NATIONAL BUREAU OF STANDARDS REPORT

10 653

EFFECT OF DATA AGGREGATION IN MODELLING

U.S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards [1] was established by an act of Congress March 3, 1901. Today, in addition to serving as the Nation's central measurement laboratory, the Bureau is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. To this end the Bureau conducts research and provides central national services in four broad program areas. These are: (1) basic measurements and standards, (2) materials measurements and standards, (3) technological measurements and standards, and (4) transfer of technology.

The Bureau comprises the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Center for Radiation Research, the Center for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of an Office of Measurement Services and the following technical divisions:

Applied Mathematics—Electricity—Metrology—Mechanics—Heat—Atomic and Molecular Physics—Radio Physics [2]—Radio Engineering [2]—Time and Frequency [2]—Astrophysics [2]—Cryogenics.[2]

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; develops, produces, and distributes standard reference materials; relates the physical and chemical properties of materials to their behavior and their interaction with their environments; and provides advisory and research services to other Government agencies. The Institute consists of an Office of Standard Reference Materials and the following divisions:

Analytical Chemistry—Polymers—Metallurgy—Inorganic Materials—Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations in the development of technological standards, and test methodologies; and provides advisory and research services for Federal, state, and local government agencies. The Institute consists of the following technical divisions and offices:

Engineering Standards—Weights and Measures — Invention and Innovation — Vehicle Systems Research—Product Evaluation—Building Research—Instrument Shops—Measurement Engineering—Electronic Technology—Technical Analysis.

THE CENTER FOR RADIATION RESEARCH engages in research, measurement, and application of radiation to the solution of Bureau mission problems and the problems of other agencies and institutions. The Center consists of the following divisions:

Reactor Radiation—Linac Radiation—Nuclear Radiation—Applied Radiation.

THE CENTER FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in the selection, acquisition, and effective use of automatic data processing equipment; and serves as the principal focus for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Center consists of the following offices and divisions:

Information Processing Standards—Computer Information — Computer Services — Systems Development—Information Processing Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System, and provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data—Clearinghouse for Federal Scientific and Technical Information [3]—Office of Technical Information and Publications—Library—Office of Public Information—Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Located at Boulder, Colorado 80302.
[3] Located at 5285 Port Royal Road, Springfield, Virginia 22151.

NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT		NBS REPORT
4314518	December 7, 1971	10 653

EFFECT OF DATA AGGREGATION IN MODELLING

Milestone 3

by
Lambert S. Joel
Applied Mathematics Division
Lead Paint Poisoning Project
Building Research Division
Institute for Applied Technology
National Bureau of Standards
Washington, D. C. 20234

Not for publication
or for reference

Sponsored by

Department of Housing and Urban Development

U.S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

ABSTRACT

The report describes, with examples, some of the effects upon model formulation and model applicability, of differences in the aggregation levels among various data sources, (primarily populations).

EFFECTS OF DATA AGGREGATION IN MODELLING

Three of the most common problems related to aggregation of data sources are 1) possible nonlinearity of the relationship between dependent and independent variables, 2) nonlinearities in the underlying relationships between populations and the characteristics which are to be used as the independent variables, and 3) homogeneities induced by grouping differing populations together.

The effect of nonlinearities in the relationship between dependent and independent variables, is that a model constructed at a given aggregation level cannot be applied at other levels. The effect of nonlinearities in the underlying relationships is that the relationships between the "dependent" and "independent" variables will be elusive, and the effect of homogeneities is that identifying differences in the "independent" variables may be suppressed entirely, making model construction impossible.

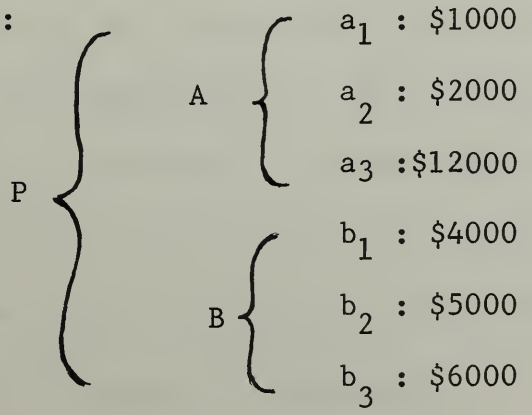
A simple example of nonlinearity will serve to illustrate the first problem. Suppose that the number of telephone calls (T) per day in a city is quadratically related to its population (P). (This is more reasonable than to assume that the relationship is linear, i.e., proportional to the population, because the number of pairs of people increases faster than the number of people, and a phone call is a pair phenomenon.) We would then have $T = CP^2$ where C is a scaling constant. Now if a pair of neighboring cities are considered as a single aggregated population according to some criterion, e.g., Minneapolis - St. Paul, San Francisco - Oakland, Philadelphia - Camden, etc. then

applying this model we would have $T = C(P_1 + P_2)^2 = C(P_1^2 + P_2^2) + 2C(P_1P_2)$ a number which could be considerably larger than $T = C(P_1^2 + P_2^2)$, calculated as a sum of the values for the pair of locations separately. If the relationship were <u>linear</u>, i.e., $T = CP$, then areas could be aggregated without penalty. This means that a nonlinear model constructed at any given level cannot be applied without carefully ascertaining that the regions to which it is applied really are equivalent demographically to those from which the model was deduced.

Problems 2 and 3 affect the character of a model that may be evolved from a data base, and in fact overlap, so that they can be discussed together.

In general, the perceived character of an observed population is strongly affected by the way the boundaries for the observations are delineated. Sample data for use in constructing models should be gathered therefore, with some circumspection.

Median values and rankings are parameters whose magnitudes depend on the level of aggregation of described populations. For a simple example consider a population P of 6 people a_1, a_2, a_3, b_1, b_2, b_3 with incomes given below and grouped as shown into two subgroups (A & B) with three people each:



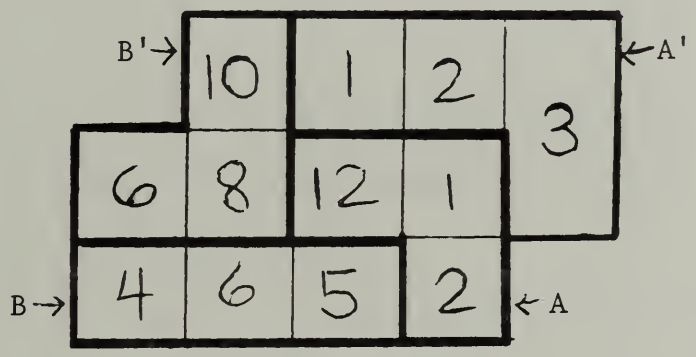
The average incomes of P, A and B, are all $5000. The median income of A is $2000, of B, $5000 and of P, $4500, i.e., the median income of the aggregated subgroups is not related to their individual medians. As for the effect of rankings, if we stratify incomes by thousands of dollars, the fractions of the populations of A and B in the lowest stratum are both 1/3 (1/3 of the people in B are in the lowest stratum for that subgroup: $4000), whereas the fraction of the population of P in the lowest stratum is 1/6. For this kind of parameter even a linear model will fail under aggregation because the parameters themselves are nonlinearly related to the populations which they describe.

Phenomena depending on concentration level, are frequently encountered, typically in epidemiological studies. A simple example here is given by two populations of families with the same total number of young children, in which the children are distributed one per family in the first population, and three in every third family in the second. A survey, for instance, of the consumption of children's clothing would probably find higher per capita consumption in the first population than in the second because of the use of hand-me-downs in the second group, even though the mean statistics of the two populations are otherwise identical. (To make the illustration cogent we can imagine that all families have the same income.)

The word aggregation in the context of a statistical study or construction of a mathematical model refers to the way the "subjects" of the study or model, are grouped in the analysis or subsequent application. The degree to which subjects, "populations" or observations are combined is called the level of aggregation.

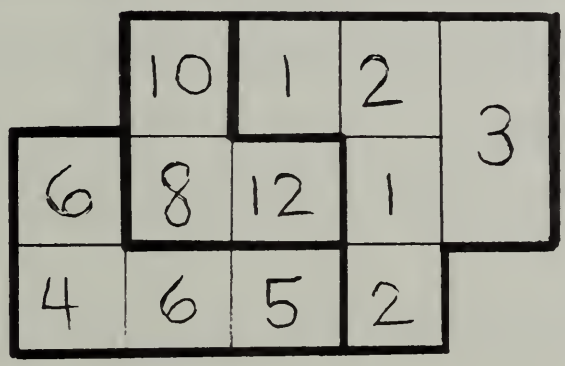
As an example, census grouping by county, state, or region defines three levels of aggregation. Similarly stratifying population by family incomes in $5000/annum steps or $10,000/annum steps defines two levels of aggregation. On the other hand, grouping the population into urban and rural classes is a non hierarchical or non stratified kind of aggregation. In this report we shall be concerned with stratified aggregations, that is with levels of aggregation.

We can see how aggregation can disguise the character of population by looking again at our first example. Let us flesh it out a little with some definitions and some geography. We will say that incomes under $5,000 are "low", from $5,000 to $9,999, "middle" incomes, and over $10,000, "high". We will also locate our population on a little map.



Now recall that the mean income of what we will now call "area" A is $5,000, the median income $2,000. Thus by one measure A is a low income area, by another, a middle income area. In fact, it can be split into a low and a high income population, and the aggregation even at this low level distorts it. Again, with regard to the distribution of incomes rather than measures like means and medians, we have added a second "area" P'. At the level of aggregation in which we compise the sub sets, P and P' look quite alike in terms of income: (1, 2, 4, 5, 6, 12) and (1, 2, 3,

6, 8, 10). Whereas in fact P is comprised of a middle income area B and a mixed high-low area A, while P' comprises the low income A' and a sort of upper middle area B'. Notice, moreover that at this aggregation the analyst looking at the "census figure" level would deduce a slightly different image of the total area if the boundary lines of the samples were the ones shown on the second version of the "map".



DATA SOURCES

Unforturately, social researchers, unlike physical scientists usually are not able to acquire data for construction and validating hypotheses by conducting controllable experiments.

Perhaps this constitutes an overstatement of the difficulty. With sufficient resources, (in time and tangibles), a data collection program specifically designed for the purpose of testing a hypothesis concerning some human phenomenon is almost always possible. "Sufficient," however, frequently means "very great". Consequently one usually turns to existing tabulated data sources, chiefly the various compilations of statistics by the U.S. Bureau of the Census. In the construction of a first stage mathematical model of incidence of elevated blood levels (EBL) in young children resulting from ingestion of lead paint, which is the subject of

this report, the effect of inconvenient sample boundaries has been a particularly vexatious problem.

The details of the model are described in a report entitled "A Model to Estimate the Incidence of Lead Paint Poisoning". It suffices here to state that the hypothesis underlying the model is that the rate of incidence of EBL in an area can be formally related to a) the fraction of aged and deteriorating housing, which determines the level of the presence of lead paint, b) representative socio-economic status, which is a determinant of this environmental hazard, and c) the degree of accessibility to contaminated paint by the "population at risk": children up to six years of age.

In the absence of an independent data collection effort for lead poisoning estimation, it is natural to inquire how well the Decennial Census can serve as a data pool for the independent, i.e., the determining variables, of a model with the underlying ideas as stated. Let us examine the structure of census compilation.

The basic unit in compilation of census population data is the block group in urban localities and the enumeration district elsewhere. Block groups are, as the name implies, collections of "bunched" city blocks, with aggregate population of about 1000. Enumeration districts are roughly equivalent rural areas. Block groups (or sometimes enumeration districts) are aggregated into clusters whose populations should be close to 4000 according to Bureau of Census demographers. These clusters are called census tracts. Census tracts are the smallest subdivisions for which computer tapes are available at present (November 1971) containing the first count population and housing tabulations of the 1970 census.

Tract boundaries were drawn originally in 8 cities for the 1910 decennial census, and the number of tracted areas has increased progressively since then. The early tracts were ethnically homogeneous, probably because of the widespread interest in the image of the United States as a "melting pot" of races and nationalities. Today responsibility for delineations is vested in local tract commissions, assisted but not bound by Bureau of the Census guidelines on population size and ethnic and socio-economic homogeneity[1], with the result that the more than 50,000 current tracts in the U.S. are haphazard in character with populations ranging from 1500 to 12000, and degrees of homogeneity largely determined by the professional compositions of the (autonomous) tract commissions.

Data requirements and other costs, i.e. the difficulties of handling 50,000 tracts rather than 250 metropolitan areas, discourage using a census tract model for a first cut estimation of the national incidence of "pediatric EBL".

On the other hand, incidence data have not been available up to this time for the purpose of calibrating a model at a higher level of aggregation (i.e., county, city or SMSA). This remark requires some amplification. Data have been gathered in some 20 cities, but in addition to inherent problems in compilation of information, we are confronted with disparate surveys without common sampling methods, common definitions of EBL, or

---

[1] Census Tract Manual, Fifth Edition, U.S. Department of Commerce, Bureau of the Census, Jan. 1966.

common boundary criteria. (Some of the regions surveyed extended beyond city lines without, however, encompassing SMSA's.)

The Lead Poisoning Prevention and Control program in New Haven, Connecticut included a survey of the city, in which incidences were tabulated by census tract. This data base has been used in the construction of a preliminary model.

Insofar as pathologies induced by the level of aggregation are concerned, the results to date have been mixed.

There is unquestionably some "homogenization" difficulty. A study reported in the Connecticut Health Bulletin[2] demonstrates that the socio-economic character of neighborhoods in New Haven, as revealed by census block group statistics, is submerged by the aggregation of the groups into census tracts. We are certain that this has been a strong (but not the only) factor in our inability to fit a model to the New Haven data at the level of precision we desire. (This report is being produced prior to attempts to validate any of our models by applying them to data from another city, Aurora, Illinois, from which survey data, by tract, have been promised to us.) On the other hand, the values predicted by these

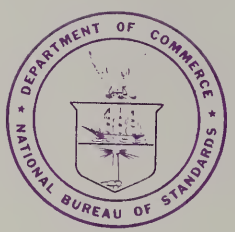
---

[2] E. Siker, J. Deshaies, S. Korper, and E. Stockwell, "Development of a Community Wide Health Information System--Neighborhood Delineations Socio-Economic Status," Connecticut Health Bulletin Vol. 84, No. 9, Sept. 1970.

models based on data aggregated over the entire city, are no worse than the tract predictions. That is, although the models are nonlinear, the initial tests don't discourage the hope that one of them, if proven valid over tracts, may be applicable to prediction at the city or SMSA level.

Homogenizations resulting from aggregating populations may actually improve the validity of model formulation in situations where there are "compensating errors" in available data. A case in point is the use of age of housing as a determining parameter for incidence of EBL. It is often assumed that a sharp (and continuing) decline in the use of lead based paint in housing interiors, in favor of titanium based paint, commenced in 1940. Thus, the age of housing is a natural kind of parameter to employ as a determinant of lead contamination. In particular, the fraction of housing units erected before 1940, which is tabulated by the census, is such a parameter. It could not be employed directly in preliminary modelling attempts because 1970 age of housing tabulations are part of the "4th count" census figures and will not be issued until March 1972. Relating 1960 housing data to current EBL incidences cannot be done reliably because of widespread "urban renewal" activities which normally eventuate in wholesale demolitions.

However, if we must fall back on out of date statistics, use of 1960 age of housing figures, at the SMSA level, (with a crudely estimated uniform attrition factor for old houses,) would be vastly superior to attempting a tract or fine subdivision model employing 1960 housing data, in which errors of large magnitude in estimates of the age of dwelling units are to be expected.


DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS