made correctly. No significant differences were found on this task among the adjacent, coincident separable, and coincident integral methods of representing reliability. The only significant difference found was for level of clustering. Clustered maps resulted in a significantly different pattern of performance—a low percentage of maps for which one error was made and a high percentage of maps for which zero or greater than one error was made. Unclustered maps resulted in a high percentage of single errors. Accuracy with which rate retrievals could be performed was not significantly different for maps with or without reliability depicted—even when maps (and HSAs) are rather small (as they are for the quarter-page maps).

Judgments of HSA reliability were somewhat less accurate (82%) than rate retrieval judgments but, again, no significant difference was found. There was a notable interaction between reliability representation and map scale. Performance was significantly less good when reliability was represented on quarter-sized maps—almost entirely because of poor performance of map pairs (with performance dropping to 64% of the maps having all three judgments correct). This drop in performance is clearly the result of the small size of some HSAs on the eighth-page reliability maps that are matched with the quarter-page rate maps.

Regional-scale tasks

At the regional level, three categories of task were posed: (1) estimating characteristics of individual regions delineated on each map, (2) comparing a pair of regions, and (3) selecting map regions. For the last task, regions were selected to meet each of two criteria; the region with "most distinct cluster of high rates" (that is, the most distinct 'hot spot') and the one with the "lowest proportion of reliable data".

Subjects saw significantly fewer within-region clusters when reliability was depicted with integral color than with separable or adjacent depiction [$F = 3.43$; degrees of freedom (df) $= 2$; $p = 0.035$]. Map pairs resulted in significantly more correct comparisons of region reliability ($\chi^2 = 8.94$; df $= 4$; $p = 0.011$), but also significantly higher ratings of reliability for individual regions ($F = 11.41$; df $= 2$; $p = 0.000$). Region reliability comparisons were also considerably more accurate when the map variable was clustered than when it was not, as were region death rate comparisons. No significant difference resulted for the task of selecting a region with a cluster of high rates, but there was a significant difference for the task of identifying the region with least reliable data ($\chi^2 = 12.74$; df $= 4$; $p = 0.013$). Accuracy on this task was highest with adjacent maps and lowest with integral colors.

Overall, adding reliability representation had no impact on estimates of mean death rate by region or on any region comparisons. Fewer clusters were seen on maps that included reliability depiction than on maps that did not ($F = 2.76$; df $= 3$; $p = 0.041$) (with most of the difference attributable to those maps using integral colors). Performance on choosing which region is more clustered is poor (just above chance levels) whether or not reliability information is included. Although this poor performance may indicate that map readers (at least untrained ones) are not able to judge regional clusters accurately, the poor performance might also result from the kind of task posed (comparison of clustering in arbitrary regions, the bounds of which may divide clusters into parts) or the way clustering was measured (Moran's I-statistic calculated independently for each region, even though the regions have different numbers and sizes of enumeration units).

National-scale tasks

For the whole map, subjects were asked to rate, on 7-point scales, the degree of clustering, number of clusters, complexity, and reliability as well as to rate subjectively the pleasantness and ease of use of each representation scheme. In table 2 (see over) we

Table 2. ANOVA (analysis of variance) results for rating scale responses.

	Reliability representation[a] F	Reliability × subject (color)[b] p	Rate × representation (color)[a,c] F	Rate × subject (color)[b] p
Degrees of freedom	2		2	
Clustering	1.53	0.220	0.03	0.970
Number of clusters	4.90	**0.009**	0.01	**0.994**
Complexity	2.08	0.128	0.83	0.438
Reliability	8.04	**0.001**	3.45	**0.036**
Pleasant	13.09	**0.000**	1.91	0.155
Easy to read	21.52	**0.000**	0.44	0.644

Note: significance at $\alpha < 0.05$ is shown in bold.
[a] Hypothesis effects which are the numerator mean squares in F-statistic calculations.
[b] Error effects which are the denominator mean square for F (these error mean squares were also used in calculation of the Tukey studentized range test (HSD) means comparisons).
[c] Subjects are nested within rate representation (color scheme).

present ANOVA results for within subjects comparison by reliability and rate representation methods. Maps using integral colors were judged to have significantly fewer clusters than maps using the alternative reliability representations. When judgments of reliability were made, adjacent map pairs were judged to be significantly more reliable (in their depiction of rates) than maps using separable (but spatially coincident) texture overlay, perhaps because the reliability information is harder to ignore when it is on the same map as the rates. Map clustering is a significant factor for all four tasks, with more clustered maps resulting in higher ratings for clustering, cluster number, and reliability and lower ratings for complexity.

Between subjects ANOVA was used to compare rating scale responses of subjects in the current experiment (for whom maps depicted reliability) to those in our previous experiment (Brewer et al, 1997) in which maps depicted only death rates (table 3). Estimates of the level of clustering and number of clusters on test maps are not significantly different for maps with reliability representation versus maps showing only death rates. There was a significant difference in complexity judgments due to higher ratings with integral color display than with maps having no reliability representation, map pairs, or maps using separable texture overlay. Since integral methods result in the visual appearance of unique symbols for each combination of variables (rate × reliability), the map becomes visually more complex than when the second variable is depicted

Table 3. ANOVA (analysis of variance) results for rating scale responses—comparison of experiment 1 (without reliability representation) and experiment 2 (with three different forms of reliability representation).

	Reliability representation		Rate representation	
	F	p	F	p
Degrees of freedom	3		2	
Clustering	1.34	0.260	0.14	0.872
Number of clusters	2.22	0.085	0.58	0.558
Complexity	3.43	**0.017**	4.02	**0.018**
Pleasant	19	**0.000**	9.14	**0.000**
Easy to read	40.54	**0.000**	5.09	**0.006**

Note: All analyses are between subjects. Significance at $\alpha < 0.05$ is shown in bold.

on a different map, on the same map using separable symbolization, or not depicted at all.

Method of reliability representation makes a substantial difference in judgment of pleasantness and ease of reading (table 2). Map pairs and visually separable texture overlay were judged to be significantly more pleasant and easier to use than visually integral color merger. For both ratings, Tukey's HSD found no difference between map pairs and separable display with both means different from integral display.

Ratings of pleasantness were significantly lower for maps that depict reliability than for those that do not (table 3). As with many other findings, however, much of this negative reaction could be attributed to maps using visually integral colors—judged to be significantly (and substantially) less pleasant than any of the other three methods. Not surprisingly, maps without reliability representation were rated as significantly easier to read than those with reliability representation (table 3). Among the reliability representation methods, there was no significant difference in ease-of-reading ratings between map pairs and visually separable texture overlay, and both were rated as significantly easier to read than maps with visually integral colors.

Discussion

Results demonstrate that reliability of georeferenced information can be successfully represented on maps without impeding map readers' ability to estimate values or identify clusters. A novel method of coincident visually separable display using texture overlay is particularly successful in allowing both death rates and reliability to be interpreted independently. This method of display facilitated map interpretation in a variety of ways (in comparison with a coincident visually integral method that relies on attributes of color to encode both death rates and reliability, and to adjacent displays of rates and reliability). Unlike integral colors, but like adjacent display, a separable texture overlay of reliability information can be disregarded when necessary in order to notice potentially important data clusters. Texture overlay, however, is more effective than an adjacent display in prompting map readers to recognize those maps, or map sections, in which data should be evaluated with caution.

For the exploratory uses that are the most important application of a mortality atlas, the preferred symbolization scheme is one that prompts map users to notice reliability (or lack of it), but allows them to look past this information to notice clusters that might prove to be important—thus visually separable representation is ideal. If presentation uses of an atlas were paramount, however, then visually integral representation might be preferred. The visually integral color scheme tested here would have the general result of preventing readers of a presentation-oriented atlas from seeing any clusters that contained a high proportion of enumeration units for which data were not reliable. The only clusters that would be visible would be ones for which data were judged to be reliable.

The research reported is among the first attempts to test methods for depicting reliability on maps (see Evans, 1997, for another). The findings of this study are encouraging—because they demonstrate that map readers (even ones who may not be as sophisticated as typical users of a mortality atlas) can cope with the added visual burden of reliability depiction on already complex 7-class maps. Although the research produced answers to key questions concerning the design of a particular atlas (one reason for the study), it also contributes toward a more general understanding of how to represent reliability in a variety of mapping (and other information display contexts). There are many research questions about georeferenced reliability representation that remain to be addressed. They include issues of symbol design (for example, assessing different texture overlay parameters and determining whether there are forms of color change that are more successful at depicting data and data reliability than those used in this study), extension of representational concepts to electronic maps (for example, how does the problem of representing reliability change in an interactive computer environment?), and understanding the implications for decisionmaking of providing georeferenced reliability information.

As increasing volumes of health data are being mapped and geographical information systems are making it possible to explore more creatively links between health and environment, it is critical that the issue of data reliability and its representation be given attention—both by those building georeferenced databases and visualization tools and by those who use them in research and policy applications. As a result of this research, data reliability depictions have been incorporated in the *Atlas of United States Mortality* recently published by the US Centers for Disease Control, National Center for Health Statistics (Pickle et al, 1997). An important next step is to investigate the robustness of the visually separable texture overlay method developed here in the context of dynamic electronic maps on CD-ROM and the World Wide Web.

Acknowledgements. The research reported here was supported by a contract from the US National Center for Health Statistics (DHHS, OASH, DAM No.9430589). Additional examples of maps used in the experiment can be accessed through the World Wide Web at: http:www.gis.psu.edu/MacEachren/MacEachrenHTML/ep/e&pweb.html. Information on the *Atlas of United States Mortality* to which the research contributed can be found at: www.cdc.gov/nchswww/nchshome.htm.

References

Berlin B, Kay P, 1969 *Basic Color Terms: Their Universality and Evolution* (University of California Press, Berkeley, CA)

Bertin J, 1981 *Graphics and Graphic Information Processing* (Walter de Gruyter, Berlin)

Bertin J, 1983 *Semiology of Graphics: Diagrams, Networks, Maps* (University of Wisconsin Press, Madison, WI)

Blot W J, Fraumeni J F J, 1976, "Geographic patterns of lung cancer: industrial correlations" *American Journal of Epidemiology* **103** 539 – 550

Blot W J, Harrington J M, Toledo A, Hoover R, Heath C W, Fraumeni J F J, 1978, "Lung cancer after employment in shipyards during World War II" *New England Journal of Medicine* **299** 620 – 624

Boynton R M, Fargo L, Olson C X, Smallman H S, 1989, "Category effects in color memory" *Color Research and Application* **14**(5) 229 – 234

Brewer C A, 1994, "Color use guidelines for mapping and visualization", in *Visualization in Modern Cartography* Eds A M MacEachren, D R F Taylor (Pergamon, Oxford) pp 123 – 147

Brewer C A, MacEachren A M, Pickle L W, Herrmann D J, 1997, "Mapping mortality: selecting hues and evaluating color schemes" *Annals of the Association of American Geographers* **87** 411 – 438

Carr D B, Olsen A R, White D, 1992, "Hexagon mosaic maps for display of univariate and bivariate geographical data" *Cartography and Geographic Information Systems* **19**(4) 228 – 236

Carswell C M, Wickens C D, 1990, "The perceptual interaction of graphical attributes: configurality, stimulus homogeneity, and object integration" *Perception and Psychophysics* **47** 157 – 168

Croner C M, Pickle L W, Wolf D R, White A A, 1992, "A GIS approach to hypothesis generation in epidemiology", in *ASPRS/ACSM/RT '92* American Society for Photogrammetry and Remote Sensing, Falls Church, VA; American Congress on Surveying and Mapping, Bethesda, MD, pp 275 – 283

Diehr P, 1984, "Small area statistics: large statistical problems" *American Journal of Public Health* **74** 313 – 314

Dorling D, 1996, "Area cartograms: their use and creation", School of Environmental Sciences, University of East Anglia, Norwich, Norfolk

Evans B, 1997, "Cartographic representation of reliability, does it benefit the user?" *Computers and Geoscience* **23** 409 – 422

Goodchild M, Chih-Chang L, Leung Y, 1994, "Visualizing fuzzy maps", in *Visualization in Geographical Information Systems* Eds H Hearnshaw, D Unwin (John Wiley, Chichester, Sussex) pp 158 – 167

Granger G W, 1955, "An experimental study of colour preferences" *Journal of General Psychology* **52** 3 – 20

Jones K, Kirby A, 1980, "The use of chi-square maps in the analysis of census data" *Geoforum* **11** 409 – 417

Lewandowsky S, Behrens J T, Pickle L W, Herrmann D J, White A, 1995, "Perception of clusters in mortality maps: representing magnitude and statistical reliability", in *Cognitive Aspects of Statistical Mapping* Eds L W Pickle, D J Herrmann (Centers for Disease Control, National Center for Health Statistics, Washington, DC) pp 107 – 132

MacEachren A M, 1992, "Visualizing uncertain information" *Cartographic Perspectives* **13** 10 – 19

MacEachren A M, 1995 *How Maps Work: Representation, Visualization and Design* (Guilford Press, New York)

MacEachren A M, Ganter J H, 1990, "A pattern identification approach to cartographic visualization" *Cartographica* **27**(2) 64 – 81

MacEachren A M, Howard D, von Wyss M, Askov D, Taormino T, 1993, "Visualizing the health of Chesapeake Bay: an uncertain endeavor", in *Proceedings, GIS/LIS '93* American Society for Photogrammetry and Remote Sensing, Falls Church, VA; American Congress on Surveying and Mapping, Bethesda, MD, pp 449 – 458

McGranaghan M, 1993, "A cartographic view of spatial data quality" *Cartographica* **30**(2/3) 8 – 19

Makuc D, Haglund B, Ingram D D, Kleinman J C, Feldman J J, 1991 *Health Service Areas for the United States* (Vital and Health Statistics, National Center for Health Statistics, Washington, DC)

Mason T J, McKay F W, Hoover R, Blot W J, Fraumeni J F J, 1975 *Atlas of Cancer Mortality for US Counties: 1950 – 1969* (US Government Printing Office, Washington, DC)

Olson J, Brewer C A, 1997, "An evaluation of color selections to accommodate map users with color-vision impairments" *Annals of the Association of American Geographers* **87** 103 – 134

Pickle L W, Herrmann D J (Eds), 1995 *Cognitive Aspects of Statistical Mapping* WP18, Cognitive Methods Staff, Office of Research and Methodology, Centers for Disease Control and Prevention and National Center for Health Statistics, Washington, DC

Pickle L W, White A, 1995, "Effects of the choice of age-adjustment method on maps of death rates" *Statistics in Medicine* **14** 615 – 627

Pickle L W, Mason T J, Howard N, Hoover R, Fraumeni J F J, 1987 *Atlas of US Cancer Mortality Among Whites: 1950 – 1980* (US Government Printing Office, Washington, DC)

Pickle L W, Mason T J, Howard N, Hoover R, Fraumeni J F J, 1990 *Atlas of US Cancer Mortality Among Nonwhites: 1950 – 1980* (US Government Printing Office, Washington, DC)

Pickle L W, Mungiole M, Jones G K, White A A, 1997 *Atlas of United States Mortality* (Public Health Service, Hyattsville, MD)

Pomerantz J R, 1985, "Perceptual organization in information processing", in *Issues in Cognitive Modelling* Eds A M Aitkenhead, J M Slack (Lawrence Erlbaum Associates, London) pp 125 – 158

Regan B C, Reffin J P, Mollon J D, 1994, "Luminance noise and the rapid determination of discrimination ellipses in colour deficiency" *Vision Research* **34** 1279 – 1299

Travis D, 1991 *Effective Color Displays: Theory and Practice* (Academic Press, San Diego, CA)

van der Wel F J M, Hootsman R M, Ormeling F, 1994, "Visualization of data quality", in *Visualization in Modern Cartography* Eds A M MacEachren, D R F Taylor (Pergamon, Oxford) pp 313 – 332

Whitfield T W A, Wiltshire T J, 1990, "Color psychology: a critical review" *Genetic, Social, and General Psychology Monographs* **116**(4) 387 – 411

Winn D M, Blot W J, Shy C M, Pickle L W, Roledo A, Fraumeni J F J, 1981, "Snuff dipping and oral cancer among women in the southern United States" *New England Journal of Medicine* **304** 745 – 749