



STATISTICAL PRIMER

State Center for Health Statistics • 1908 Mail Service Center • Raleigh, NC 27699-1908 • 919/733-4728
www.schs.state.nc.us/SCHS/

No. 15

May 2002

Mapping Mortality and Morbidity Rates

by

James L. Wilson
Paul A. Buescher

Introduction

Maps are powerful tools used to analyze and compare a wide range of geographically referenced data. The last third of the twentieth century has witnessed the increasing use of maps to portray statistical information. Maps of mortality and morbidity are useful for assessing relationships with potential risk factors. They can also be used to track progress in public health programs and policy development.

A map is a model of the world. Like any model, a certain amount of information is omitted to make the map comprehensible to the viewer. Because of this simplification, it is necessary for a map to include tools such as symbols and keys to assist the map-reader in interpretation. The methods used in map construction and the tools used to interpret them are part of the repertoire of modern *geographic information systems (GIS)*. GIS stores, retrieves, and integrates disparate geographical data to create map *layers*. These layers portray information about a wide range of topics, using a set of conventional symbols. Layers form the basis of a *thematic* map.

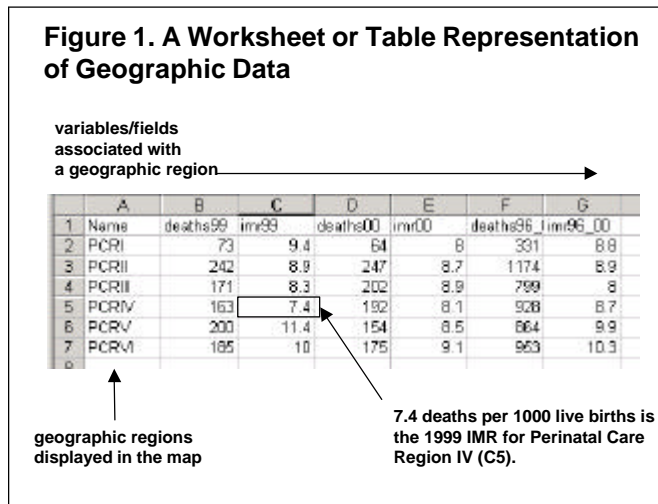
The focus of this *Statistical Primer* is the interpretation and analysis of a specific type of thematic map that portrays mortality and morbidity rates, which is produced at the State Center for Health Statistics (SCHS) in North Carolina. To this end, the *Primer* will introduce basic concepts such as geographic units of analysis, data classification, and some uses of symbols. Map interpretation basics will also be introduced. The maps in this Primer were created using ArcGIS, ESRI Inc., 2001.

Choropleth Maps

Most of the maps developed by the SCHS are based on geographic areas, which are usually represented as *polygons*. They can be political and administrative entities like counties, census tracts, cities, or minor civil divisions. Areas can also represent biological and environmental phenomena like drainage basins, soil types, vegetation zones, climate areas, and geologic structure. Each county, for example, can have thousands of variables assigned to it that describe a diverse set of characteristics such as population structure, economic data, housing characteristics, or birth rates. These variables, or attributes, can then be arranged into a



spreadsheet or table, with geographic areas as rows and variables as columns (Figure 1). Through GIS, the attribute data can be linked to the geographic data used to draw the object (e.g., point, line, or polygon) using a field that is common to both datasets.



Counties with similar characteristics can be built into larger spatial units. Areas of the same level of geography, such as the census tract, can be aggregated or joined together to form a county. Counties can then be joined together to create, for example, State Economic Areas (SEAs) or Health Service Areas (HSAs). The process of building larger areas from smaller ones based on similar characteristics is known as *regionalization*. Similar characteristics can include market areas, economic conditions, population structure, or mortality rates. Regionalization is also useful for combining areas that have small populations. Small populations typically generate small numbers of events such as births and deaths, and therefore unstable vital rates. This problem is especially acute for mortality due to rare diseases. Regionalization involves aggregating the *numerators* (events) and *denominators* (populations) used for calculating rates. (For a discussion of the effects of small numbers on the calculation of rates see *Statistical Primer 12: Problems with Rates Based on Small Numbers* at www.schs.state.nc.us/SCHS/pubs/.) Once a region is formed, the remnant internal boundaries can be “dissolved,” yielding the new spatial entity. **Figure 2** illustrates this process.

Using geographic areas to show levels of numeric data is known as *choropleth mapping*, a way of portraying data for areas. The SCHS employs choropleth maps in many of its publications to depict rates of morbidity and mortality. The publication *Leading Causes of Death, Volume II* and the maps from the *North Carolina Health Atlas* (www.schs.state.nc.us/SCHS/) are examples of this type of mapping. Many national atlases that present health statistics use choropleth maps. The basic “building blocks” for these maps are usually counties, but other units are used as well, such as ZIP codes, HSAs, and SEAs. For recent examples of region building, see the *Atlas of Cancer Mortality in the United States 1950-94*,¹ *Atlas of United States Mortality*,² and *The Dartmouth Atlas of Health Care*.³

There is a long tradition of using the county as the basic geographic unit for many choropleth maps. An important reason for this is that a wide variety of data have been collected at this level of geography. The county is an administrative and political unit but it is also comprised of smaller geographic units such as census tracts, block groups, blocks, and minor civil divisions. County boundaries are fairly stable over time, while areas such as ZIP codes and congressional legislative districts do not necessarily follow county boundaries and can change over time because they are dependent on population dynamics.

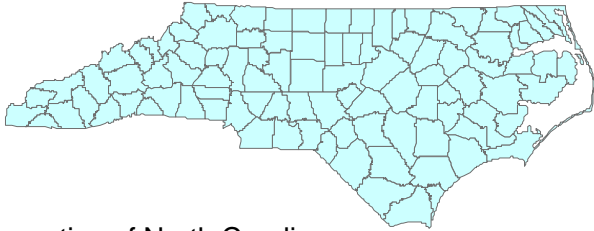
The county is also central to many aspects of social, economic, and political life. Most of the 100 counties in North Carolina have their own health departments, cooperative extension services, chambers of commerce, and many school districts have been consolidated to the county level. Although counties are defined formally as political and administrative units, they are also functional units because people live, work, and identify with them.

Mapping Rates and Ratios

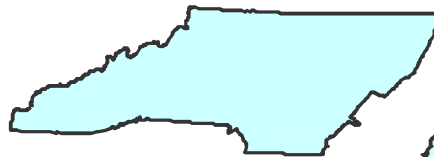
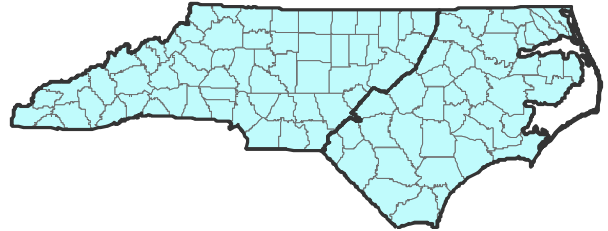
Map statistics at the SCHS are based on two underlying components: numerators and denominators. Numerator data include the counts of vital events such as deaths and births. Denominator data are often the populations for geographic areas (e.g., states, counties, zip codes, census tracts, block groups). The number of vital events that occur within a county for a given

Figure 2. Examples of Regions

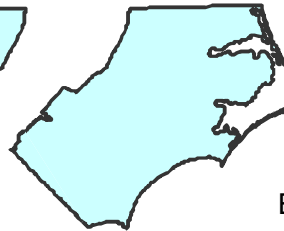
A. Creating Two Separate Regions



The counties of North Carolina



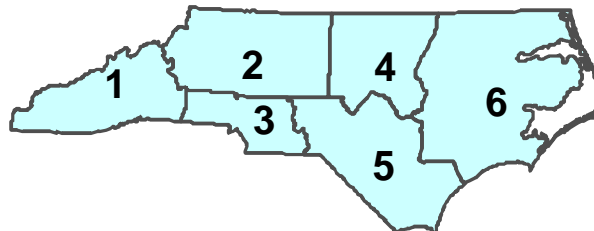
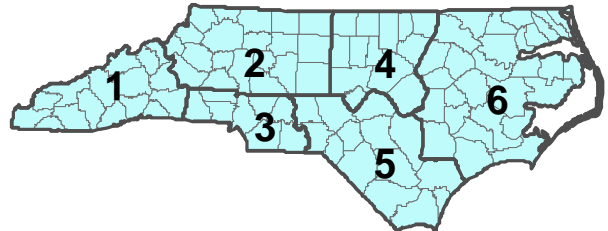
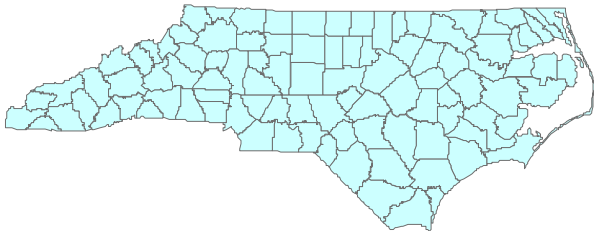
The rest of North Carolina



Eastern North Carolina



B. Creating Perinatal Care Regions



In creating regions, the numerator and denominator data are aggregated (summed) to the new level of geography. Regionalizing is useful in situations where base populations are small and few events occur. Sudden Infant Death Syndrome (SIDS) is a good example of events that occur each year in relatively low and fluctuating numbers, so it would be more appropriate to map the SIDS death rates to Perinatal Care Regions rather than counties, thereby increasing the stability of the data.

period of time is a function of the underlying size of the population, although other factors such as the age structure of the population are important.

Choropleth maps are commonly used to portray rates and ratios. A rate is the number of events in an area for a specified time period divided by the population of the area, with the result often multiplied by a constant (such as 1,000). Rates differ from ratios, which are comparisons between two different quantities or two different populations. For example, the Infant Mortality Rate (IMR) is actually a ratio since two different quantities, infant deaths (d) and births (b), are being compared for the events occurring during a given time period: $IMR = (d \div b) \times 1000$. Unlike a ratio, a rate's denominator (or population at risk) also includes its numerator: $Rate = (D \div (D+S)) \times 1000$. For example, D might be the total number of deaths in a population in a year, and D+S is the total population for the year (those who died in the year plus those who survived).

A *crude rate* is calculated by dividing the number of events (the numerator) by the population (the denominator) for a specified period of time. A major advantage to mapping crude rates is that it assists the reader in identifying where relatively large numbers of negative health events are occurring. A drawback to mapping crude rates is that because the county populations differ by age or other factors, the rates may not be very useful for comparing underlying health risks among areas. However, when used in conjunction with a map of *adjusted* rates, information about age-related versus other factors affecting mortality or morbidity is revealed more clearly. (See *Statistical Primer 13: Age Adjusted Rates* for details, at www.schs.state.nc.us/SCHS/pubs/.)

Developing Map Categories

As previously mentioned, maps are models, and modeling usually involves some type of simplification process. Not everything about the world can be represented on a map, and in statistical mapping usually one or two topics or themes are shown. There are 100 counties in North Carolina; to show the numerical value of the rate for each county on the map would make the discrimination of spatial patterns nearly impossible. To assist in pattern recognition, numerical values such as rates are grouped using a classification

method. There are several classification methods that are commonly used. This *Primer* will focus on the *natural break*, *equal interval*, *quantile (equal count)*, and *statistical (using standard deviations)* methods. Each method has certain advantages and disadvantages, which will depend on the distribution of the data and the intent of the mapmaker.

Data such as mortality or morbidity rates can be arranged as a series of high to low values. Generally, the observed distribution is somewhat less than continuous; what emerges are several “clumps,” groups or clusters of similar rates. The breaks (or data cut-points) between any two groups within the distribution are considered natural, since the observer has imposed no categorization scheme other than the number of breaks desired. The mapmaker determines the number of *natural break* categories, either using a predetermined difference or tolerance between any two groups, or letting the GIS software determine mathematically the boundaries among the groups. The use of natural breaks will usually result in non-uniform data intervals and numbers of observations per category.

Within the data distribution there may be *outliers*, values that are extreme and can be found at either or both ends of the distribution and that do not fit well in the categorization scheme. Rather than exclude these data, as might be done in some statistical analyses, they are typically included into a special category for mapping. Outliers are a problem for all classification schemes, but most of a distribution can be accommodated well with commonly used classification techniques.

For relatively uniform and rectangular distributions (as opposed to curved or bell-shaped distributions), *equal interval* or *quantile (equal count)* categories may be appropriate. For classifying rates by equal intervals, the distributions of rates or values are partitioned into two or more categories, with each category having the same data range or interval. For example, the minimum and maximum rates for a distribution are 12.0 and 91.0. The range of data values is therefore 79.0. If four categories are chosen, each category will have an approximate range of 20.0 ($79.0 \div 4$). The boundaries for each equal-interval category would then be 12.0-31.9, 32.0-51.9, 52.0-71.9, and 72.0-91.9. Although the category intervals are uniform, the distributions of data

within them may not be. There are often fewer observations in the extreme categories.

To create categories based on the quantile method, the ordered observations are grouped into categories possessing approximately the same number of observations. In North Carolina there are 100 counties, so if five map categories are required, each category will have twenty counties. Odd numbers of observations or ties in their values may result in categories that have different numbers of observations. Both the equal interval and quantile methods ignore the natural distribution of the data. However, their chief advantage is that they can be used to show changes over time when the categories are kept constant.

The *statistical* methods of classification found in most geographic information systems are based on the normal distribution of data. A normal distribution of data has a bell shape (arched), with the average or mean value of the data at the center of the bell (the highest point of the arch). Most observed data distributions do not conform neatly to that classical normal shape, but are assumed to do so since the observations are a finite number and can be regarded as a sample of a theoretical normal distribution. The normal distribution is often *standardized*, which is a technique used to “calibrate” or “re-scale” the distribution so that the mean is set to zero (0.0). One standard deviation is the unit distance from the standard mean (0.0). Plus or minus one standard deviation is the distance from the mean that contains about 68% of all the observations. Plus or minus two standard deviations contain about 95% of the observations. Extreme observational values are found at the “tails” of the distribution and can sometimes be as far as three and four standard deviations away from the mean or central value. Map categories are created using the values of standard deviations as the category boundaries.

A common problem in using the statistical method of classification is that not all phenomena being mapped possess a bell-shaped, normal distribution. Outliers can potentially *skew* the mean of the population distribution. The consequence of creating map categories from standard deviations is that while the bulk of the observations will fall within one and two standard deviations, there may be one extreme category with a few observations.

The purpose of map classification is to render the data into a simpler form for analysis and to delineate meaningful patterns. The trade-off is the loss of detailed information about the data, but the gain is in information about the spatial relationships that may exist.

Once the data are categorized, the next step is to select the appropriate symbols.

Map Symbols

Maps incorporate symbols to represent real-world phenomena. Symbols such as different colors, shading, line, and dot patterns can be used. Map symbols often have a logical connection to the phenomena they represent. For example, blue shades typically symbolize water, green and brown represent different categories of elevation or altitude, and blacks and reds almost always represent transportation networks or borders. Basic geometric objects such as the point, line, and polygon symbolize a wide range of both real and abstract phenomena. Points can represent public health events such as tuberculosis cases or the locations of towns. Lines can represent connections among points of interest like the social network of HIV cases or the transportation network that exists among places. Polygons are used to symbolize political entities such as counties or census tracts.

Patterns and shading can be used to symbolize quantities in choropleth mapping. Patterns can be made to represent varying quantities by adjusting the density and sizes of lines or dots (pixels). Increasing pattern density and thickness usually represents increasing quantities or rates. Color can also be used to represent quantities when the intensity is modified. Each color shade can be associated with a map category.

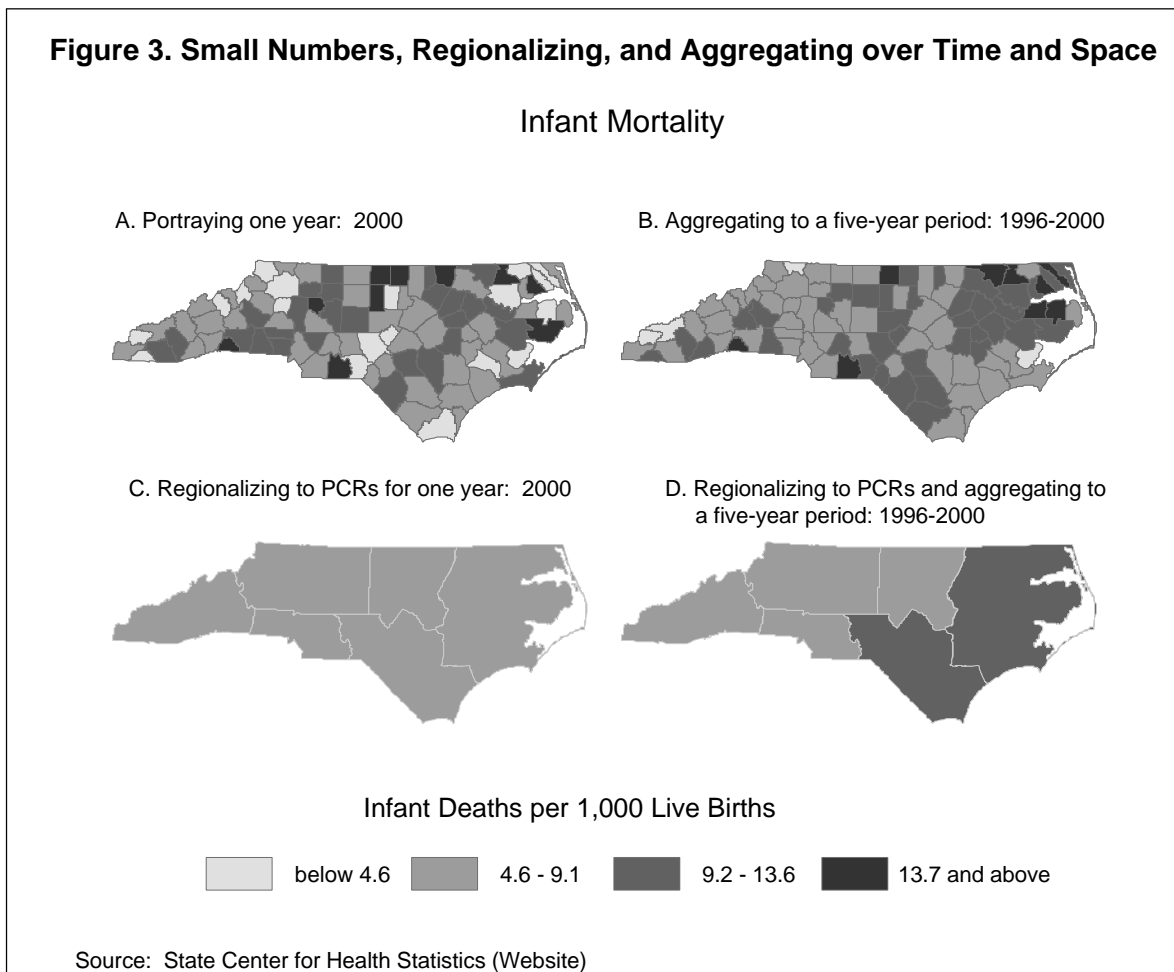
In general, only one color (or hue) should be used to represent quantities in a single-variable choropleth map. Different colors should not be used to represent different categories on a single-variable map. A color can be made to represent varying quantities by making it lighter or darker. However, in certain circumstances such as two-variable mapping, two colors may be appropriate. The use of more than one color is generally reserved for maps that have qualitative differences

or dimensions in the phenomena such as vegetation or soils, or when they possess more than one quantitative attribute.

Interpreting and Analyzing Choropleth Mortality Maps

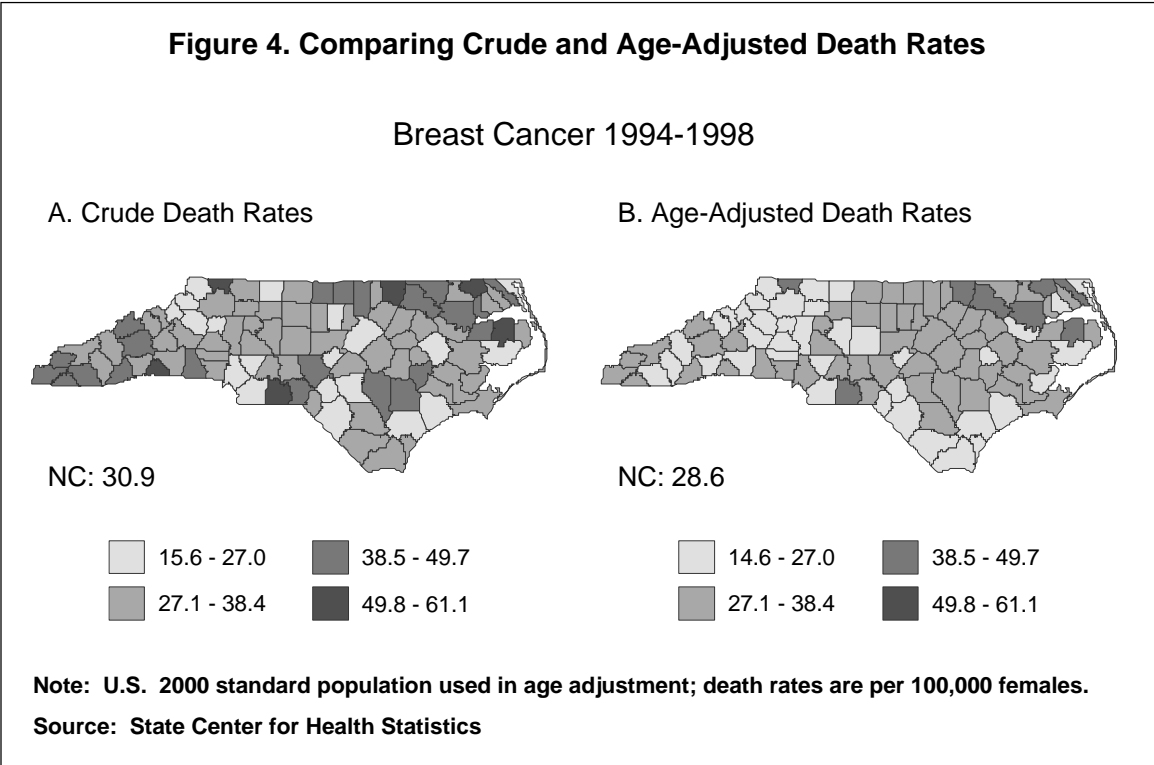
Most of the maps produced at the SCHS are choropleth maps using counties, or some region based on counties such as Perinatal Care Regions (PCRs), as the geographic units of analysis. For these types of maps, different categories are symbolized by varying shades of a particular hue, with darker shades representing higher rates and lighter shades representing lower rates. The following examples help illustrate some of the concepts introduced earlier by using a series of maps of real-world data.

Figure 3 shows how map patterns can be affected by the aggregation of data in both time and space. Infant mortality occurs in relatively low numbers compared to some other adverse birth outcomes such as low birth weight. Therefore, problems often arise when comparing infant mortality rates from one year to the next at the county level because the change from one year to the next can vary dramatically. For example, in one year in a county there were 260 births with three infants dying in that year. These numbers produce an Infant Mortality Rate (IMR) of 11.5 $((3 \div 260) \times 1,000)$. The next year, the same county had two infant deaths and 260 births, resulting in an IMR of 7.7. This represents a 33% drop in the IMR, but there is a question whether this change was the result of the inherent random fluctuation of small numbers of infant deaths or a true change in the underlying IMR.



There are several methods available that can be used to stabilize randomly fluctuating rates and their resulting map patterns. Two simple methods are: 1) aggregate the events (infant deaths) and the population generating them (births) across several years, and/or 2) aggregate with respect to geography; for example, aggregate individual counties into Perinatal Care Regions (PCRs). In Figure 3, Map A shows county IMRs for just the year 2000. Map B shows the county map pattern when the 1996 through 2000 infant deaths and births are aggregated to produce the IMRs. Comparing these two maps shows that a geographic pattern is more easily discernible through temporal aggregation. Map C shows geographic aggregation into PCRs for the year 2000, which reveals no pattern if the same map categories are used. Map D shows the effects of aggregating in both time and space. The pattern in Map D is that the south-central and eastern portions of the state exhibit rates substantially higher than the overall five-year state IMR of 9.1. However, the contrast between Map C and Map D could also reflect a reduction over the five-year period of the differences in IMRs among the PCRs. The general effect of aggregation, in both time and space, is loss of some of the “noise” in the data, which provides more meaningful patterns.

Figure 4 portrays the difference in patterns between 1994-1998 crude and age-adjusted death rates for breast cancer. The classification in the legend, based on the equal interval method, is kept constant except for the lowest values. Age adjustment produces the rates that would be expected *if* all counties had the same age distribution as a standard population. At the SCHS, the standard population used is the U.S. 2000 population. Age adjustment using an external standard permits comparisons of rates among all counties, showing the effects on mortality of factors other than the age distribution. If the same standard population is used to adjust rates for different time periods, then temporal change in the rates (due to factors other than age) can be measured. The pattern in Map B demonstrates how age adjustment has the effect of moderating the extreme values of the county distribution of rates, producing fewer “hills and valleys” in rates across the map. In fact, in Map B there are no counties in the highest category. Map B also shows that the concentration of higher rates tends to be in the eastern and non-metropolitan counties of the state, when age distribution is held constant.



A map of crude rates draws attention to areas where problems are occurring in terms of magnitude and can assist in decision-making about allocation of health care resources. Age-adjusted maps are useful for making comparisons across time and space, and for showing the impact of factors other than the age distribution on mortality.

Figure 5 portrays four methods commonly used for classifying mortality rates. Maps A through D show the effects the different classification methods have on the spatial pattern of rates. The rates themselves are not altered, but the category values or “cut points” defining the intervals are, in contrast to the method used in Figure 4. This results in counties changing categories. Each classification method emphasizes a particular aspect of the underlying data distribution.

Map A, using the *natural break* method, shows in a simple way how counties with similar values are arranged in the map. Counties with similar values may also be neighbors, thereby creating natural regions of rates. In the State Center for Health Statistics publication *Leading Causes of Death*, more sophisticated statistical clustering methods are used for classifying rates, but the underlying idea of grouping like values is the same. The chief drawback of the natural break or clustering method is that it is more difficult to compare maps from two different time periods, since the values that delineate the class breaks will generally change.

With the *equal interval* classification method, the numbers of observations are usually fewer in the extreme classes since most data values tend to lie in the middle of the distribution. This can be seen clearly in Map B, where the middle categories have the greatest numbers of counties and the extreme categories have the fewest. This method is useful for monitoring changes in rates in the lowest and highest categories over time when the class breaks are kept constant. Map B shows that most of the extreme highest county rates are found in the northeastern part of the state, while the extreme lowest rates are found in the northwest.

Quantile classification yields map classes that have approximately equal numbers of observations (Map C). Because North Carolina has 100 counties, each of the four categories will have about 25 counties. In contrast to the equal interval method, the quantile method is useful for tracking shifts in membership in the middle categories. The quantile method is a simple way to begin a series of maps whose observations are being tracked over time without making any *a priori* assumptions about the underlying distribution of the data. In the example provided here, the data are simply ranked and partitioned into four categories, each containing 25 percent of the observations. In Map C, one can see that most of the highest-rate counties for breast cancer are located in the eastern portion of the state, while the counties with the lowest rates are found in the west. However, county rates classified into the same categories are not necessarily similar to one another, as they would be under the natural break or statistical clustering methods.

For the animated time series maps of communicable diseases (rabies, tuberculosis, syphilis, and gonorrhea) that are found on the SCHS website (www.schs.state.nc.us/SCHS/healthstats/healthatlas.html), the quantile (or equal count) method was used for establishing the class breaks. The class breaks are held constant over the time period, which allows the dynamic shifts in class membership to be viewed from one year to the next.

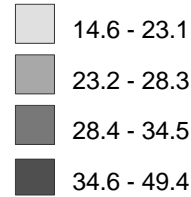
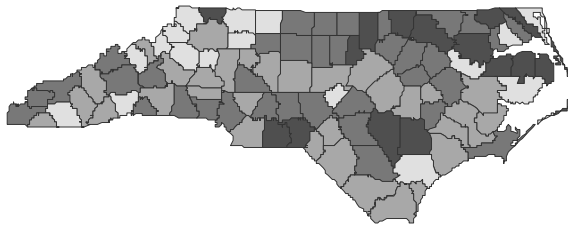
The *statistical* method classifies map data by the distance of the observational value from the distribution’s mean using standard deviations. The middle category (-0.5 to 0.5 standard deviations) will usually contain the mean value. As can be seen from Map D, most counties are fairly close to the mean. Based on the normal distribution, we would expect 6.7 percent of the rates to fall in the highest category (> 1.5 standard deviations), in this case six to seven counties. Map D bears this out, showing six counties in the highest category. The counties with the highest rates are again found in the northeast and south-central portions of the state, with relatively low values found in the northwest. Statistical maps, though a little more difficult to interpret, are useful for drawing attention to extreme cases.

Figure 5. Examples of Classification Methods

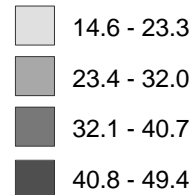
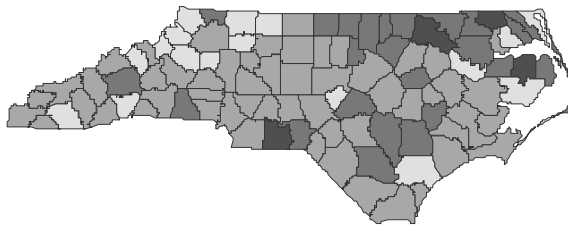
Breast Cancer 1994-1998 Age-Adjusted Death Rates

North Carolina 28.6

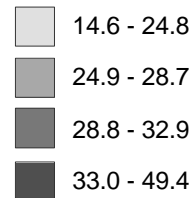
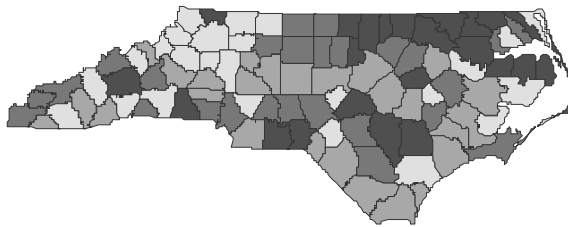
A. Natural Break



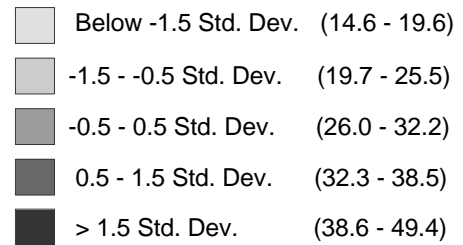
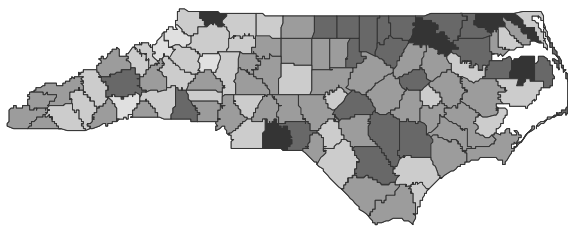
B. Equal Interval



C. Quantile



D. Statistical: Standard Deviations



Note: U.S. 2000 standard population used in age adjustment; death rates are per 100,000 females.

Source: State Center for Health Statistics

The best classification method will depend on the nature of data being mapped as well as the intent of the map designer. A table of the rates being mapped should accompany the map when feasible, as well as an explicit statement about which classification method is being used. This will enable the map-reader to make a more informed judgment about the observed patterns.

Conclusion

Maps provide a means to view the geographic relationships of health data. The relationships observed on the map may lead to hypotheses about which risk factors are contributing to the pattern. For example, high mortality and morbidity rates from a variety of causes are concentrated in the eastern portion of the state. One could hypothesize that this is associated geographically with poverty, educational levels, and/or access to health care. How well do the patterns of these potential risk factors correspond to the observed patterns of health events?

Maps reduce the multidimensional world to points, lines, and areas in a two-dimensional medium. Classification simplifies the underlying data distribution into several categories. The result is to distill complex data into simpler patterns as an aid to health program planning and policy making.

References

1. Devesa SS, Grauman DJ, Blot WJ, Pennello GA, Hoover RN, and Fraumeni JF. *Atlas of cancer mortality in the United States 1950-94*. NIH Publication No. 99-4564. Bethesda, MD: National Institutes of Health, National Cancer Institute, 1999.
2. Pickle LW, Mungiole M, Jones GK, White AA. *Atlas of United States mortality*. Hyattsville, MD: National Center for Health Statistics, 1996.
3. Wennberg JE. *The Dartmouth atlas of health care in the United States*. Chicago, IL: American Hospital Publishing Inc., 1996.

Note about the Authors: Jim Wilson is a Research Associate with the Center for Health Services Research and Development at East Carolina University, and also Adjunct Assistant Professor of Geography. He was at the State Center for Health Statistics at the time of this work. Paul Buescher works at the State Center for Health Statistics.



State of North Carolina
Michael F. Easley, Governor

Department of Health and Human Services
Carmen Hooker Odom, Secretary

Division of Public Health
Leah Devlin, D.D.S., M.P.H., Acting State Health Director

State Center for Health Statistics
Gustavo Fernandez, Ph.D., Director
www.schs.state.nc.us/SCHS/

The NC Department of Health and Human Services does not discriminate on the basis of race, color, national origin, sex, religion, age or disability in employment or the provision of services.

Department of Health and Human Services
State Center for Health Statistics
1908 Mail Service Center
Raleigh, N.C. 27699-1908