DOCUMENT RESUME

ED 344 920 TM 018 252

AUTHOR Jang, Younghee; Radio, Joni L.
TITLE Problems in Merging District- and Community-Based Data Sets.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Apr 92
NOTE 15p.; Paper presented at the Annual Meeting of tne American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Census Figures; *Community Characteristics; *Data Analysis; Databases; Elementary Secondary Education; *Evaluation Methods; Geography; Information Utilization; Management Information Systems; National Surveys; Research Methodology; *Research Problems; *School Districts
IDENTIFIERS Census Mapping Project; Data Sets; Educational Information; *Information Integration; TIGER Database File

ABSTRACT

Integrating information from local school district and community data sources is essential to understanding the relationships between them. A major problem in merging such data concerns the geographic incongruities of the boundaries of school districts (local education agencies) and the boundaries of communities. This paper focuses on resolution of this problem at the national and state levels. The recent attempt of the Southwest Regional Laboratory to resolve this problem through the Census Mapping project is described. It is evident that the near-optimal solution lies in a complete blocking of the United States and the identification of all the blocks in every school district to allow an accurate aggregation of Census data pertaining to each local education agency. The complexity of this effort has led the Census Bureau to develop and implement a digital cartographic database, the Typological Integrated Geographic Encoding and Referencing (TIGER) system. School district information from the states is being included. A lot of staff power is required to construct the database, and the district equivalency file will not be available until 1993. Great progress will be made if Congress passes a Uniform Data Act that requires a universal format for data from states receiving federal funds. Better forecasting and planning and more equitable distribution of funds will result when more timely and accurate data are available. There is a 14-item list of references. (SLD)

Problems in Merging District-And Community-based Data Sets

Younghee Jang and Joni L. Radio
Southwest Regional Laboratory
AERA Annual Meeting, San Francisco, CA

April 1992

This paper was developed under a subcontract from the Far West Laboratory operating under a contract it holds with the U.S. Department of Education's Office of Educational Research and Improvement. However, the contents do not necessarily represent the policy of Far West Lab nor of the Department of Education, and endorsements of the contents by either should not be assumed.

Introduction

Demographic, social, economic, and housing characteristics of the community (city) influence the funding, policies and priorities of a school district, otherwise known as a local education agency (LEA). Therefore, without an understanding of the relationship between communities and LEAs, determining significant indicators of success is difficult. The LEA sets and carries out program priorities and maintains longitudinal achievement data. The U.S. Census provides data about socioeconomic factors of the surrounding communities. Integrating information from these sources is crucial to researchers who want to use such district and community data, as well as to Southwest Regional Laboratory's (SWRL) Metropolitan Educational Trends and Research Outcomes (METRO) Center.

SWRL's METRO Center addresses schooling problems of educationally disadvantaged children in the Western's region's metropolitan areas. One of the studies in the METRO Center is the Successful Indicators Study (SIS). The goal of SIS is to develop indicators within a school district and community that result in a positive climate for improving the achievement level of the Western region's educationally disadvantaged children.

As a first step to integrating the two data sets, SWRL identified the boundaries of both the LEAs and communities, and explored any overlap that exists. Because most of the LEAs are not coterminous with any of the area organization units the Census Bureau used, the Census data cannot be applied directly to LEAs.

The data for the SIS study come from many sources. The three primary sources are the 1980 Census, the 1989-90 California Basic Education Data System (CBEDS), and the Public School Directories of the four states being studied (i.e., Arizona, California, Nevada, and Utah). SWRL also used the metropolitan statistical area (MSA) maps the Department of Commerce published to identify the metropolitan areas for the study. All of the communities within each MSA were found by using detailed maps. The public school directories for each of the states were used to identify the LEAs followed by a detailed mapping between the communities and the LEAs.

In some instances, the communities and LEAs match appropriately. In other cases, a LEA consists of several communities. Neither of these cases leads to any significant problems as the Census Bureau provides aggregate data for communities with populations over 10,000. However, because the Census data for a given community are broad, problems do arise when a community is served by many school districts. This is quite common in Arizona and California, with 62% and

42% in this category respectively, because of the immense population growth around cities and their surrounding areas in these two states in the past several decades. As an example, California's total K-12 enrollment is at least four times greater than the total K-12 enrollment for Arizona, Nevada, and Utah combined, and Arizona's K-12 population is more than that of Nevada and Utah combined. However, this overlap between community and school districts is not significant for Nevada because of its system of countywide LEAs. Unless the Census data of the localities within the same community are similar, the attribution of the community characteristics to the LEAs is problematic.

The major problem with merging these types of data is the geographic incongruities of the boundaries of LEAs and the boundaries of communities. Hence, a primary methodological concern is the resolution of the noncoterminous nature of some of the LEAs and communities. This paper focuses on the resolution of this problem at the national and state level. First, brief historical information is provided to orient the readers on past attempts at resolving this problem. Next is a description of the geographic organization used by the Census Bureau and concerns that exist about this method of organization. This section is intended to help the reader place SWRL's work on the Census Mapping Project into context. SWRL's recent attempt at resolving this methodological problem via the Census Mapping Project is described in detail. Finally, the paper ends with a presentation of the importance of timely and accurate databases, and needed developments.

Historical Information

In 1970, the National Center for Educational Statistics (NCES) contracted with the Census Bureau for the development of a standard set of maps showing the boundaries of LEA with 300 or more students. The 1970 Census geographic units were allocated to each of the mapped districts and resulted in the School District Geographic Reference File. Census data were restructured to be applicable to each district. A number of subsequent studies and NCES reports were based on these data.

A 1978 congressional mandate led to the 1982 Census Mapping Project the Council of Chief State School Officers (CCSSO) coordinated. States provided the maps with school district boundaries. For the first time, boundaries of the nation's 16,038 existing school districts were mapped.[1]

The importance and usefulness of these national efforts resulted in the Stafford-Hawkins Act of 1988, which specifically requires NCES to submit a report to Congress on a decennial basis.

> On April 1, 1993, and every 10 years thereafter, the Center shall submit a report to the appropriate committees of the Congress concerning the social economic status of children who reside in the areas served by different local education agencies. Such report shall be based on data collected during the most recent decennial Census.

SWRL carried out the mapping project (Census Mapping Project Guideline, 1990) in California for NCES in preparing the 1993 report. This effort is described below. In the next section, a description of the Census geographical organization is given to facilitate an understanding of the source of the problem working with Census data.

Geographic Organization

For the SIS project, the LEAs of interest include those of the four states' major cities and "edge cities" (Tushnet, 1992). These are found as part of an MSA or contiguous to an MSA. To illuminate the problem and SWRL's proposed solution, a brief outline of the geographic organization used by the Census Bureau is presented.

The U.S. Office of Management and Budget (OMB) first defined the concept of Standard Metropolitan Statistical Areas (SMSA) in 1949 to be used in its Census publications. It represents an area with "a large population nucleus together with adjacent communities that have a high degree of integration with the nucleus" (p. 20, Frey & Speare, 1988). For the entire United States (except New England), SMSAs have been defined in terms of counties or county equivalents. The longitudinal nature of decennial data is useful for comparative purposes only if some stability is ensured. Of all the geographic entities, the county boundaries seldom change, and it also is often the smallest geographical unit for which many types of data are tabulated.

---

[1] For more details, refer to the references on the 1970 and 1982 NCES files on school districts and their user's guide. In particular, the Technical Documentation on the Census of Population and Housing, 1980: Summary Tape File 1F, School Districts (STF-1980) provides a useful collection of relevant information.

In the fifties, as the country became increasingly urbanized around the major cities in several parts of the country, it became difficult to determine the boundaries of metropolitan areas when they merge into one contiguous region as in the case of New York and Northeastern New Jersey, and Chicago and Northwestern Indiana. In the 1960 Census, the concept of Standard Consolidated Area (SCA) was introduced to provide a larger aggregate unit than MSA to address the situation of adjacent MSAs that were closely integrated. There were two SCAs in 1960. In 1975, SCA was renamed as Statistical Consolidated Statistical Area (SCSA) when definite criteria of size and integration were established. In 1980, there were 16 SCSAs comprised of 48 SMAs with at least a million people each.

In 1983, OMB revised the definitions of SMSAs and renamed these areas as MSAs. A SMSA with over 1 million population with two or more counties was divided into two or more Primary Metropolitan Statistical Areas (PSMA) if local criteria support such subdivisions with the former SMSA known as a Consolidated Metropolitan Statistical Area (CMSA). The 1983 revision resulted in 253 MSAs and 19 CMSA comprised of 60 PMSAs.[2]

The Census geographical organization of the entire country for data information and data summary is given in the following hierarchy, with the 1983 changes incorporated.

- States or State equivalent
  - Consolidated Metropolitan Statistical Area (CMSA)
    - Metropolitan Statistical Area (MSA or PMSA)
      - Remainder of State (non-MSA)
        - County (County segment in New England)
          - Minor Civil Division (MCD—present in only 20 States)
            - Remainder of MCD or remainder of county
              - Tract (BNA)
                - Block Group (BG)
                  - Block (ED)

Census provides summary data at each level of the hierarchy, with a block representing the smallest unit of information.[3] For the purpose of this paper, part of the explanation on a block, a block group, block numbering area, and tract are presented below.

Block—Normally a rectangular piece of land, bounded by four streets. However, a block may also be irregular in shape or bounded by railroad tracks, streams, or

[2] For further details, refer to the authoritative monograph by Frey & Speare (1988).

[3] A glossary of these terms is provided in STF-1980.

other features. Blocks do not cross the boundaries of counties, Census tracts, or block numbering areas (BNAs).

Block Group (BG)—A combination of Census blocks that is a subdivision of a Census tract or BNA and is defined in all areas where block statistics are collected.

Block Numbering Areas (BNA)—An area defined for the purpose of grouping and numbering blocks in block numbered areas where Census tracts have not been defined—typically, in non-SMSA places of 10,000 or more population and in contract block areas.

Census Tract—A small statistical subdivision of a county. Tracts generally have stable boundaries. When Census tracts are established, they are designed to be relatively homogeneous areas with respect to population characteristics, economic status, and living conditions.

To get an idea of the relative sizes of these entities, in the 1980 Census there were approximately 2.6 million blocks, almost 200,000 block groups, and over 43,300 tracts. Thus, on average, a tract is 5 times larger than a block group, and a block group contains 13 blocks. This geographical organization of the country for the purpose of Census is determined by adopting geographical entities with stable boundaries and in the lower units of track, block number, and blocks, as determined by the imposition of a certain uniformity in geographical extent and features among similar units. On the other hand, the boundaries of LEAs are established through political, social, and historical factors. Hence, incongruities between the boundaries of LEAs and Census units are expected. In the next section, concerns with the Census organization is addressed.

Concerns About the Census Organization

For the SIS project, once the LEAs of interest are identified from the MSAs and their contiguous areas, the nature of the boundaries of the LEAs in relation to the Census organization units must be ascertained to obtain accurate demographic data from the U.S. Census. In some cases, this is relatively simple. For example, in Nevada, each of two MSAs constitute an LEA with its boundaries coinciding with county boundaries. Thus, aggregating the Census data for the LEA is straightforward. However, in those cases where a community is served by several LEAs and the boundaries of the LEAs cut through the Census units, the solution is more difficult. It turns out that a uniform solution is possible for all these situations.

The basic statistical requirements on the geographic organizational units of Census are uniformity in size as in block groups, homogeneity in composition as in tracts, and stability as in counties. The size of communities (cities) varies tremendously over the country, and their growth or decay over time are dynamic. Hence, the city as a unit is not quite suitable for the purpose of the Census. On the other hand, most school districts result from the political and social efforts of a community (often a city or some incorporated entity like a township) or a group of communities to educate its children. As a consequence, their size and boundaries are more in line with the boundaries of cities. However, since the Census Bureau provides aggregated data for cities over 10,000, this is not a major issue with most LEAs.

The boundaries of an LEA seldom coincide with the boundaries of a Census unit unless the LEA is large enough to encompass an entire county or several counties. In those cases, there is no obstacle. The situation that represents the largest dilemma is where several LEAs serve the same community. This, coupled with the fact that in most of the cases the boundaries of an LEA cross the boundaries of the Census units, such as tracts, block groups, blocks, and in some cases, even the boundaries of neighboring counties, requires an approach to the demographic data of a school district at the finest level—that is the block level. The only remaining problem working at the block level is the so called "split block" problem when the LEA boundary cuts through a block. The Census Bureau calculates the portion of the contribution of the split block to the aggregated data for the LEA using the ratio of the area of the split block within the LEA boundary to the total area of the split block. This is often referred to as the proportional-to-area formula.

It is evident that the near-optimal solution to problems arising from the noncongruence of the boundaries of the Census units and school districts lies in a complete blocking of the country and the identification of all the blocks in every school district. This process allows an accurate aggregation of the Census data pertaining to each LEA. However, this is an obstacle of immense magnitude. Illinois (Pohlmann & Chaudhari, 1981) attempted to produce a school district-Census geo-reference file in which each Census block group or enumeration district is matched with the appropriate school district(s). One of the goals of Illinois was to allow Illinois to arrive at a more equitable distribution of federal and state funding for school districts. There were serious limitations in the results. The quality of the Census maps was poor, with errors, omissions, and inconsistencies. There was not a complete set of school district maps in Illinois. There were "serious unresolved differences in the maps of adjacent districts" (Pohlmann & Chaudhari, 1981, p.6). Completely identifying the blocks in each school district was too labor intensive and beyond the resources of the project, and the choice of block groups and enumeration districts resulted in a difficult estimation of population distributions when school districts split these Census units.

The problems at the national level were similar to those Illinois encountered, except the magnitude was far greater. However, the 1970 and 1980 efforts of NCES resulted in a greater awareness of the states in the need for better district maps, and resulted in the 1982 Census Mapping Project CCSSO coordinated. States and most of their counties began a process of mapping out their school districts, which continues today. The availability of accurate and up-to-date district maps is crucial to successfully maneuver around this obstacle.

Another obstacle in the SIS project involves the organizational changes in LEAs over the decade. Given the growth in some metropolitan areas in the past decade, some LEAs were merged to form new LEAs. For example, in California, 23 new districts were formed. Using data from the California Basic Educational Data System (CBEDS) and the Public School Directories over the last decade, SWRL identified the districts from which a new district was formed. For data comparison purposes, the 1990 data of the new district can be compared with the aggregated 1980 data of the old districts. Therefore, once this identification was made for each new district, this obstacle was removed.

In the next section, the solution to the problem is described. Although the outline of the solution is obvious, the actual solution has to wait for the developments in several areas. The efforts in each of these areas are described.

SWRL's Solution to the Problem

The solution clearly lies in the complete blocking of the entire country and identifying all the blocks within each and every school district. To do this would require the confluence of several developments: a sophisticated computerized system to handle the amount of data, accurate and up-to-date district maps for all the districts in every state, and the mapping of these districts onto Census maps so the blocks within each district can be determined.

For this and other purposes, to handle the massive amount of information processing, the Census Bureau developed and implemented a digital cartographic data base called the "TIGER" (Topological Integrated Geographic Encoding and Referencing) system. This system incorporated the most up-to-date information from U.S. Geological Survey Files and Census Bureau's Geographic Base Files. From the TIGER database, the Census maps used for the school district mapping project are generated. An automated system ensured greater accuracy, and most inaccuracies in the Census maps encountered in earlier efforts have been eliminated.

As a result of the 1982 Census Mapping Project, the state data centers began to play a crucial role. Each state assembles the set of maps showing the boundaries of its school districts. These district maps are more up-to-date in their boundaries. However, some problems remain. Over a decade, the boundaries of some of the school districts have undergone significant changes. Some districts merge to form unified districts. The most complete information is generally available from the county superintendent offices. But not all counties have the information on maps. A secondary source of information is the annually updated Public School Directories of the four states being studied (i.e., Arizona, California, Nevada, and Utah).

Once the maps with school district boundaries are available, they are transcribed and color-coded onto the Census maps by following a detailed process specified in the 1990 National School District Program Census Mapping Project Guidelines for Participation (CMPG, 1990). Each county has a set of maps associated with it. The number of maps vary from county to county depending on the density, but typically the number is into the hundreds. An index sheet shows the number of maps, called parent sheets, which cover the entire county and their spatial relationship to each other. For example the number of 1989 Census maps for Riverside County, CA, is approximately 900. Cities or densely covered areas whose details cannot be shown on the scale of the parent maps have inset sheets so that the block numbers can be clearly read. Each inset area has a number of inset sheets associated with it. There are numerous annotating rules to follow, one of which is "school district codes must be assigned for all parts of a school district shown on any number of map sheets" (p. 6, CMPG, 1990).

There is significant improvement on one of the earlier problems of block splitting by district boundaries from 1980. "For the 1990 decennial Census, the Census Bureau delineated Census blocks nationwide. Therefore, the change of school district boundaries coinciding with Census block boundaries is much greater than it has been in the past" (p. 6, CMGP, 1990). It is clear that problems encountered in the earlier efforts of 1970 and 1980 have resulted in a greater convergence of Census and district boundaries for the 1990 Census. However, some split blocks still remain, and the Census Bureau will provide data to NCES on these splitting blocks using the proportional-to-area formula to allot for population assignments. However, states can submit a population proportion based on local knowledge and agreement of the affected school districts.

The massive amount of data clearly dictates the next key to the solution is an automated system. TIGER is a fully automated geographic support system. It is a digital geographic data base covering the United States, Puerto Rico, and the outlying areas. The records in the file represent roads, streets, and other map features of significance to the Census programs and political/statistical boundaries used in Census data tabulation. With this system, the role of the states is therefore to provide accurately annotated district boundaries on Census maps. These annotated maps, after regional review, are forwarded to digitizing sites for editing and digitizing, which requires remote access to the TIGER database in Charlotte, NC. Equivalency Files for school districts are created and forwarded to the Census Population Division for further review and analyses. The Census Analysis Branch produces the tables from the processed files received. The tabulations are sent to NCES for final review. Data products will include data tables, many relating to education specific issues, for each school district. NCES intends to provide the data on the school districts in each state on separate CD ROMs.

It is evident that the merging of district- and community-based data sets has undergone significant advancement in the past 20 years, with a solution that is the result of national and state cooperation combined with the power of automation. Central to this solution is the Mapping Project in which every state assists the national body to determine the composition of the appropriate blocks in each and every school district. As the use and importance of database grows, the Mapping project becomes even more significant if timely and accurate data on school districts are available.

Significance and Use of Database

A consequence of the Mapping Project is a more equitable distribution of Chapter 1 funds ($3 billion annually) because the law requires allocation to the geographic level of either the county or school district. Federal and state educational agencies universally recognize the importance of accurate and timely demographic data. The California Department of Education continues to develop the CBEDS database to provide information on staff, enrollment, finance, facilities, curriculum, and community demographics related to public elementary and secondary education. The CBEDS data are collected on "Information Day" each October from local school administrators and professional staff. The files typically are available within two months. An early report by the California Department of Education (Wang, 1980) argued strongly for establishing a formula and policy to address the many problems of an increasingly diverse population and a state faced with changing demographics. As mentioned earlier, Illinois carried out a district mapping project in 1981 to reconcile the Census data. The Wisconsin School Evaluation Consortium (Landon & Shirer, 1981) published a district data base handbook to provide guidance to districtwide steering committees about how to compile a database. These are just samples of state efforts in attempting to establish a database for educational evaluation, planning, funding, and research.

The National Institute of Education sponsored a study (Burstein, 1983) on the use of existing databases in program evaluation and school improvement, and the possibilities for the future. The author of this study concluded that information maintenance and use in local districts was more a happenstance of competing priorities and human resources and limited technical expertise. In addition, the cost of computing and storage was quite significant in the early eighties. Another problem was information interchange and sharing because different agencies adopted different organizations for their database. Clearly, local and isolated state efforts were insufficient to address the problem of a universal database for school districts.

In November 1984, the CCSSO voted to work actively with the NCES to ensure that reporting of data from all sources is accurate and timely. The primary goal of CCSSO's Education Data Improvement Project was improving the NCES's common core of data, collected annually from state agencies, that is more comprehensive, comparable and timely. Profiles contain information on the federally funded programs: Chapter 1 of the Education Consolidation and Improvement Act, Bilingual Education, Migrant Education, Special Education, Vocational Education, and Food and Nutrition Services. These profiles are analyzed to provide across-the-state operational definitions and comparability (Triplett, 1986). Feedback from the states to the federal level will lead to better national legislation and programs. Throughout the eighties, the awareness of the database's significance was growing. We expect that this trend will continue to accelerate in the nineties.

Future Developments

This paper examined the problem in merging of district- and community-based data sets. It is embedded in the more general framework of reconciling the data from school districts and the data from the Census when the geographic entities are not coterminus. The cooperative efforts between the states, the Census Bureau, the NCES, the development of the powerful TIGER system, and the blocking of the entire nation have all contributed to a solution of this problem. However, the solution is static and still entails a lot of staff power to complete the mapping process. The results are less than timely as the district equivalency file will not be available until 1993. The state data on school districts are still not in a universal form, which make the analysis of across-the-state data rather difficult, and lessen the impact of state data on Congress.

The cooperative efforts of the CCSSO is continuing. However, a tremendous headway will result if Congress passes a Uniform Data Act, which requires all data reported by states receiving federal funds be in some universal format. The initial conversion cost can be shared between the states and the federal government. Present and future technology in computer networks and a distributed database will open up an entirely different world of information sharing. The advent of graphic terminals and computer graphic software will allow updating of school district maps and their transcription onto Census maps far less labor intensive, and hence provide the data in a more timely manner. Recall that most Census maps for each county run into the hundreds, with scores of school districts in each county.

We have made great strides in the past 20 years. This trend will continue and accelerate as computer systems are developed and refined. We can all look forward to more forecasting, better planning, and equitable distribution of resources when decisions are based on timely and accurate data.

References

Burstein, L. (1983). *The use of existing data bases in program evaluation and school improvement. Methodology project.* Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles.

California Department of Education. (1983). *California Basic Education Data System data users' guide.* Sacramento: Author.

Census Bureau. (1980). *Census of population and housing: Summary school districts* (Tape File 1F). Washington, D.C.: Author.

Census Bureau. (1990). *National school district program Census mapping project guidelines for participation.* Washington, D.C.: Author.

Frey, W. H., & Speare, Jr., A. (1988). *Regional and metropolitan growth and decline in the United States.* New York City: Russell Sage Foundation.

Landon, G. L., & Shirer, W. B. (1981). *Wisconsin School Evaluation Consortium,* Madison.

National Center for Educational Statistics. (1970). *Census fourth count (population) school district data tapes.* Washington, D.C.: Author.

National Center for Educational Statistics. (1970). *User's manual for 1970 Census fourth count (population) school district data tapes.* Washington, D.C.: Author.

National Center for Educational Statistics. (1980). *Census of population and housing: School district equivalency files: MARF 3 MARF 4.* Washington, D.C.: Author.

National Center for Educational Statistics. (1980). *User manual for Census of population and housing: School district equivalency files: MARF 3 MARF 4.* Washington, D.C.: Author.

Pohlmann, V. C., & Chaudhari, R. E. (1981). *School district-Census geo-reference file for Illinois for 1980.* Normal, IL: Center for the Study of Educational Finance, Illinois State University.

Triplett, S. E. (1986). *Federal programs information on school and school district universe files. Improving universe data on schools and school districts.* Washington, D.C.: State Education Assessment Center, Council of Chief State School Officers.

Tushnet, N. C. (1992). *Overview of the study of successful practices in metropolitan districts. Symposium: District and community factors that enhance the achievement of disadvantaged students.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wang, P. (1980). *The changing ethnic population and its impact on school districts in California.* Sacramento: California Department of Education.