

## Conversion of Percentage Distributions to Weighted Index Numbers of Geographic Significance

JOHN FRASER HART, Indiana University

This paper is essentially a progress report concerning techniques for geographic manipulation of total percentage distributions. Although the percentage distribution here used for illustrative purposes is the age composition of the total population of Indiana counties in 1950, this is but one of many percentage distributions in which geographers have been or might be interested. The Census of Population alone, for instance, provides data on age by color and sex, on school enrollment and years of school completed, on residence, and on the employment status, occupation group, and industry group of the labor force, both for cities and for counties. Similarly, one may obtain data on housing conditions, or on major types of land use, and these are but a few of the types of percentage distributions which might interest geographers.

Our primary problem in utilization of such data is the cumbersome nature of existing techniques for their manipulation. Geographic analysis of these data has thus far been relatively unsophisticated; most commonly a single category from a single percentage distribution is selected for comparison either with other categories from the same distribution, or with selected categories from other distributions, but we have had few comparisons of two entire percentage distributions such as, for instance, an analysis of the relationship between variations, at the county level, of the age-sex distribution and the industry group distribution.

This problem can best be demonstrated by specific illustrations of the techniques which geographers have used in analyzing percentage distributions. It is no lack of modesty, but rather a desire to avoid even the appearance of criticizing others, that has prompted me to draw all these illustrations from my own efforts in this direction.

The simplest technique of analyzing the geography of any percentage distribution is to map each of its individual components. In considering land use patterns in Indiana, for instance, one may shade all counties falling within a certain percentage range with respect to each category (1). Each of these maps can be compared with any other map, of course, but this would necessitate a tedious, cumbersome, and usually highly subjective analysis.

The same objections apply, perhaps with even greater force, to the use of pie diagrams. One obtains a vivid graphic representation, to be sure, when the percentage of the manufacturing force engaged in the textile industry of the Carolina Textile Belt is shown in a pie graph, but it is almost entirely graphic (2). Further comparison with other distributions is extremely difficult, if indeed, it is possible at all.

Our concern with the use of such single categories might be considerably less were they not the basis for most of our functional classifications of urban places (3). Although these classifications purport to indicate the urban function, in the majority of instances they are based almost entirely upon the superabundance of a single occupational or

industrial category, and completely ignore lesser variations within the employment structure. In my functional classification of southern cities, for instance, there are four cities—El Paso, Tex., St. Augustine, Fla., Huntington, W. Va., and Bluefield, W. Va.—which are classified as transportation and communication centers because at least fourteen percent of the employed labor force in each is engaged in these industry groups. Yet the percentage engaged in mining, for instance, ranges from nil in St. Augustine to 0.5 in El Paso, 0.8 in Huntington, and 6.8 in Bluefield, and the percentage engaged in manufacturing varies from 5.2 in St. Augustine to 8.3 in Bluefield, 11.8 in El Paso, and 26.5 in Huntington.

It is variations such as these that have so often frustrated attempts at rational classification of urban function on the basis of employment data, and several attempts have been made to utilize the entire urban industrial structure rather than a single category. We have two techniques, the pie diagram and the columnar diagram—with variants—for portraying entire percentage distributions.

The pie diagram is most impressive, and it conveys a great amount of information within a relatively restricted space (1). But how useful is it for comparative purposes? One can examine the map of major land uses in Indiana, for instance, and see that there is considerable variation in acreage of cropland from one county to another, but it is virtually impossible to make any but the most subjective comparisons between counties in different parts of the state.

A similar objection can be raised to the columnar diagram, even when the columns are arranged in the formal rhythm of the age-sex pyramid (4). The map showing age-sex pyramids for each county in Indiana is quite an impressive sight, but it actually tells us relatively little. One can entertain himself mightily, of course, in examining the pyramids for the various counties and exclaiming over their vagaries, but it is difficult to compare one county with another, even when they are adjacent, and any comparison between the overall pattern of this distribution and any other distribution would appear beyond the ability of mere human beings. In other words, the map is an end in itself rather than a tool for further analysis.

The failure of the age-sex pyramid map as an analytical tool, in fact, prompted a retrogression to the mapping of single categories within the age-sex distribution, in the hope of obtaining new insight into variations in the percentage distributions of the population between the various quinquennial age-sex categories (Fig. 1). But the 1950 Census of Population provides data for seventeen separate quinquennial categories, and as it became obvious that the number was too great for ready comprehension they were grouped into four major categories in terms of demographic potential.

People under 20 may be considered children, and as a general rule they are not producing significant numbers of offspring. Counties with high percentages of children are most numerous in southern Indiana, and in the northwestern part of the state; there are relatively few counties in which children comprise less than a third of the population.

The years 20-44 are the child-bearing ages, and these are also the working years and the migrating years. The counties with the highest

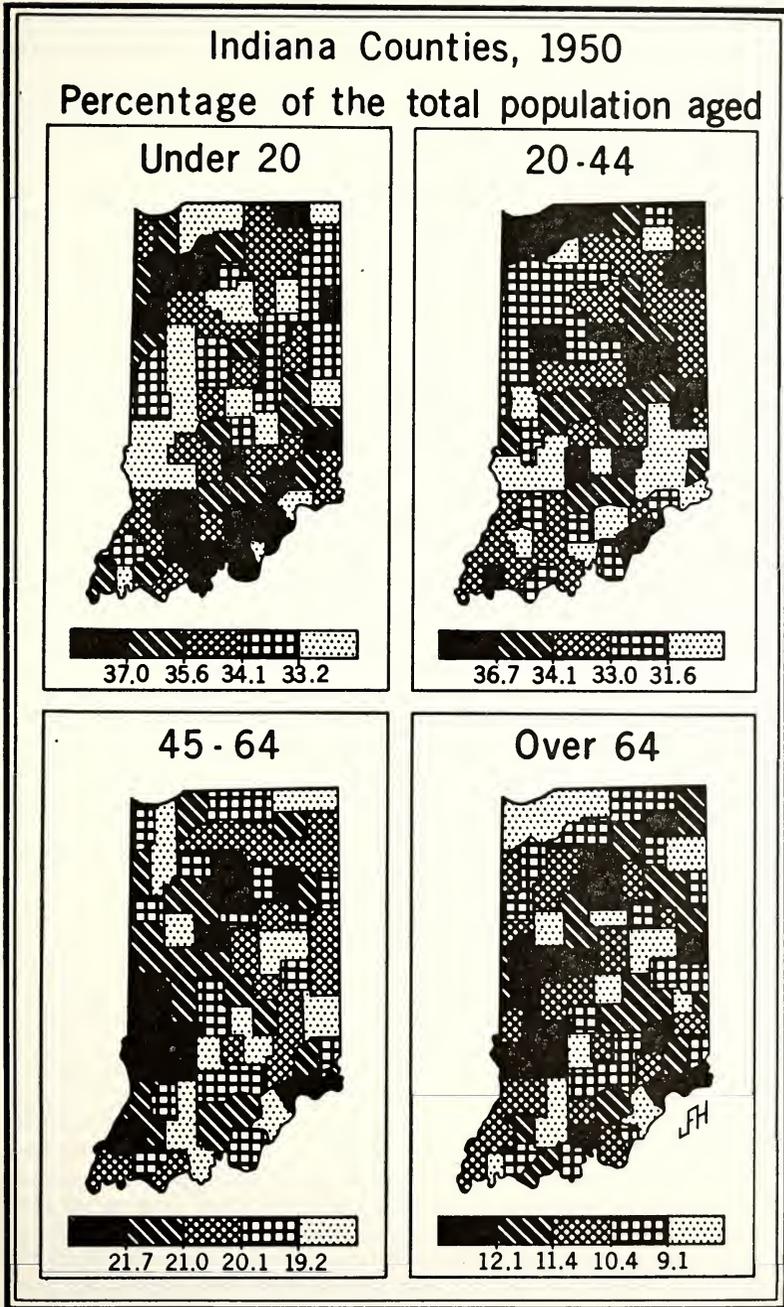


Figure 1. Percentage of the total population, 1950, in each of four age groups, by counties. The class intervals in each age group are based on quintiles.

percentages of their population in this age group are the urban counties, and they are noticeably concentrated in the center of the state and in the northern tier of counties.

The years 45-64 are the middle ages, and the people in this age group are normally past the child-bearing stage of life. Unusually large concentrations of people in this age group may indicate economic stagnation, as in the coal mining areas of western Indiana; the younger people have left, but the middle aged hang on to their homes and familiar associations. A population with a high percentage of its people over 45 is not only an aging population, but it is a dying one, because so many of its people are past the age when they could produce the children needed to maintain the population.

Despite the increasingly pejorative connotations of, and the numerous new euphemisms for the term, "old age" is commonly applied to the year over 64. Unlike the middle aged, a considerable percentage of the old population are no longer part of the labor force, and many of them have retired to areas away from those in which they were once employed. Their distributional pattern is complex, although they are least numerous, per capita, in urban areas.

This last observation, when coupled with the observation that persons in the 20-44 age group are noticeably concentrated in urban areas, can lead us to a useful generalization. Any percentage distribution may be likened to a closed hydraulic system; if it is pressed at one place it will expand at another. And so with the age distribution, or any other percentage distribution; a high percentage in one category will leave fewer people for the remaining categories, and all categories are intimately inter-related in that they must add up to one hundred percent.

Then why not, it might be argued, assign a distinctive weight to each category, multiply the percentage figure for the category by its weight, and total the products to obtain a weighted index for the distribution? It is impossible for two differing percentage distributions to have the same weighted index number, and a single number representing the entire distribution would facilitate comparison with other percentage distributions, or with any other distribution.

The possibilities are intriguing, and I personally believe that the use of properly weighted index numbers is destined to become an increasingly important tool of geographic analysis. Before such numbers can be used effectively, however, we require more information as to the best system of assigning weights to individual categories. It would be possible, of course, to assign each category a distinctive weight at random, and this would produce distinctive and unique index numbers; for instance, we could assign a weight of 1 to percentage aged under 20, 2 to percentage aged 20-44, 3 to percentage aged 45-64, and 4 to percentage aged over 64.

The usefulness of index numbers obtained by such a random weighting system would be inhibited, however, by the fact that they would vary without meaningful pattern. It would be preferable to assign weights so that all counties with low index numbers would be relatively similar to each other, and quite different from the counties with high index numbers; counties with intermediate numbers, of course, would be transitional between these two extremes.

McIntosh demonstrates that this actually can be done with some plant communities (5). He marked individual celluloid strips with the percentage of black oak, white oak, red oak, and sugar maple in each of a number of different stands. When he arranged the strips in order of decreasing black oak and increasing sugar maple, he found that both white oak and red oak peaked at distinctive intermediate points, and he was able to assign each species a weight which had meaning in terms of the inherent character of each stand. A stand dominated by black oak, for instance, had a low weighted index number, whereas a sugar maple stand had a high weighted index, and stands dominated by white and red oak had intermediate indices. These weighted indices proved a useful tool in investigating the relationship between plant associations and soil acidity, in the study cited.

This method of assigning weights to individual categories appeared so promising that it was used in an attempt to assign weights to individual age groups in Indiana (Fig. 2). The counties were ranked in order of increasing percentage of the group aged 20-44, and each of the other three age groups in the county is indicated by a dot at the appropriate percentage. The resulting pattern obviously is not so simple as that found by McIntosh, and a considerable degree of subjective interpretation is necessary. It would appear, for instance, that the percentage of aged people decreases as the percentage aged 20-44 increases, and a similar but slighter decrease is noticeable in the 45-64 age group. It would be a brave man indeed, however, who would attempt to assign weights to individual age groups on the basis of this graph, and hence it has been regretfully laid aside, at least temporarily, because it requires excessive subjective interpretation. Casual observation, however, has indicated that this technique might prove much more useful, and rather less subjective, if quinquennial age groups were used in place of those shown here. There appear to be some interesting relations between quinquennial age groups spaced at twenty-year intervals—or a generation apart—but my investigations along this line have not yet progressed to the point that a formal report is justified.

A second technique of assigning weights, which was used by Hagood in devising her rural level of living index, would appear to merit consideration by geographers. She selected a comparatively large number of characteristics of rural population and housing, and correlated each characteristic with every other (6). The resulting table of correlations shows that some characteristics are so closely allied that either may be used, and the table also provides a basis for weighting the characteristics which are combined in the rural level of living index.

TABLE I

Coefficients of Correlation between the Percentages of the Population in each of four age groups, Indiana counties, 1950

<i>Group aged</i>	<i>Group aged</i>		
	Under 20	20-44	45-64
20-44	0.66		
45-64	0.75	0.71	
Over 64	0.72	0.75	0.79

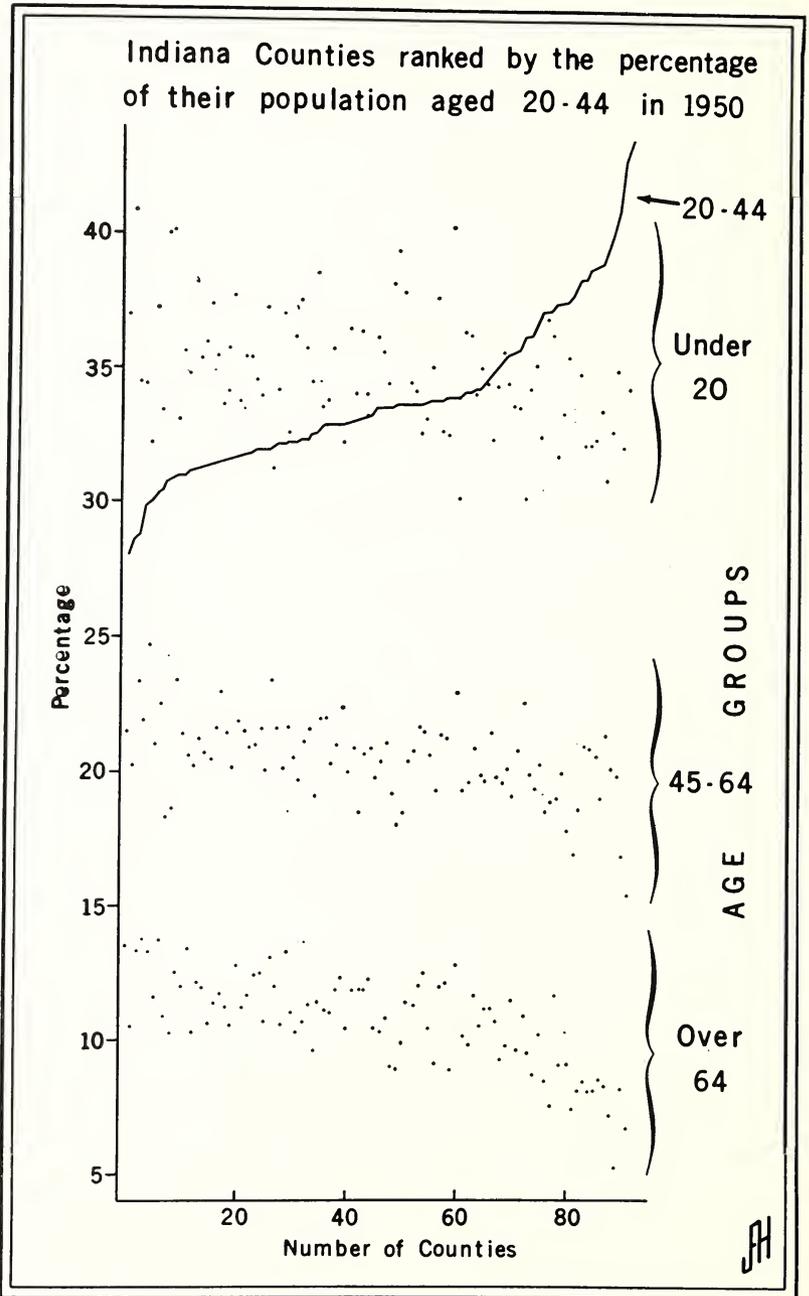


Figure 2. Distribution of the population of Indiana counties in three age groups when the counties are ranked in order of increasing percentage in the age group 20-44.

In an extremely restricted sense I have tried to apply her technique to the age structure of Indiana's population in 1950. I used only four age groups, instead of the seventeen for which data are available, and within each age group I divided the counties into quintiles by percentage so that coefficients of correlation could be computed by the simple method explained by Freund (7). There was a relatively high degree of correlation, possibly because of the restrictions imposed by the method of computation. The closest correlation was between the 45-64's and the over 64's, whose coefficient was 0.79; at the other end of the scale the coefficient between the under 20's and the 20-44's was only 0.66 (Table I).

But here again it seems that more fruitful results would have come from use of all available age groups, with absolute percentages rather than quintile groups. The major obstacle, of course, would be the cumbersome nature of the calculations necessitated, for 136 coefficients of correlation would have had to be computed. It does seem to me that the potential results would appear to warrant these calculations, although I believe that it would be extremely helpful, if not mandatory, to have mass data processing equipment available. If the job is undertaken, I would further suggest that it be done for a representative sample of all counties in the United States, rather than for Indiana counties alone. Hagood, for instance, devised her weights on the basis of a sample of 200 counties, little more than double the number in Indiana.

I do believe that the correlation of each quinquennial age group with every other for such a sample would reveal that some age groups are so closely related that they might be used interchangeably in an index of age structure, and that relations between other groups are such that they might be weighted to provide a rationally weighted index of the age structure of any unit of area. This index, being a single number, could then be used for further comparison and correlation with other data, whether statistically or cartographically. And I believe that a series of such indices, for the various sorts of percentage distributions cited at the beginning of this paper, would enable geographers to obtain new insight into unsuspected relationships between percentage distributions.

#### Literature Cited

1. JOHN FRASER HART. 1959. Forest land use in Indiana. *Indiana Academy of the Social Sciences, Proceedings, New Series*, 3:65-72.
2. EUGENE MATHER and JOHN FRASER HART. 1954. Industry in the Deep South and the Border States. *Tijdschrift voor Economische en Sociale Geographie* 45:108-112.
3. JOHN FRASER HART. 1955. Functions and Occupational Structures of Cities of the American South. *Annals of the Association of American Geographers* 45:269-286.
4. JOHN FRASER HART. 1958. Age Pyramids for Indiana's counties and larger cities. *Indiana Academy of Science* 67:187-193.
5. ROBERT P. MCINTOSH. 1958. Plant Communities. *Science* 128:115-120.
6. MARGARET JARMAN HAGOOD. 1943. Development of a rural-farm level of living index for counties. *Rural Sociology* 8:171-180.
7. JOHN E. FREUND. 1952. *Modern Elementary Statistics*, New York: Prentice-Hall. 270-276.