DEMOGRAPHIC DATA FOR LOCAL AREAS

HARLEY B. MESSINGER
INSTITUTE FOR HEALTH RESEARCH, BERKELEY

1. Data sources

1.1. *Census.* The 1970 Census offers much better access to local data than previous decennial censuses. Examples are "block-face" coding and the preparation of census tract directories. The latter permit coding street addresses to census tracts and were never before available for so many cities at one time, the former would permit obtaining data from housing along designated streets with several levels of auto traffic density for a study of air pollution effects on people living at the several levels.

1.2. *Local records.* Vital statistics records have long been popular for epidemiologic studies but recently the value of other routinely collected local items has become appreciated. Santa Clara County in the San Francisco Bay Area has linked files for different departments by coding census tract onto many county records.

1.3. *Surveys.* While much can be done with data gathered for other purposes, an epidemiologic study will usually have to collect information directly as well. Having an interviewer survey each individual household no longer is considered necessary to get valid data. Hochstim's Cervical Cytology Study in Alameda County showed how to increase efficiency of data collection by use of the telephone or the mails [2].

2. Data handling

2.1. *Data banking.* Larger studies may require huge data files meriting the label of "data bank." At one extreme, too little attention is paid to planning for this aspect and the study founders in a mass of doubtful data; at the other, the gathering of information becomes an end in itself and more is collected than can possibly be analyzed with funds available. With the growing realization that protection of confidentiality requires positive measures, even more care will be required to keep data processing budgets within bounds. In the "link file" system used by the American Council on Education for a longitudinal study [1] the identity of a questionnaire respondent is kept secret by having the file linking him to his particular set of answers stored in a foreign country.

2.2. *Record linkage.* Besides the aggregation of statistics by area, the linkage of records for individuals is also important, for example, births and early deaths. An early example of this was the British Columbia Population Study based on records of 114,000 marriages in that province over a ten-year period. Births resulting from these unions were linked with early deaths and also with records of handicapped children in a separate registry [4].

2.3. *Mapping.* Data for small areas may be displayed by mapping as an alternative to tabling. Patterns in census tract distributions of a disease, say, might resemble those of certain air pollutants. In the case of many chronic diseases, such a spatial correlation might not appear until age adjustment of the disease rates had been made. Trend-surface mapping [3], an application of the general linear model to map analysis, offers an approach to systematic use of this technique.

2.4. *Restructuring analytical units.* In census tracting an area, one tries to divide up larger political units such as counties into smaller homogeneous pieces using judgment to set boundaries. Even though this objective can be reasonably well achieved at one point in time, the passage of the years tends to remove any advantages of judiciously chosen partitions over randomly chosen ones. Despite this, census tracts remain more homogeneous than larger units simply because of their size and the fact that demographic characteristics show strong spatial correlation. Thus there is the opportunity to group similar census tracts into "pseudo-counties" as an alternative to the existing political units which may vary markedly in size and may be heterogeneous. An air pollution study in California would not be likely to use Los Angeles County as one unit nor would the census tract be the answer if reliable incidence figures for most chronic diseases were needed. Clustering of contiguous tracts to maximize within-cluster homogeneity would give a better areal unit.

3. Data problems

3.1. *Population mobility.* The high residential mobility of urban populations in the United States is well known. Much of the movement, however, is within the boundaries of political units such as counties or, at worst, standard metropolitan areas in the case of the larger urban agglomerations. Thus a study with coverage of a large city may be able to keep track of a population sample over a period of a year or two provided they have planned for the staff needed to do the tracing.

3.2. *Updating the census.* In an ongoing study based on census data, thought should also be given to detecting major changes in the physical characteristics of the local areas involved. A freeway, from the moment when it's first proposed can have a dramatic effect on the local units in its path. How could such complex changes be comprehended without costly periodic community surveys? And that is only one influence! If one tried as well to keep up with the less dramatic changes, the total effort would dominate all but the largest project. A good case can be made for replacing decennial censuses with continual sampling surveys even if it takes a constitutional amendment!

3.3. *Ecological correlations.* Criticisms have been raised as to the validity of correlations based on aggregated data and the tendency has been noted for their magnitude to increase as one moves up to higher levels as from census tracts to counties to states. They may even change sign as in Buechley's example of the negative correlation of income with cirrhosis of the liver mortality at the census tract level changing to a positive one at the state level [5]. Within a city, it's the "poor" tracts where the disease is common but it's the "rich" states that have the large urban centers with the "poor" tracts inside them. Both correlations are valid but it is easy to try to interpret these associations as if they were based on data for individuals.

4. Analytical methods

4.1. *Factor analysis.* Factor analysis provides an objective way to find patterns of association in large sets of demographic variables. Most studies have employed only data from the decennial censuses. To these demographic measures can be added ones relating specifically to health [6], [7] such as the liver cirrhosis mortality rates already mentioned. The association with economic status seen in states of the United States, incidentally, is not found in prefectures of Japan, cirrhosis in the latter country appearing in a different factor.

4.2. *Cluster analysis.* Just as multiple regression is much easier to interpret if the predictor variables are really stochastically independent, so clustering of areal units on the basis of a few orthogonal factor scores facilitates description of a cluster of census tracts. The late Professor Tryon pioneered in this technique not only in his primary field of psychology but also in the analysis of census tract data. For instance, comparisons of census tract "types" in San Francisco before and after World War II may be found in his book [8].

REFERENCES

[1] A. W. Astin and R. F. Bourch, "A 'link' system for assuring confidentiality of research data in longitudinal studies," *Amer. Educ. Res. J.*, Vol. 7 (1970), pp. 615–624.

[2] J. R. Hochstim, "A critical comparison of three strategies of collecting data from households," *J. Amer. Statist. Assoc.*, Vol. 62 (1967), pp. 976–989.

[3] W. C. Krumbein and F. A. Graybill, *An Introduction to Statistical Models in Geology*, New York, McGraw-Hill, 1965. (See Chapter 13.)

[4] H. B. Newcombe, "Population genetics: population records," *Methodology in Human Genetics* (edited by W. J. Burdette), San Francisco, Holden-Day, pp. 92–113.

[5] A. Pearl, R. W. Buechley, and W. R. Lipscomb, "Cirrhosis mortality in three large cities: implications for alcoholism and intercity comparisons," *Society, Culture, and Drinking Patterns* (edited by D. J. Pittman and C. R. Snyder), New York, Wiley, 1962. (See Chapter 19.)

[6] E. S. Rogers and H. B. Messinger, "Human ecology: toward a holistic method," *Milbank Fund Quart.*, Vol. 45 (1967), pp. 25–42.

[7] E. S. Rogers, M. Yamamoto, and H. B. Messinger, "Ecological associations of mortality in Japan and the United States: a factor analytic study," *Population Problems in the Pacific* (edited by M. Tachi and M. Muramatsu), Tokyo, 1971. (See Chapter 24.)

[8] R. C. Tryon and D. E. Bailey, *Cluster Analysis,* New York, McGraw-Hill, 1970. (See references to the "Social Area" problem.)

Discussion

Question: John R. Goldsmith, Environmental Epidemiology, California Department of Public Health

The Current Population Survey and National Health Survey collect both demographic and health data for intercensus years. Data on a sample basis is valid for the country as a whole, but not for separate states, counties, or cities. With some additional resources, such data could be obtained. We obtained it for the Health Survey data for California from 1954 to 1960 and found this useful. The California Legislature is also considering support for a demographic service.

In response to the question as to whether census data are only of value for "fishing expeditions," my opinion is that such data are of great value in testing specific hypotheses. We have used them in health survey studies, in the two community mortality study, and in testing for the effect of demographic variables on other effects.

Reply: H. Messinger

We had thought about using National Health Survey data for a study with states as areal units but could not because comparable items would have been required for at least the conterminous states. The NHS questionnaire is a useful model, in any case, and the "five-foot shelf" of reports issued so far can serve as benchmarks for local surveys.

The term "fishing expedition" should no longer be one of opprobrium. In some areas, such as genetics, most experimental problems can be formulated as clear-cut tests of Hypothesis A against Hypothesis B. In other areas, especially those involving the behavior of humans as individuals and as groups, problems are not as easy to structure. At one stage, census data may be used as an aid in formulating hypotheses; at another, in testing hypotheses as Dr. Goldsmith points out.

Question: J. Neyman, Statistical Laboratory, University of California, Berkeley

Is it practicable to obtain local demographic data indicating, for example, that a given locality, originally comparable, is in the process of turning into a slum? (Or vice versa?)

Reply: H. Messinger

As of the date of a census, it's possible to identify a group of tracts as a slum (having defined this term) from the population and housing data collected, by computing various variables like family income, persons per room, percentage houses dilapidated, percentage unemployed, and also health data like mortality and morbidity rates. I would do this by methods mentioned in Section 4 of my remarks, but other ways could be devised. Then the problem is reduced to the one already touched on of updating the census. With change occurring at present rates, looking at the situation every ten years is inadequate if we are expecting to achieve any control over the process of urban decay.

Ways of extending the analytical methods of Section 4 to encompass change may be found in Harris [1], especially Tucker's article on three mode factor analysis [3]. A well documented study focusing on urban change is Murdie's study of Metropolitan Toronto from 1951 to 1961 [2].

REFERENCES

[1] C. W. Harris (editor), *Problems in Measuring Change*, Madison, University of Wisconsin Press, 1963. (See Chapters 6 to 10.)

[2] R. A. Murdie, "Factorial ecology of Metropolitan Toronto, 1951–1961," University of Chicago, Department of Geography Research Paper No. 116, 1969.

[3] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change," Chapter 7 of Harris, *Op. cit.*