GRAPHICAL METHODS FOR PRESENTING STATISTICAL DATA: PROGRESS AND PROBLEMS

Roberto Bachi
Hebrew University (Israel)

THE PROGRESS ACCOMPLISHED IN AUTOMATION OF GRAPHS AND ITS IMPLICATIONS

Statistical maps are of great interest both to cartographers and to statisticians. Cartographers and geographers consider such maps as a very important category of thematic maps. Statisticians view them as applications of the graphical-statistical method to the special case in which the data to be graphed are detailed geographically. Professor Jenks and Professor Robinson have discussed statistical maps from the first viewpoint. I shall discuss them from the second. In so doing I shall draw largely from the conclusions of a meeting on geographical statistical methods held recently at the 40th session of the International Statistical Institute in Warsaw 1/ and from the experience gained from an exhibition of automated statistical graphs and of new graphical symbols held at that session.

The most impressive and encouraging development that has occurred in statistical graphics in the past two decades or so is the tremendous increase in technical facilities for the automated production of statistical graphs of many types (diagrams, bar and pie charts, stereograms, statistical maps, etc.). There is no need in this symposium on computer-assisted cartography to discuss such facilities or even to list them because they are well known to all participants.

However, it may not be superfluous to perhaps stress some of the consequences of this development. The most obvious of these is the enormous growth in the mass production of graphs resulting from the use of cheap and quick methods. Additionally, important shifts in the uses of graphs, resulting from applications to new fields beyond the traditional ones (illustration of findings, spreading knowledge among the general public, and teaching), should be noted. For instance:

1. Graphs are beginning to be applied as tools which may assist in solving the difficult problem confronting many statisticians today -- that of mastering the enormous quantity of data produced by the computer.
2. The automated graph has great potential as a research tool. Bringing onto the computer's screen parallel sets of data in graphical form and graphically comparing actual data to fitted models can help to obtain an overview of the data, to explore their meaning, to discover regularities and irregularities, to suggest, test or discard scientific hypotheses.
3. Animation of graphs may constitute an additional tool of research, mainly in regard to dynamic phenomena.
4. Graphs may be of help in communicating to politicians and to business and administrative executives the statistical information needed to make decisions.

By the way, graphs used for the various purposes may require different properties. For example, while graphs for executives should be easily understandable, graphs for internal use by statisticians or scientists may be more sophisticated and may not require the same degree of aesthetic properties. On the other hand, pleasant and attractive graphs, such as those obtained today by the automatic use of color, may have important applications as illustrative, educational and mnemonic tools.

However, all graphs should be produced in a way which guarantees objectivity and honesty in the transmission of information, and which gives a trustful representation of the data, apt to be interpreted quickly and in a more or less similar way by their readers, To reach this aim they should stand up to acceptable scientific standards.

DANGERS IN THE CURRENT DEVELOPMENTS OF GRAPH PRODUCTION

Unfortunately this is not true of a considerable part of the graphs produced today. From this viewpoint, it appears that we are not yet fully prepared to meet the challenge of mass production of graphs and of their use for very responsible purposes, such as those mentioned above. Actually in many cases no clear rules for the production of graphs seem to exist and, if they exist, they are largely neglected by producers of graphs. Widespread production of bad graphs may have, among other consequences, executives using such graphs making the wrong decisions and this may, in the long run, bring discredit upon the whole graphical method.

The situation seems to be partly related to insufficient education on the part of graph makers and partly to insufficient development of scientific research and thinking in the graphical field.

If we try to summarize today the "state of art" of statistical graphics, it appears that the common graphical methods enable us to represent in a satisfactory way only a rather limited part of the many types of statistical data currently produced.

EXAMPLES OF LIMITATIONS OF COMMON GRAPHICAL METHODS

To appreciate some of the limitations of common graphical methods it is sufficient to consider a few examples:

(1) Consider first the common diagrammatic method, which is considered the best from a scientific point of view and which renders invaluable services to statistics. This method is severely limited due to two well-known reasons: (a) the diagram employs two coordinates (x,y) to represent respectively the values of the characteristic according to which the data are ordered, and the statistical data. Therefore, while one linear series covers only one line or column (one dimension) in a table, its graphical presentation requires two dimensions in the plane. This creates serious difficulties when we try to represent graphically in the same graph many parallel lines or a composite series. The use of a logarithmic scale can help to find a practical solution to this problem, but it is warranted for only certain types of data. (b) The lack of generally accepted criteria for linking the scale of

x and y is a problem which has intrigued statisticians for a very long time. Every graph maker knows this problem and is aware of the fact that by stretching and compressing the scales, he can completely change the aspect of the graph. The consequences of this have been put in evidence by the impertinent and yet so pertinent remarks by Huff on "how to lie with statistics." 2/

(2) If it is desired to represent a function of two variables, $z=f(x,y)$, the best solution -- in theory at least -- is given by stereograms. Stereograms can be plastically built and may be very useful; however, their construction and use is not simple. Preparation of two-dimensional graphs, showing in perspective the stereogram, can be performed with the help of the computer. However, their interpretation is not easy because not all readers are really able to follow from such graphs how z changes in function of both variables; moreover, the aspect of the graph may change according to the perspective used. Use of contour lines to represent the stereogram avoids difficulty. However, here also considerable skill is required of the reader to interpret the graph, mainly when "depressions" and "elevations" are found in the same graph and if they can be distinguished only by reading the numbers indicating the levels of each line.

In consequence of that we lack a commonly accepted and simple device for graphically representing sets of statistics of great importance such as grouped correlation tables or contingency tables.

(3) Many of the choropleth maps actually produced today are far from being satisfactory for the reasons mentioned by Professor Jenks in Monday's plenary session. Another defect of choropleth maps is explained below.

It is generally accepted today that any series of average values ($\overline{X}i$), referring to the regions i in which a territory is divided, can be presented graphically by covering the area (Ai) of each region i on the map with an appropriate pattern, shading or color.

This method is correct whenever the $\overline{X}i$ is an average value referring to the area of region i and whenever the conditions of the areas of the regions are the object of our research. Consider, for instance, Graph 1, which shows the average value of agricultural production per hectare (=10,000 square meters) in each of the provinces of Italy. Here

$$\overline{X}i = \frac{Xi}{Ai}$$

where Xi is the total value of the agricultural production in i.

In order to simplify our discussion let us suppose that the pattern used to represent $\overline{X}i$ has a blackness proportional to $\overline{X}i$, and let us forget the inability of the human eye to judge correctly the degree of blackness.

Under such hypotheses our map has some very useful and simple properties:

a. The visual impression received by considering the territory of i on the map is proportional to the product of the blackness per unit of area of i and of the area of i, viz., to

$$\overline{X}iAi = Xi$$

which is the total value of the agricultural production in i.

b. The visual impression received by considering the entire map is proportional to

$$\Sigma\overline{X}iAi = \Sigma Xi = X$$

viz. to the total value of the agricultural production in Italy.

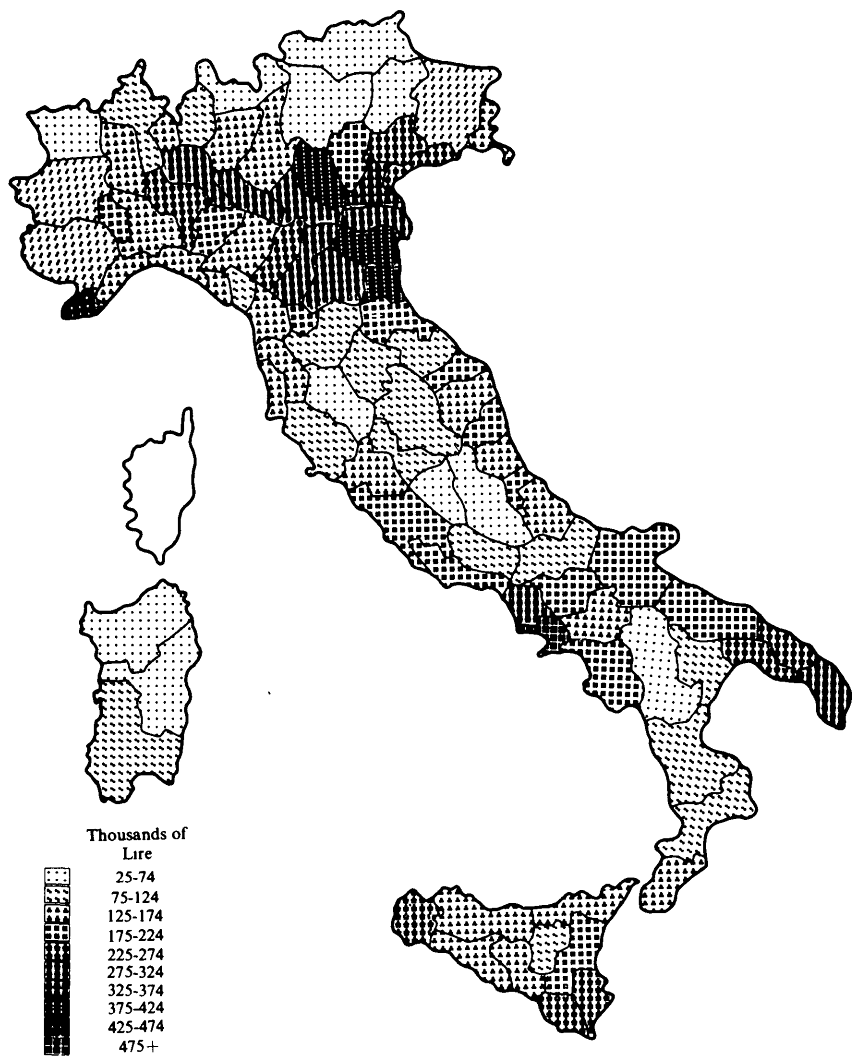
c. Suppose that maps are constructed in the same way and with the same scale for the value of agricultural production in Italy for different years. The comparison between such maps will convey correctly not only information on the changes occurred in the course of time in the average value of agricultural production in each province, but also on the total changes of the value in each province and in the entire country.

However, let us consider now the very frequent case in which the regional averages $\bar{X}_i$ do not refer to the area of i but to its population, whose conditions we wish to investigate. For instance, let us suppose we wish to study: (1) the average income of the population in each region i, (2) the birth rate per 1,000 inhabitants of i, or (3) the proportion of people aged 65 or over among the population of i. If we write

$$X_i = \frac{X_i}{P_i}$$

where Pi is the size of the population of i, Xi represents in the above examples: (1) the total income produced in i; (2) the total number of births in i; and, (3) the number of people aged 65 and over in i.

Adopting the usual method of graphical presentation has in this case the following consequence. The visual impression received by considering the territory of i on the map is again proportional to $\bar{X}_i A_i$. However, as $\bar{X}_i = \frac{X_i}{P_i}$, $\bar{X}_i A_i = \frac{X_i}{P_i} A_i$, which does not correspond to the object of our research, and is often of little or no meaning. Properties a, b, and c listed above are lost, and the map may give very misleading impressions. In order to appreciate the dangers of this graphical presentation let us consider a few common examples:



Graph 1 Italy (1965). Value of gross production from agriculture and forestry per hectare of territory
Source of data "Moneta e Credito", September 1966 *Scale* 5000 Lires = 1 GRP unit

- Graph 2 represents the median income of families in each of the States of the United States as published in the Report of the U.S.A. Census of 1950. As the number of patterns is too small, and the patterns used are arbitrary, the map has been redrawn in Graph 3 by using Graphical Rational Patterns (See the section entitled "NEW GRAPHICAL SYMBOLS. GRAPHICAL RATIONAL PATTERNS," for an explanation of such patterns), proportional to the average 3/ income $\overline{X}i$ and repeated (according to the method generally followed) over the entire territory of each State. To understand the implication of this method compare in Graph 3 Utah and New Jersey, which have a very similar median income per family. Utah is by far more impressive than New Jersey in the map as it has a territory almost 11 times larger. However, as New Jersey has a population almost 8 times that of Utah, its importance (in regard to the economic conditions of the people) is much larger.

- Consider the beautiful maps of the Urban Atlas produced by the Bureau of the Census for presenting the data of the census of 1970. (See Figures 5 to 8, pp. 256-259.) In such maps many census tracts can roughly be considered to have a population with the same order of magnitude. In the maps which describe the conditions of the population (such as percentage of population in each age group, average income, educational conditions, etc.) each tract i should contribute to the creation of the general visual impression in a way roughly proportional to $\overline{X}i$. Actually it contributes in a way proportional to $\overline{X}iAi$.

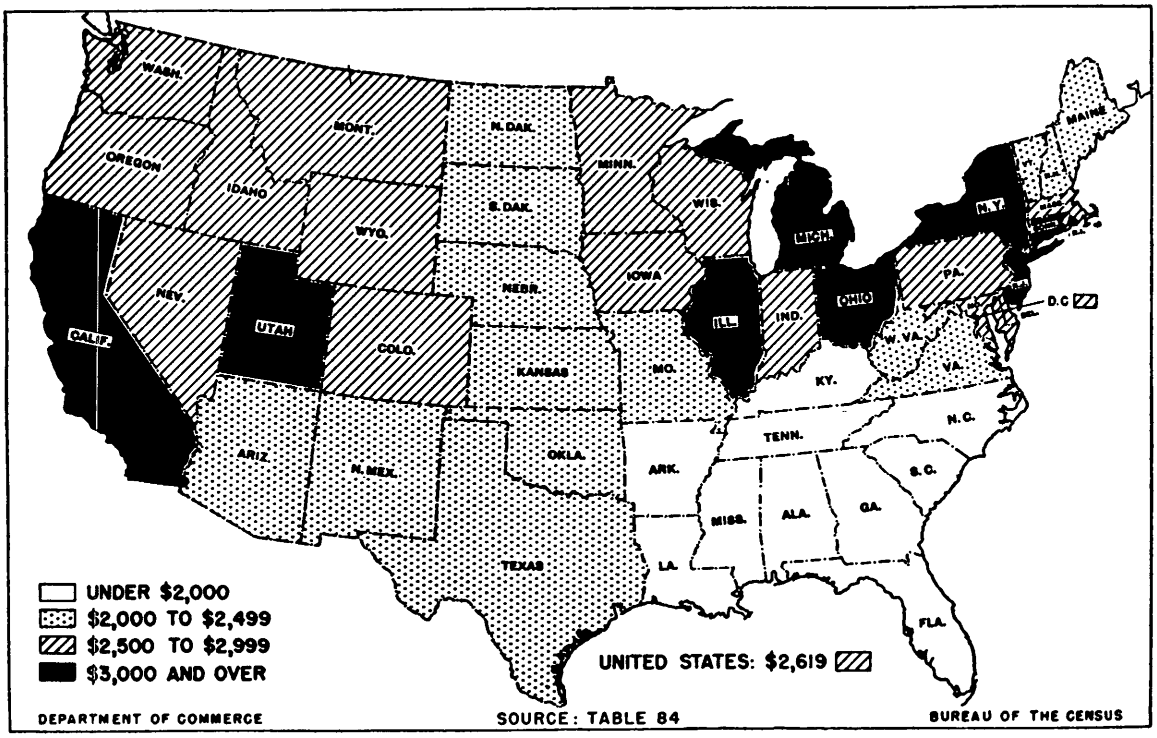
  Generally speaking, this implies that wide suburban tracts have a disproportionately large importance in catching the eye of the reader in comparison with the small central tracts. It may happen that in a town only a few large suburban tracts dominate the entire map, while tens of central tracts have an almost negligible visual importance. As a consequence, the average visual impression that we receive from the entire map may be very different from the average $\frac{\Sigma Xi}{\Sigma Pi}$ which is given in the legend of the map.

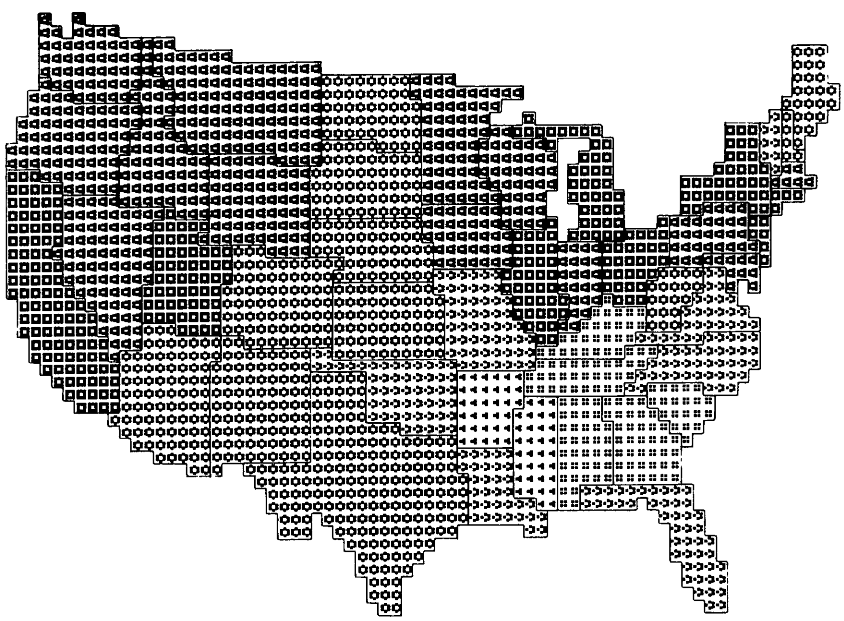
- A similar situation is found whenever the population is very unevenly distributed (which is very often true). Then the eye of the reader is caught by the $\overline{X}i$ in the wider regions, which are often those with low population density (being desert or semidesert or mountainous, etc.). By contrast, the reader may neglect to consider the areas, generally small, where the majority of the population is concentrated and which have decisive importance in determining the average $\overline{X} = \frac{\Sigma Xi}{\Sigma Pi}$ for the entire population.

POSSIBLE ELIMINATION OF PRESENT LIMITATIONS OF GRAPHICAL METHOD

The above criticisms of the usual graphical methods should not bring us to the negative conclusion that most graphs cannot be based on good scientific criteria. The contrary is probably true. Many of the existing limitations can probably be eliminated, if we are ready to make efforts to seek solutions. In the following a few examples are given of work in that direction. The proposals given are to be considered as provisional and may still need study, discussion, and improvements in the technical implementation of some of the tools introduced. However, they seem to show that attempts at improvement are feasible.

Graph 2: MEDIAN INCOME IN 1949 OF FAMILIES AND UNRELATED INDIVIDUALS, BY STATES: 1950





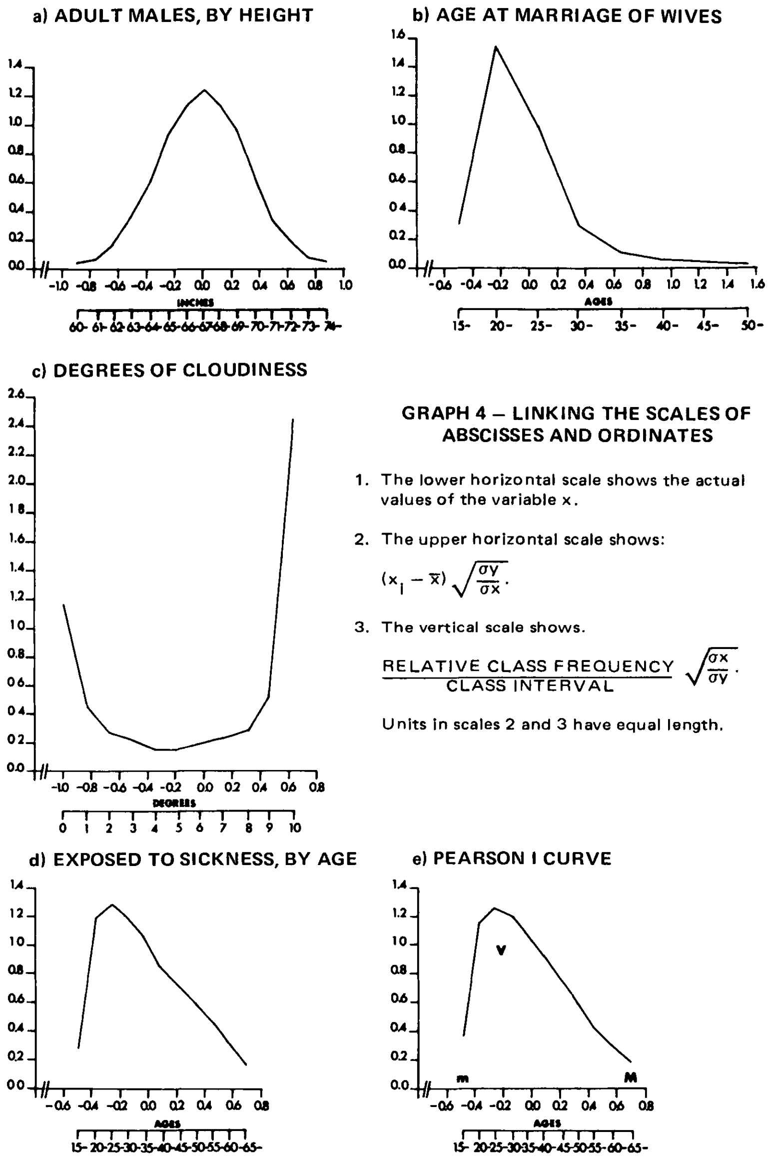
Graph 3: USA (1950) Median income per family

Patterns (X_1) 1200-1600 1600-2000 2000-2400

2400-2800 2800-3200 3200-3600

LINKING x- AND y-SCALES IN DIAGRAMS

In the following an attempt (taken from an article -- in preparation -- by Bachi and Samuel) will be presented in regard to possible linking of x- and y-scales in graphical presentation of one-dimensional statistical distributions (probability density functions, empirical frequency distributions, or histograms). Let us limit ourselves here, for the sake of simplicity, to the case in which the main aim of graphical presentation is that of enabling us to compare quickly the shape of distributions irrespective of their location and dimensions. In the (x,y) plane consider a density function f(x) as shown by Graph 4e which determines a figure V defined between x = m and x = M and between y = 0 and y = f(x). Let σ_x and σ_y denote the standard deviations of the elements of V along the x and y axes respectively.



GRAPH 4 – LINKING THE SCALES OF ABSCISSES AND ORDINATES

1. The lower horizontal scale shows the actual values of the variable x.
2. The upper horizontal scale shows: $(x_i - \bar{x})\sqrt{\frac{\sigma y}{\sigma x}}$.
3. The vertical scale shows. $\frac{\text{RELATIVE CLASS FREQUENCY}}{\text{CLASS INTERVAL}}\sqrt{\frac{\sigma x}{\sigma y}}$.

Units in scales 2 and 3 have equal length.

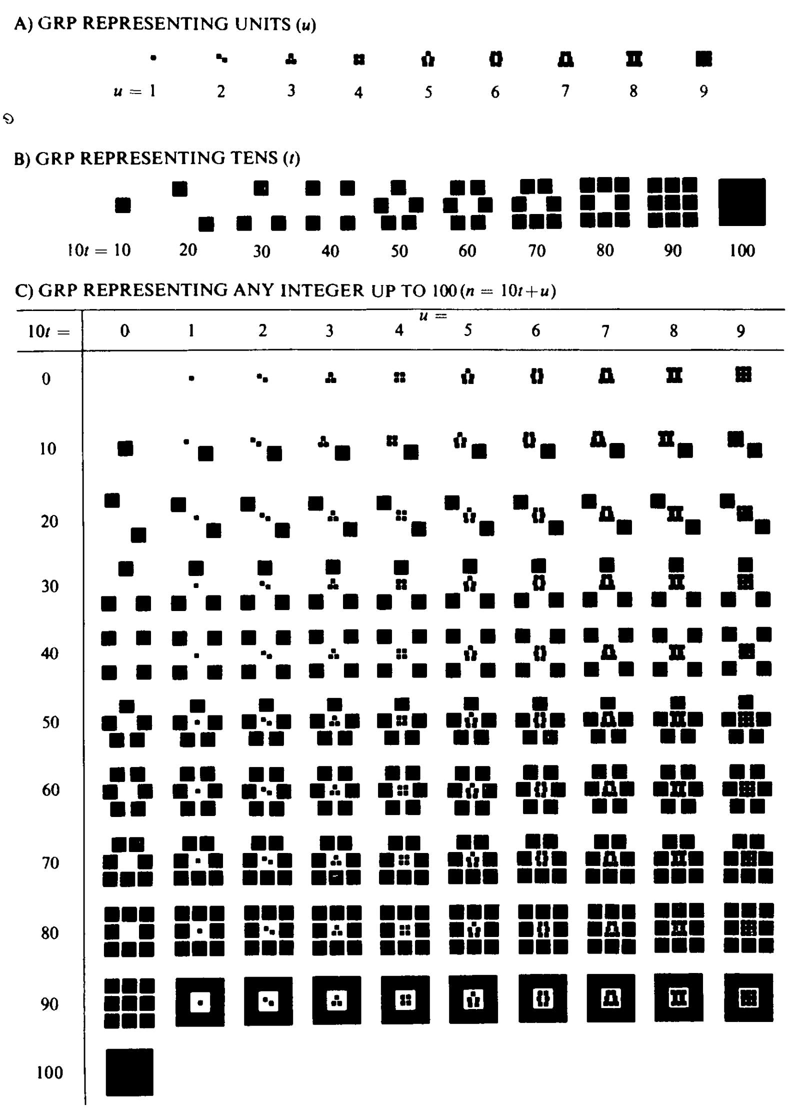
In order to standardize graphical presentation, it is proposed that x- and y-scales be based on the simple criterion of giving the same length to represent σ_x and σ_y. This means that in any of the usual diagrams the original x-scale is to be corrected by a factor (σ_x/σ_y), and the original y-scale should be corrected by a factor (σ_x/σ_y). This correction can easily be performed also in regard to any empirical frequency distributions by applying appropriate formulas. The illustrations in Graph 4 show examples of empirical and theoretical distributions drawn according to the above rules and indicating respectively: (a) the height of 8,585 British male adults (see Yule and Kendall, 1950, p. 82); (b) the ages at marriage of 315,300 English wives (ibid., p. 201); (c) the degree of cloudiness in 1,715 observations at Greenwich (ibid., p. 96); (d) the distribution of age of 2,995,724 persons exposed to risk of sickness (Elderton, 1953, p. 7); (e) Pearson I curve fitted to (d) (ibid., p. 62). The method proposed may be connected with the use of certain parameters of shape of the distributions, which cannot be discussed here. The method enables us to discover quickly similarities of shape between empirical distributions and between them and theoretical models. For instance, all empirical distributions similar to a normal curve will have an appearance near to that of 4a; the empirical distribution 4d appears to be very similar to the theoretical distribution 4e, etc.

NEW GRAPHICAL SYMBOLS. GRAPHICAL RATIONAL PATTERNS

In the past decade or so various proposals of new graphical symbols have been put forward to fill the gap in the current graphical methodology. Some of them, such as Chernoff faces 4/ have been discussed in other papers at this symposium. Others, such as Bertin's distinguishable circles, 5/ have been mentioned. Here we will describe a system called Graphical Rational Patterns (GRP) 6/ which may help in seeking solutions to some of the problems mentioned above.

In its simplest form, the GRP is a pattern representing any integer $n = 10t + u$ by u unitary square marks of an area a (which indicate the units) and t square marks of area 10a (which indicate the tens). Correspondence between n and the pattern representing it is ensured in two ways: (1) the blackness of the pattern is n times that of the pattern representing 1; (2) the marks forming each pattern are clustered in a way which enables us to easily read the value of the pattern whenever the need arises. The patterns are drawn in such a way as to require very little space and, as a matter of convenience, they may be enclosed within a small square frame (see Graph 5). Besides the single patterns described above, repeated patterns can be used, in which the symbol is repeated many times either (a) in a line to show that a value n is to be attached to a given line (such as a traffic line or an isoline) (see Graph 6); or (b) over an area to show that a value n is to be attached to a given region in a map or a given portion of a diagram (see Graph 7). The scale of the patterns can be extended to $n > 100$ (for instance up to $n = 1{,}000$).

The scale can be adapted to represent any value $0 \leq nk \leq 100k$ by attributing any arbitrary value k to the elementary unit square of area a and multiplying by k the values of the patterns. By printing the GRP in color (or in black over a colored field) or by enclosing it by frames of various types, it becomes possible to add qualitative information indicated by colors or type of frame to quantitive information given by the GRP. Proper methods can be used to adapt GRP to show negative versus positive data. Various attempts have been made to automate GRP: (a) A scale of n = 1,.......,10 has been built by combining dot, line, and other symbols of the high speed line printer into a form more or less similar to that of the

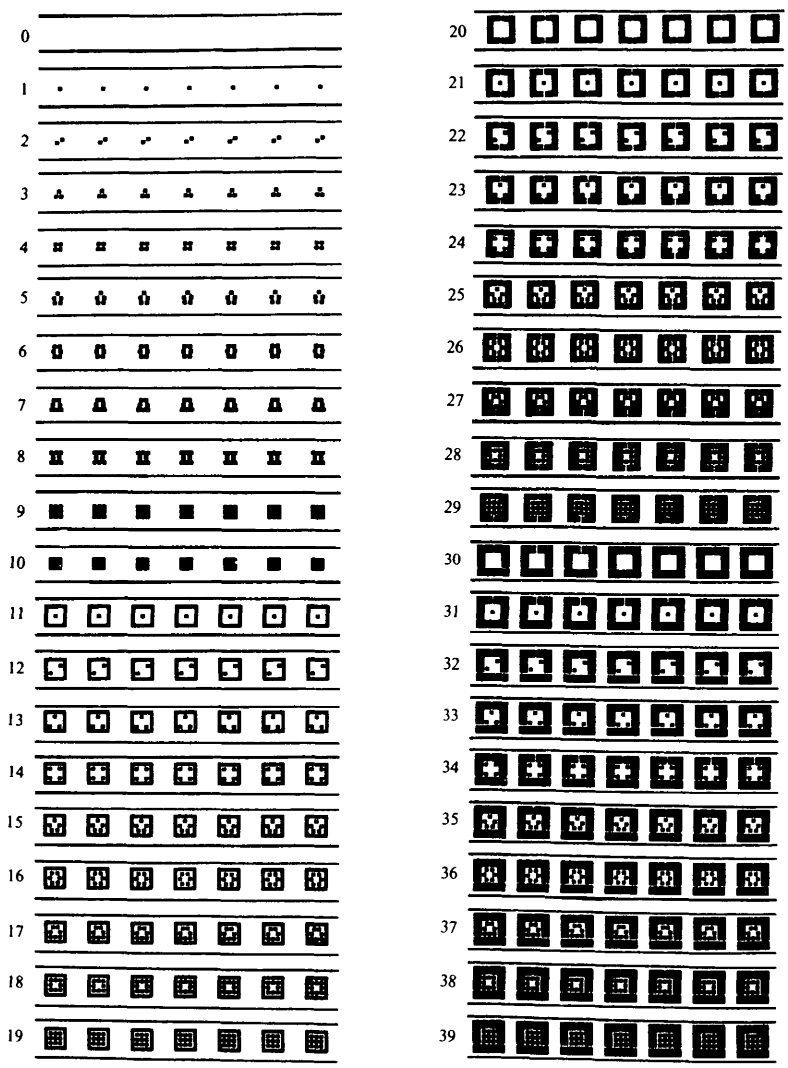


Graph 5

Basic scale of single GRP showing integer numbers up to 100 original GRP, and a scale of tens has been prepared by repeating in a convenient number and way the symbol for 10; (b) A scale of GRP has been built by the plotter. However, as (a) was not entirely satisfactory from a scientific viewpoint, and as production of (b) was too slow, the GRP are now produced (c) on a photo-electric typesetter which enables us to obtain the GRP in the desired form and size on any desired spot on a film. Moreover, (d) GRP are being put at present on CRT.

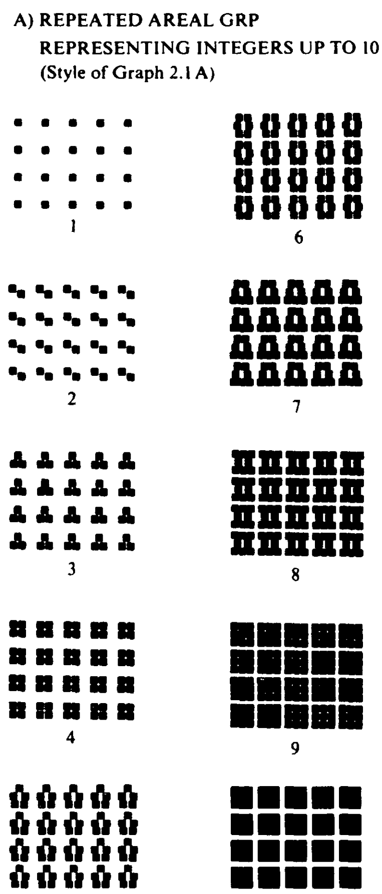
The properties of the GRP system, as compared to those of the common diagrammatic system can be seen from Graph 8, where a few selected numbers n are represented by the two methods. It is seen that (a) both methods enable us to represent

A) REPEATED LINEAR GRP OF VARIABLE SIZE REPRESENTING INTEGERS UP TO 39
(STYLE OF GRAPHS 2 1A and 2 3A)



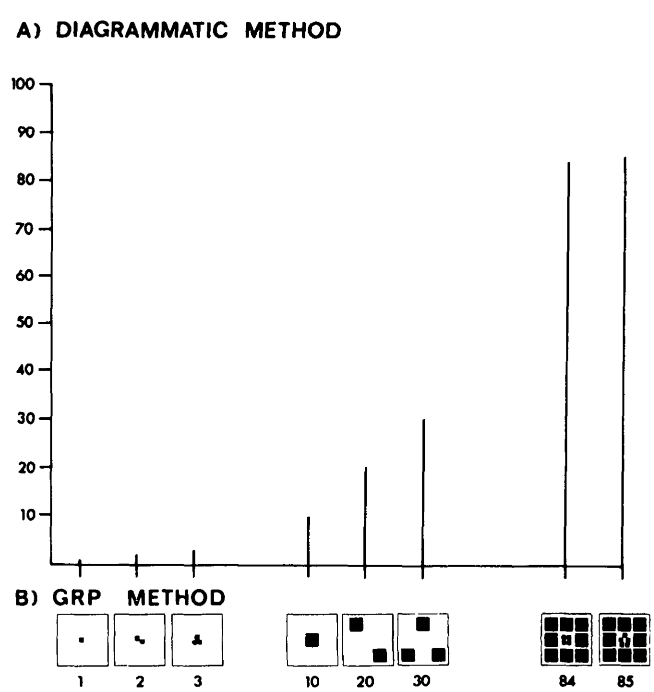
Graph 6 Scales of repeated linear GRP

in an accurate way the numbers n. In the diagrammatic method correspondence is between n and the length of the ordinate: reading requires comparison with the scale. In the GRP system correspondence is between n and the blackness of the pattern. After a little practice, it is possible to read n directly from the pattern. (b) While the diagrammatic method demands two dimensions for representing a series of n, the GRP method demands one dimension only. Therefore, many series of data can be represented in one GRP graph, although this is not possible in common diagrams. (c) After the reader has mastered the rules of formation of GRP, he can appreciate even small differences between patterns. Thus in Graph 8 the differences between n = 84 and n = 85 is clearer in the GRP than in the ordinate. (d) While the diagrammatic method does not show equality of ratios such as 2/1 = 20/10, 3/2 = 30/20, the GRP do (as do semi-logarithmic charts). (e) On the other hand, the GRP are clearly at a disadvantage in comparison with the usual diagrammatic method in the following respect. In the usual diagrams it is often legitimate to join ordinates of consecutive data and to judge differences between



Graph 7

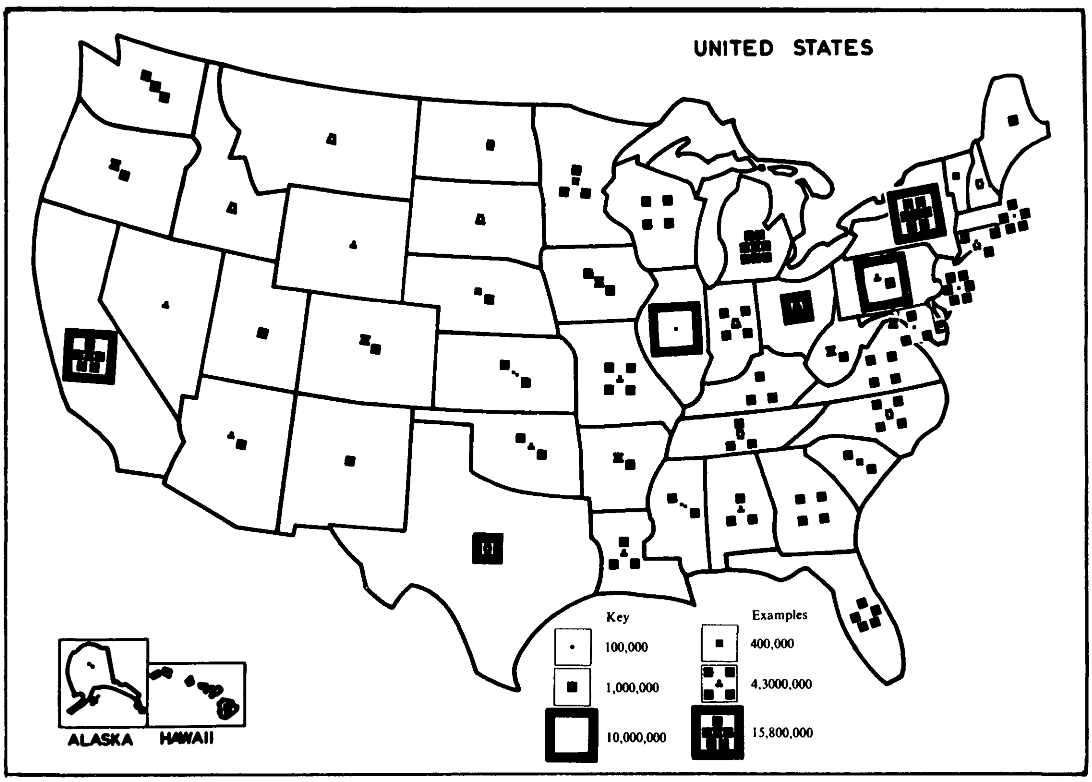
Scales of repeated areal GRP.



Graph 8 : Comparison of GRP Method and Diagrammatic Method them (or in semi-logarithmic charts, ratios) on the basis of inclination of the diagrammatic line. This additional powerful tool of information is not available in the GRP system. However, as GRP and diagrammatic methods can be integrated and used together, it may become possible to utilize in a proper way the advantages of both systems.

In the following a few examples of applications of GRP to statistical maps are shown.

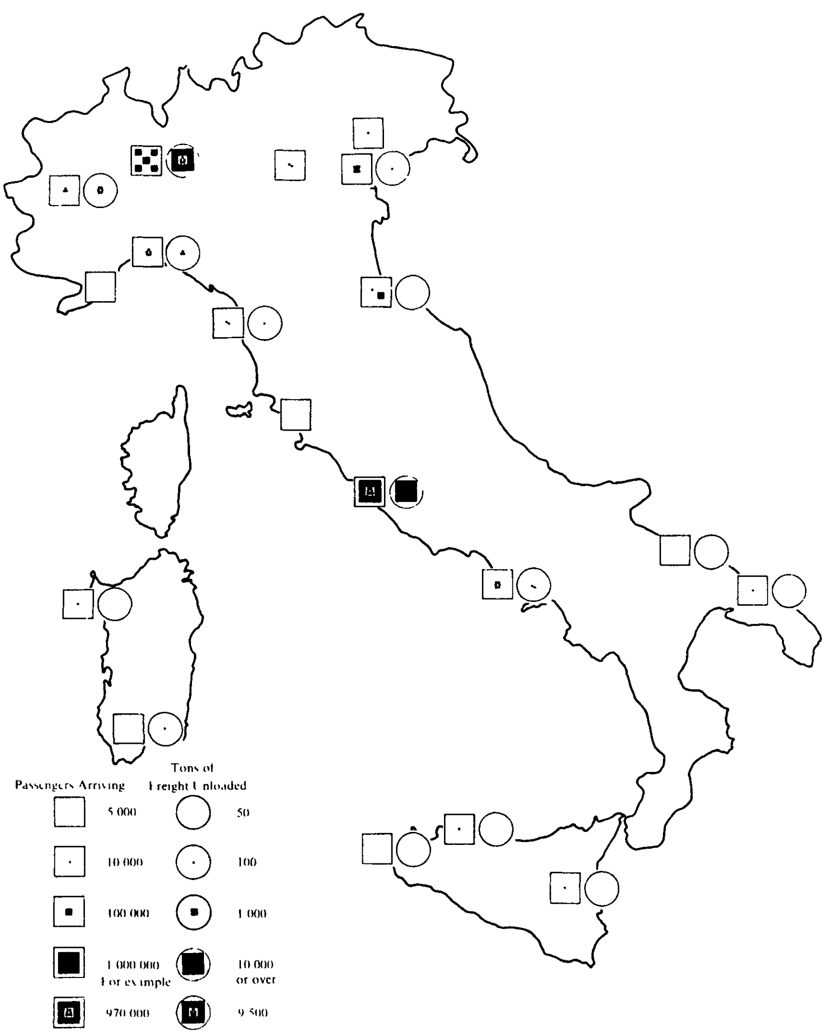
Graph 9 presents a map of the absolute size of the population in each of the States of the U.S.A.; this map is built only for readers to whom the internal divisions of the U.S.A. are known and meaningful. As the population symbol is put at the center of each State, and as the States have different areas, the map is not intended to show variations in densities over the areas of the U.S.A. Despite this, it gives some broad information on prevailing patterns of population distribution. Maps of population densities can be built with repeated GRP by a method similar to that illustrated in Graph 1.



Graph 9 U.S.A. (1960). Population in each state. *Source of data:* Statistical Abstract of the United States, 1962. Washington, Bureau of the Census, p. 10. *Scale:* 100,000 inhabitants = 1 GRP unit.

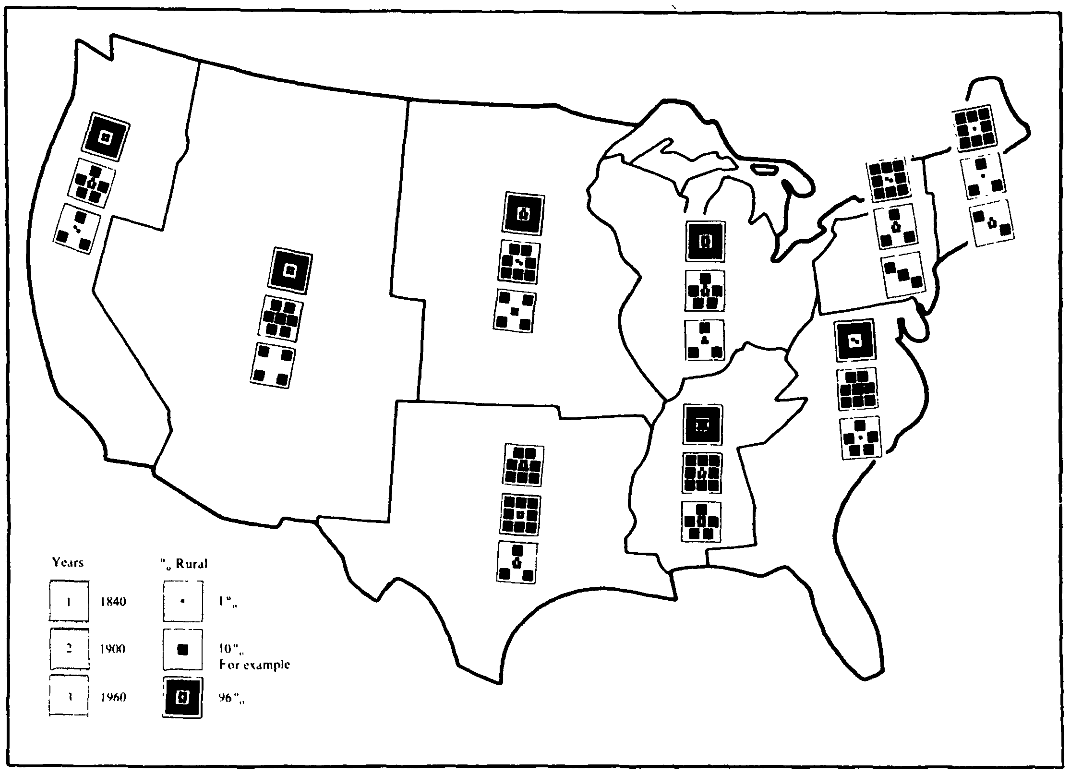
Graph 10 gives an example of presentation of two geographical series in the same map: traffic of passengers in Italian airports is given within a square pattern; traffic of merchandise is given within a circle. If instead of these different frames, colors are used to distinguish each series, far better results can be obtained.

Graph 11 illustrates an application to a series in which the data (indicating percentage rural in each of the divisions of the U.S.A.) are classified both by geographic regions and by time (censuses of 1840, 1900, 1960). Here too the use of colors would greatly improve the presentation by faciliting the association-dissociation processes.

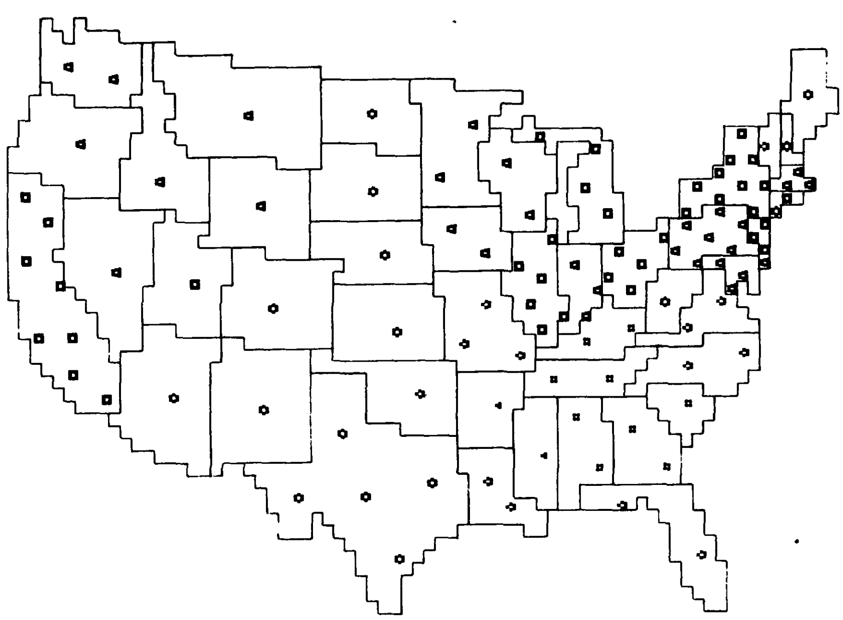


Graph 10 Italy (1964) Traffic in main airports *Source of data* Annuario statistico italiano, 1965 Roma, Istituto Centrale di Statistica, p 296 *Scale* 10,000 passengers arriving = 1 unit in square GRP 100 tons of freight unloaded = 1 unit in circular GRP

In Graph 12 a map is shown representing the same data as Graphs 2 and 3. Here median income in each State i is given by a pattern proportional to it. The number of patterns in each i is proportional to its population of families Pi. In order to keep in line with symbols previously used, let us forget that the map shows median income and suppose it indicates arithmetic average of income per family. Then averages can be indicated by $\overline{X}i = \frac{Xi}{Pi}$, where Xi is total income. The visual impression given by each State is proportional to the product

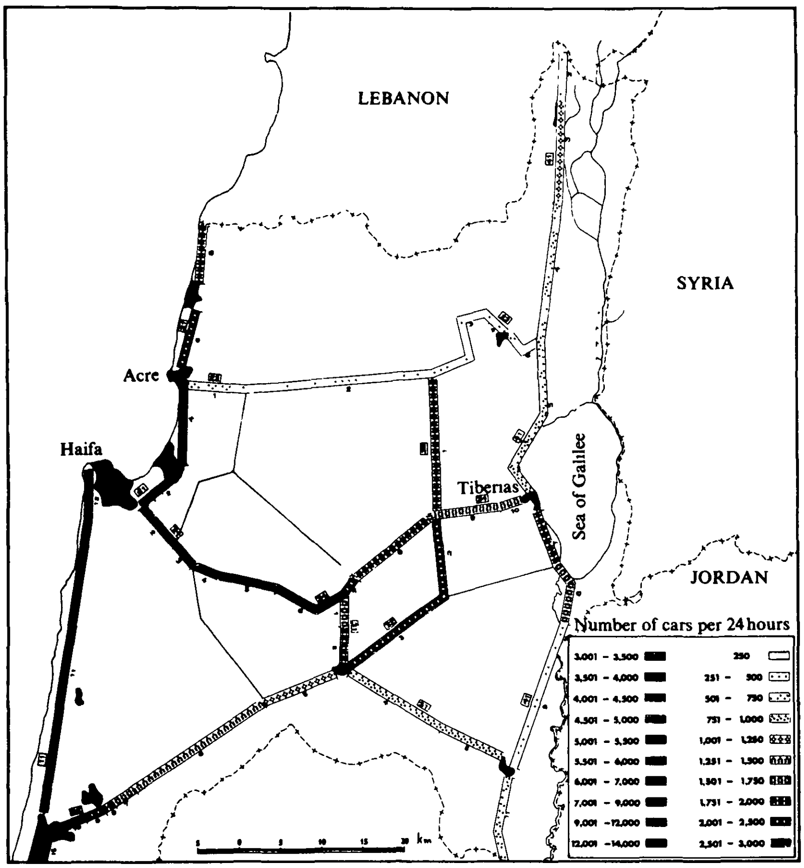


Graph 11
U S A (1840, 1900, 1960) Percent rural population.



Graph 12: USA (1950) Median income per family

Patterns (X_1) 1200-1600 1600-2000 2000-2400 2400-2800 2800-3200 3200-3600 between $\overline{X}i$ and the size of population Pi, viz., to $\overline{X}iPi$ and thus is solved into XiPi = Xi. Then properties similar to those indicated under (a), (b), (c) for Graph 1 are found.



Graph 13

Israel: Northern part of the country (April 1962–March 1963). Traffic density on non-urban roads.
Source of data Sample traffic counts performed by the Central Bureau of Statistics.

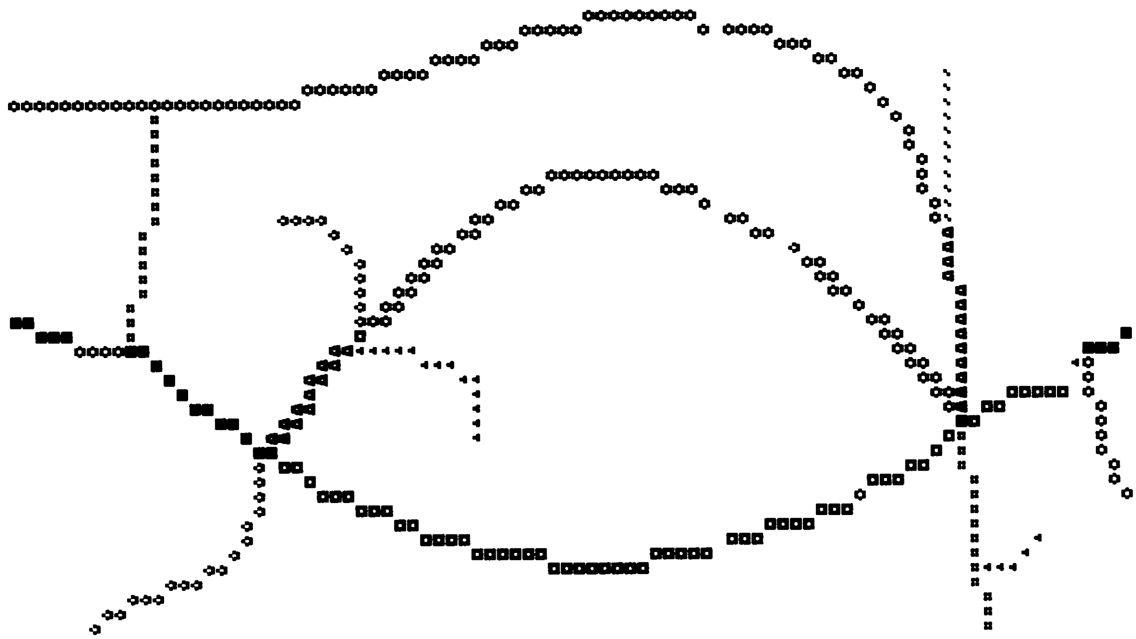
While all other graphs presented before are hand-drawn, Graphs 12 and 3 are computerized.

Graphs 13 and 14 show applications of repeated GRP to represent intensity of traffic; they show respectively intensity of traffic on part of a country (Northern Israel) and part of a town (Jerusalem). Graph 14 is computerized. GRP presentation seems advantageous over the usual presentation of intensity of traffic by bands of differing width which are not always easy to be interpreted (see an example in Graph 15).

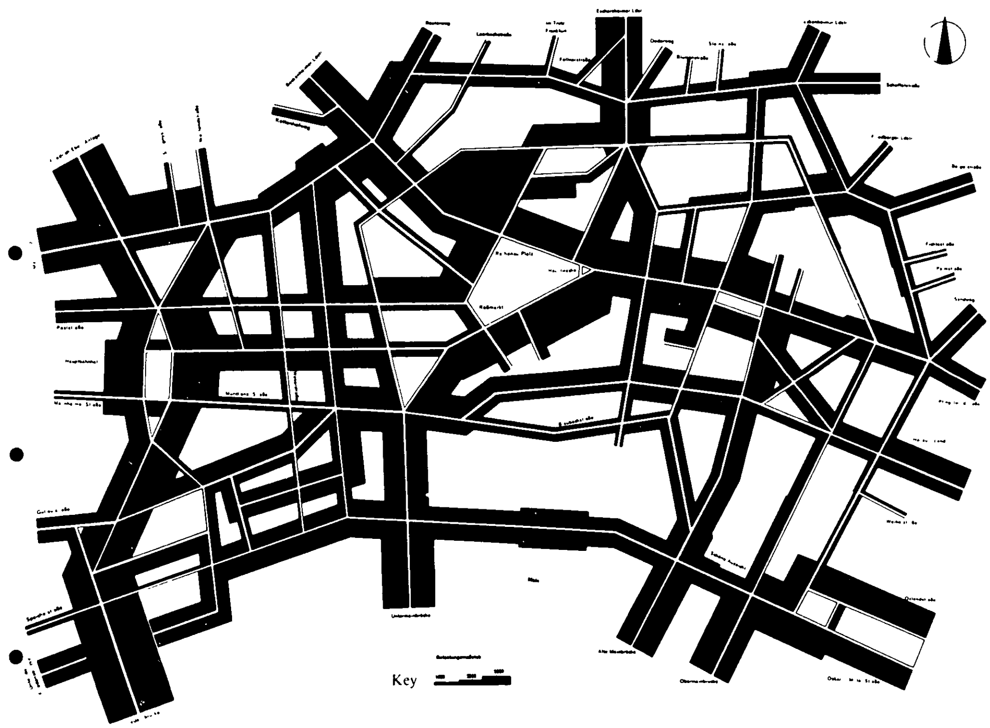
The above examples seem to suggest some cautious optimism with regard to the feasibility of improvement of graphical methods and indicate that further research work in this direction may be worthwhile.

NEED FOR ORGANIZED EFFORTS FOR IMPROVING GRAPHICAL METHODS

Another reason for cautious optimism for the future can be found in the fact that some awareness for the need of systematic efforts in the graphical statistical field seems to be felt today by various national and international institutions.



Graph 14



Graph 15

Frankfurt am Main (1961) Traffic in part of the town

Graph reproduced by permission of the Municipality of Frankfurt from Gesamt Verkehrs Planung Frankfurt am Main, 1961 (Magistrat der Stadt Frankfurt am Main 1961).

A very clear and encouraging example of this is given by the stand taken recently in this field by the U.S. Bureau of the Census.

Therefore, it is perhaps not utopic to think that in the course of time the challenge put on graphical statistical methodology by the developments due to automation may be successfully met. It is to be hoped that in the years to come the entire field may be reappraised.

In such reappraisal, the following aims seem to be of particular importance:

(a) Criteria to which scientifically dependable graphs are to conform must be investigated and restated clearly. (b) Properties, limitations, and pitfalls of each of the existing graphical systems should be thoroughly reinvestigated. (c) Also the problem of building rational scales for the graph, apt to ensure good correspondence between the data represented and the graphical symbols used, needs deep rethinking and finding of general solutions valid for all types of graphs. (d) Systematic exploration of future needs should be performed in order to encourage development of new graphical methods capable of standing up to scientific requirements.

This reappraisal is a major enterprise in which statisticians, geographers, cartographers, psychologists, and people dealing with information theory and methods should cooperate. To reach it, better communication should be established too between theoreticians, technicians, and builders of hardware and software, and between producers and users of statistics and graphs.

REFERENCES

1. Provisional copies of the opening paper of this meeting (R. Bachi, "Graphical Methods: Achievements and Challenges for the Future") were distributed at the International Symposium on Computer-Assisted Cartography. The final report on the meeting will appear in the Bulletin of the International Statistical Institute.

2. See: D. Huff, How to Lie With Statistics. New York: Norton and Co., 1954.

3. To simplify the discussion, we disregard here the fact that the data refer to the median income and not to the arithmetic average of incomes.

4. See: Journal of the American Statistical Association, Vol. 68, No. 342, p. 361.

5. J. Bertin, Semiologie graphique. Paris: Mouton & Gauthier Villars, 1967.

6. R. Bachi, Graphical Rational Patterns:A New Approach to Graphical Presentation of Statistics. Jerusalem: Israel Universities Press, 1968.