PERSPECTIVES ON GEOGRAPHIC ASPECTS OF INFORMATION SYSTEMS

BARRY WELLAR
Editor

INSTITUTE FOR SOCIAL
AND ENVIRONMENTAL STUDIES

THE UNIVERSITY OF KANSAS
December 1972

Intentionally

Blank

Page

PREFACE

This collection of papers originated with the

Special Session

GEOGRAPHIC RESEARCH
IN AN INFORMATION SYSTEMS CONTEXT

Organized and Chaired by

Professor Barry J. Wellar, University of Kansas

and held during the

Annual Meeting of the Association of American Geographers

Kansas City, Missouri

April 24, 1972

Intentionally

Blank

Page

TABLE OF CONTENTS

Intentionally
Blank
Page

INTRODUCTION

Geographers, social scientists in general, and persons associated with federal, state, local and other public agencies are increasingly looking to information systems technology as they attempt to inventory, describe, analyze, plan and/or predict spatial distributions of people and phenomena. The importance (or popularity) of the topic is evidenced by the fact that some 800-plus references which deal directly or indirectly with geographic aspects of urban and regional information systems have been compiled in a recently completed bibliography.[1] The rapid rate of growth of the field is illustrated by the fact that at least half the references are less than five years old.

When it is considered that the data elements, data items and data formats which comprise flows within and between systems experience some five interrelated, iterative phases—specification, acquisition, processing, dissemination and application—are involved in at least five stages of data base and information system evolution—conceptualization, design, development, implementation, maintenance—and, if in a public sector context, have intra- and inter-agency government implications, it is not difficult to imagine the number of related open literature references increasing by several degrees of magnitude in the span of a year or two.

Although problems of information overload and weak or non-existent communication within and between groups pursuing information systems topics are not at issue here, they were instrumental in the organization of a Special Session held during the 1972 Association of American Geographers Meetings in Kansas City,

---

[1] Barry S. Wellar and Thomas O. Graff, *Bibliography on Urban and Regional Information Systems: Focus on Geographic Perspectives*, Council of Planning Librarians Exchange Bibliographies 316-317 (Monticello, Illinois: Council of Planning Librarians, P.O. Box 229, September 1972).

Missouri. That is, just as many geographers have not yet been involved in information systems research, so do a number of members of the public sector remain to be exposed to the activities of a professional geographical organization. The session served, therefore, as a means of assembling several representatives of academia and the public sector for the purpose of sharing thoughts about a topic of mutual interest, geographic aspects of data elements, items, and formats in an information systems context.

The nature of geography being what it is, a multi-faceted exploration of spatial distributions, the presence of geographers on an AAG panel requires no elaboration. The involvement of government officials in the session requires further explanation, however, in terms of how the present writer viewed their current and future roles and impacts in information systems R & D, and the relationship between information systems R & D activities and geographers in particular and social scientists in general.

First, statistical series most often used by professional geographers in universities, industry, commerce and the public sector are those generated by agencies of the federal government, with state and local governments also being major sources of urban, regional, and other geographic data. It behooves professional geographers, therefore, to appreciate, if only partially, the why's and wherefore's associated with specification, acquisition, processing, dissemination and application of the data elements, items and formats in the various series and programs of the agencies and governments.

Second, during the past several years the federal Urban Information Systems Inter-Agency Committee (USAC) has been supporting municipal information systems R & D projects in six cities (Charlotte, North Carolina; Wichita Falls, Texas; Reading, Pennsylvania;

Dayton, Ohio; St. Paul, Minnesota; and Long Beach, California).[2] Prior to the USAC program other cities had attempted R & D projects on their own,[3] and at present still more cities (and states) are engaged in efforts to utilize information systems to assist in the operations, control, planning and research activities of municipal (and state) governments.[4] If we are to obtain the reliable, disaggregate, dynamic data bases necessary to adequately empirically test even the more simplistic models and hypotheses of urban and regional change, it seems apparent that the USAC MIS projects must be regarded as one of the more promising avenues. In fact, if one perceives a desired capability as being able to actually depict the "city as data," then there is no ready

---

[2] Developments associated with the USAC projects have appeared in the proceedings and periodicals of such organizations as Urban and Regional Information Systems Association (URISA), American Institute of Planners (AIP), International City Managers Association (ICMA), American Society of Planning Officials (ASPO), Public Administration (PA), and Association for Computer Machinery (ACM). For details about the original intent of the USAC effort, see *Request for Proposals No. H-2-70 for Municipal Information Systems* (Washington: Department of Housing and Urban Development, 1969).

[3] See, for example, U.S. Department of Housing and Urban Development, *Urban and Regional Information Systems: Support for Planning in Metropolitan Areas* (Washington: U.S. Government Printing Office, 1968) and Barry S. Wellar, *Evaluation of Selected Major Information System Research and Development Projects: Implications for the Wichita Falls, Texas, MIS, The Wichita Falls Consortium Phase I Report, Volume IV, Section 2*, PB 206 789, No. 18 (Springfield, Va.: National Technical Information Service, 1970).

[4] Reports concerning many such efforts are contained in the proceedings and issues of such organizations as URISA, ASPO, ACM, FJCC, ICMA, and AIP.

alternative.[5] Hence, the prospect of being able to use the products and processes of information system technology to merge theory and empirical testing at a previously unattainable level, is one which should cause more than a stir of excitement among professionals, geographers and others, regardless of their affiliations.

Third, there is ample evidence that more federal, state and local agencies are continuing to become involved with increasing numbers of programs with an urban or regional thrust (revenue sharing notwithstanding), which invariably means a geographic component. The objectives of many of these programs have as a common goal the betterment of our mental, physical and social well-being. As has been noted numerous times, one of the major deterrents to the effective conclusion of the programs has to do with data problems, such as temporality, aggregation, incompatibility, confidentiality, and so on.[6] A challenge to geographers, and social scientists in general, is to bring the power of their training and methodological capabilities to bear upon these numerous and sometimes seemingly intractable problems. The above is a call for relevance to the extent that information systems technology, while by no means a panacea, gives every indication of providing a means for bridging some of the gaps between urban and regional theory and applications, research and development.

In the face of an almost overwhelming body of literature, it is fortunate that the dialogue among persons representing universities, private industry and commerce, and the public sector has experienced a

---

[5] The concept of "the city as data" has been advanced in urban, transportation, and information systems courses taught by the writer at the University of Kansas. It has proven to be an excellent vehicle for introducing and using the systems approach to formulate and develop statements of problem dealing with singular or multiple components of the urban system.

[6] See, for example, Barry S. Wellar, "Data Standardization," *The Wichita Falls Consortium Phase II Report, Volume XII, Conceptualization Themes* (Springfield, Va.: National Technical Information Service, 1970), 7-52 and Barry S. Wellar, *op. cit., Evaluation of Selected Major Information System R & D Projects: Implications for the Wichita Falls, Texas, MIS.*

dramatic upswing. Regardless of the reasons for the increasing degree of interaction and interfacing, the fact remains that it is occurring, as evidenced by interdisciplinary and inter-agency participation in meetings of such organizations as AAG, AIP, ACM, URISA, HRB, ASPO, FSUC, PA, IS, ICMA, FJCC, RSA, IGU, and so on. This is not to say, of course, that all communication is effective, but rather that persons are crossing disciplinary or jurisdictional and functional or subject matter boundaries, and are exchanging ideas and methodological perspectives.

The present collection of papers represents a further step towards increased interaction between members of public sector agencies and university persons engaged in research with a geographic or spatial component to its data base. Although the papers do not readily lend themselves to one-sentence descriptions, the following synopses appear to highlight the thrusts of the contributions.

GARRISON and HUMMON range over a variety of subject matter related to needs and capabilities for generating small area data, and thereby provide a number of rubrics for the other contributors. EARICKSON focuses on data elements, social indicators and methods of analysis associated with *metropolitan* (information and other) systems. MAGRAW discusses some of the many activities (or lack of them) of *state* government, the often silent partner in the federal-state-local triumvirate, and raises a number of salient points which probe deeply into the problems and pragmatics of developing truly integrated information systems. MITCHEL provides a brief, concise USAC (*federal*) context for the evolution of automated data processing in local government. The latter three speakers address themselves more to needs than to capabilities, but touch upon both aspects in varying degrees.

The final contributions by MEYER and CORBETT deal primarily with capabilities for generating, processing and disseminating census data. Although not intended to be direct responses to any of the preceding writers, both discuss ways and means which are being and can

be utilized by the Census to resolve current and projected needs of small area and other data users.

Barry Wellar
1972
University of Kansas
Lawrence, Kansas

MEETING SOCIAL SCIENCE AND SOCIAL DECISION-MAKERS' NEEDS FOR SMALL AREA DATA*

W.L. Garrison and Norman Hummon**

Abstract

The concern of this paper is with interrelations among small area data systems and the ways in which problems are defined and studied. A point of departure is taken as given: that there are interdependencies among data systems, theory building, and the priorities given to problems. A change of status in any one of these affects the others, and the change of status of interest here is that of the availability of data. The extent to which small area data systems have changed, and the needs for small area data capabilities comprise the focal issue of this paper: the alignment of small area data needs and capabilities.

In preview, our thesis is that the impact of small area data systems upon problem definition and analysis and related matters is (1) uncertain because the capability is fragile and (2) could be great given the structure and needs of social questions. Our hope is that through the processes of need, recognition, and selling, the capability will be strengthened.

---

*This work was partially supported by a grant from the Alfred P. Sloan Foundation to the School of Engineering, University of Pittsburgh.

**Professors, School of Engineering and Department of Sociology, respectively, University of Pittsburgh.

The term "small area data" has come into wide use during the last decade. It has no precise definition, but generally, data aggregated by states are not considered to be small area data, and while metropolitan data or functional economic area data may be so described, the term is usually applied to data pertaining to parts of such areas. Also, the term is used to refer to data delivery and data management capabilities created by the Bureau of the Census, and in urban information systems studies of the last decade: in particular, the capabilities embodied in address coding guide (ACG) and dual independent map encoding (DIME) techniques, and so forth. This paper is concerned with some of the implications of these capabilities, but since the capabilities themselves have been well documented, they will not be reviewed here.[1]

In preview, our thesis is that the impact of small area data systems upon problem definition, analysis, and related matters is (1) uncertain because the capabilities are fragile, for both institutional and theoretical reasons, and (2) could be amplified if the structure and need of social questions are taken into account. Our hope is that through the processes of need, recognition, and selling, the capabilities will be strengthened. For reasons amplified below, the alignment of small area data needs and capabilities is incomplete and the directions in which to proceed are unclear. However, the definition of problems and the alignment process offer important and challenging opportunities. In order to deal with these issues, we will proceed with discussions of the development of small area data capabilities, suggestions for "ideal" small area data systems, and finally an assessment of current and future developments of small area data capabilities.

The Development of the Small Area Data Capability

The chief force for the development of small area data capabilities was and remains the practical matter of the efficiency of data

---

[1] For information contact the Census Use Study, Data Users Services Office, Bureau of the Census, Washington, D.C., 20233. See also the annual *Proceedings* of the Urban and Regional Information Systems Association.

collection, namely: the presumed efficiency of using the mail-out, mail-back format for enumeration in the 1970 Census of the Population and Housing as opposed to having enumerators make calls upon households. The mail-out format was thought to be less expensive and more accurate. Address lists and the coding of those address lists to the areas for which data are ordinarily aggregated and released were required in order to implement the mail Census. The mail-out, mail-back format was set up to produce Census output exactly the same as that which would have been produced by enumerators—and it has in many respects—to meet the needs for comparability with previous Censuses.

Three factors of varying importance can be identified that affected the development of small area data capabilities. First, various Census users expressed demands for data in finer detail (or in different arrangements) from traditional Census publications. These demands grew, in part, out of concern with certain kinds of social, urban development, transportation, education, health, and other problems which require analysis and solution implementation at a rather fine geographical scale. Demands were also expressed by researchers who sensed opportunities for new work if data could be made available at somewhat lower levels of aggregation.

We arbitrarily define two claimants for these capabilities as (a) applied and (b) scientific with the caution that this is a convenient characterization, but overly simplified. These claimants did not cause the small area data capability to come into being and their effect to date on that capability has not been great, although they may have some effect in the future. A persuasive piece of evidence that their effect has been minimal is provided in a recent report of the Commission on Federal Statistics which reviewed the small area data situation and found it wanting.[2] A quote from the Commission Report will make the point:[3]

---

[2] *Federal Statistics, Report of the President's Commission*, I, II (Washington: U.S. Government Printing Office, 1971).

[3] *Ibid.*, I, 126-7.

> Many peoples in state and local governments have questions that cannot be answered with existing data. As with the Federal policy makers, however, their questions are often proved to be inchoate, they generally lack analytic models, and there is substantial evidence that the problems in questions have not been thought through. A role for the Federal Government in providing statistics cannot [sic] hardly be defined when technical issues of what to measure have not been resolved.

In short, the Commission found no substance in claims of applied and scientific small area data users. Whether or not there is substance is quite another matter, and that issue is in a certain way not germain.[4] The point is that the case was not strong and the Commission could ignore it. Related evidence is the matter of the 1975 Census of the Population. It appears that in spite of the many requests for such a Census, it will not come about. In short, the needs for finer-grained small area data have not been convincingly shown.

The second factor impinging on the development of small area data capabilities evolves out of the long standing interest of the Census in how its products are used, and the fact that they have generally lacked such utilization information. Increased concern with planning, programming, and budgeting, and increased concern with the needs of local decision-makers accelerated interest in how Census products were and might be used. This factor is in some ways the complement of the first factor, and its impact on the development of small area data capabilities has also been quite limited.

It should be noted that the problem of justification of activities is much simpler for the Bureau of the Census than it is for most Federal agencies. This is particularly so in the instance of the Census of Population and Housing because the Census of Population is mandated

---

[4] We tend to view the Commission Report as rather inchoate. We believe that the evidence available to the Commission and represented in its own background papers (Vol. II) make the case for more frequent Censuses and a Federal role in small area data matters, and we find the Commission's rejection of both cases artificial.

by the Constitution for the purpose of apportioning congressmen among the states. Furthermore, the Bureau of the Census is a professional organization with a long tradition of coupling to professional groups, and justification for its existence and style of operation, given its internal value systems, is assumed and never questioned. To the extent that there is a general effort at justification, it seems to be a public relations activity.[5]

The most important factor affecting the development of small area data capabilities was the design of an efficient means of conducting the mail-out, mail-back Census. To this end, address coding guides were prepared as part of the pre-Census program of map revision, search for agreement on the content of questionnaires, pre-Census publicity, pre-tests, and training, all of which are continuing activities of the Census Bureau but which are accelerated prior to the implementation of a particular Census. Preparation of address coding guides depended upon local help although address lists were purchased. After some experience with address coding guides that assigned ranges of addresses to blocks, Census enumeration districts, Census tracts, etc., potential sources of error were discovered, so the dual map encoding capability with its more accurate X-Y coordinate system was developed to improve the quality of assigning addresses to areas.

In short, the development of small area data capabilities evolved out of quite practical reasons—the presumed efficiency of a mail-out, mail-back Census—rather than emerging from defined needs for small area data or from the desire of the Bureau of the Census to improve its products. While this evolution may seem somewhat reversed, it is nevertheless a beginning, and as the discussion in the next section indicates, we think it is a promising beginning.

Scenarios for "Ideal" Small Area Data Systems

In the discussion of the factors affecting the development of small area data capabilities, we identified two kinds of users of small area

---

[5] These remarks describe the situation generally and it is most improper to imply that there is no interest whatsoever in the use of Census products; the point is it is not an organizational characteristic.

data: applied users or social decision-makers, and scientific users or social theory builders. In order to outline the problems of designing small area data systems, it is useful to examine the uses of small area data from the scientific and decision-making perspective. The social decision-maker sees immediate and specific problems; the social scientist attempts to discover the relationships within and how social systems operate. Needless to say, these perspectives do not yield a common focus because they do not imply identical data sets and capabilities. However their degree of overlap provides a basis for problem definition.

The Social Scientist

Social scientists are concerned, of course, with the study of social systems, so presumably social data sets should monitor social systems. A description of the basic characteristics of social systems, then, should outline the kinds of information (e.g., data elements and items) to be incorporated in a social data set.

Social systems are finite. Therefore, the components of social systems, people, organizations, and larger collectivities, can be enumerated. Such enumeration provides the basis for a typological view of social systems. Among the more important results gained from this exercise is the determination of the size of the system and of its components. Population counts are good examples of these kinds of data. But typological data often ignore other important characteristics of social systems. Social systems are dynamic. By this we mean, social units, whether people, organizations, or larger collectivities, are constantly acting, interacting, and transacting. From the perspective of the social scientist, activity data are just as important as typological data, yet social data sets generally emphasize the latter. The reasons for this are many; among the more important are the relative ease of collecting typological data and what might be labeled historical inertia of our most important set of social data, that of the Census. Too, the legal basis of the Census is the constitutional requirement for regular population counts.

The last basic characteristic of social systems may seem too obvious to merit explicit recognition. However, the fact that social systems are located in time and space creates one of the most important

demands upon data capabilities. Monitoring social systems from the social scientist's perspective involves collection of data reporting on the activities of social units at specific times and places. We can be more explicit about the kinds of events and activities monitored by such a social data capability. First, it is necessary to define space-time reference coordinates. Then the various kinds and numbers of social units can be mapped onto their appropriate locations. The choice of the appropriate location points is not altogether arbitrary because certain places are much more important to social units than others. For instance, data about places of residence, work, commercial activities, and recreation are among the more important possibilities. An ultimate product of such a data system capability would be the ecological densities of various social units. It should be stressed that such a procedure is more than disaggregation of typological data. Locations in a time-space coordinate system are not characteristics of social units, and, therefore, do not add categories to the typologies of social units; *they are basic characteristics of the social system.*

Among the more important classifications of social units are the demographics of the population. Age, sex, race, family attributes, and nativity define pertinent typologies that can be mapped into the space-time coordinate system. Typologies can also be constructed for human collectivities. Typologies of social organizations, particularly employing organizations, could be constructed and datum such as size could be mapped into the space-time coordinate system. But, even simple typologies of social organizations border on the second kind of information, data on the behavior and transactions of social units.

The most basic kind of behavioral data describe what individuals and organizations do. Because of the importance of work on all facets of social organizations, occupational classifications may well be the most important behavioral data included in a social data system. Directly associated with occupational data is employment data which connect an individual's general work activity to a specific employing organization. Fertility, mortality, marital and migration behaviors, all basic social processes, should also be part of the activity-related data in a social data system. Two other kinds of activity data should be mentioned. First, educational attainment summarizes the creation and transferral of skills so important to the operations of a social system.

Second, data about the incomes of workers are transactional data important to the analysis of many aspects of social systems.

At the social organizational level, data about the production, service and communication activities of formal organizations should be part of the social data system. Much of the economic activity of the social system falls into this category. Data about institutionalized educational, health, and welfare activities are also relevant. Although there is often considerable overlap, the activities of governmental institutions represent another major category of collective social activity.

The comment about time and space as not being merely additional categories for disaggregation of typological data applies more strongly to activity data. The reason for this is that relationships between social units, particularly transactional activities, are often defined in time and space. Treating time and space as merely additional categories often loses the important relational information that social data systems should contain. If social science is to explain the complex relationships inherent in social systems—the causal chains of inputs, outputs, feedbacks, etc.—it is necessary that social data systems contain an adequate amount of the time-space relational-type data. Of course, it is recognized that a maximal social data system from this perspective would be isomorphic to its social system. Such a data system is neither possible nor desirable. The design of data systems involves the selection of typologies and activities that reflect maximal information about the social system. Because omniscience is not a widely shared capability, the design of better social data systems should be an iterative process which incorporates new knowledge. Flexibility in development and design of social data systems is important for another reason. As stressed repeatedly, social systems are dynamic, and, therefore, the basic composition of social units and their associated activities change. A social data system must be capable of detecting change and adapting to it.

The dynamic character of social systems poses many problems for the design of data systems. Since information that can be used comparatively through time is much more valuable than a series of non-comparable "snapshots," one design criterion is the development of measurement systems that are relatively stable yet allow considerable variation within the measurement spaces. A second important criterion

involves the possibility of transforming historical data into comparable formats of current data. A third criterion, which is particularly important for small area data systems, is the ability to maintain spatial mapping capabilities in the face of constant social change. This third criterion is in some ways the most challenging to meet, yet success would magnify the importance of social data systems because it allows social change to be mapped; it provides information not only on what social behaviors are changing, but where change is taking place. While design flexibility is important for all social data systems, it is imperative for small area data systems because in small areas the potential for social change is greatest.

The Social Decision-Maker

The second perspective that we will use to focus on the nature of small area data systems is that of the social decision-maker. Two points are particularly relevant in using this perspective. Social decision-makers need and use data to solve specific perceived problems. Also, social decision-makers invariably operate within some institutional context.

Often the problem-solving activities of social decision-makers concern the administration and allocation of social services. Small area data systems can make two important contributions. First, an assessment of human needs based on small area data can be used to estimate the areal demand functions for the services. This information is useful in setting up the administrative processes which deliver the services. Facilities and personnel can be allocated to increase accessibility and, therefore, equity of delivery of services. Also, since many social services are delivered by routinized or bureaucratic decision-making systems, small area data systems would be useful in the construction of equitable decision-making rules which reflect local area needs.

We have used the phrases "can be used," "would be useful," and "can make." In their present form, small area data systems do not normally meet the allocation decision-making needs of social decision-makers. This is evident from the fact that decision-makers collect their own data as part of the decision-making process. Countless special surveys have been conducted by and for social agencies to meet their needs. In many cases, agencies "interview" people before

dispensing their services, e.g., hospitals, welfare departments, schools. While these "interviews" are normally made to determine eligibility, solvency, etc., and the information would not be available in a general social data system for reasons of confidentiality, such information does have implications for small area data systems.

Before expanding on these implications, it is necessary to introduce the importance of the institutional context of social decision-makers. The kind of organization for which the social decision-maker works in large part defines the kind of information required to make decisions. Welfare agencies require data about employment status, marital status, family characteristics, etc.; hospitals require data on health, economic status, etc. While most institutions collect general data from clients, they also require special data unique to their own needs. We suggest that these unique requirements provide one of the main reasons for the multiplicity of social data systems. Thus we find ourselves in the paradoxical situation of having too much social data due to extensive duplication, and too little social data because of a lack of a broadly based, coordinated, areally defined data system.

Overlaps

The implications of the social decision-maker's perspective for the design and use of small area data systems overlap the social scientist's perspective in several ways. First, social decision-makers need information about what people (clients) do: do they work, have babies, go to school? The specialized data that is collected for the decision-making process often centers on this behavioral information. For planning and administrative purposes, the social decision-maker is concerned about changes in the activity patterns of clients. For example, rising unemployment presumably requires modifications in the delivery of employment services and unemployment benefits. Thus, the social decision-maker's perspective of the dynamic nature of social systems results in requirements for small area data systems that are quite similar to those of the social scientist.

Because social decision-makers attempt to solve the problems of specific people, social data systems which do not reflect the basic areal distributions of the characteristics of people are of little use in

delivering services to those who are in need. The social decision-maker must know not only that problems exist, but who the people in need are, and where they live. These requirements parallel the importance of a space-time coordinate system discussed from the social science perspective.

The last major implication of the decision-making perspective concerns the importance of flexibility in small area data systems. As mentioned above, flexibility is required to monitor social activities and social change. Flexibility of another kind could well be of even greater importance for the design and construction of small area data systems. It may be possible to resolve the data paradox mentioned above by means of extremely flexible data handling technologies. Institutions constantly monitor many social characteristics of their clients. Since these institutions have specific information needs, the number of data collection formats probably equals the number of institutions. It is unreasonable to expect these organizations to adapt their data collection formats to a common format suitable for the construction of a small area data system, so we are left with the technological problem of transforming institutional formats into a common format. The solution of this problem could result in a broadly based "real time" small area data system. Such a system would contain substantial activity information, because of the constant institutional inputs. Such a system would magnify the importance of the time dimension, giving real meaning to the nature and characteristics of social change. Such a system would iteratively approach the requirements and needs for small area data that have been outlined from the perspectives of the social decision-maker and the social scientist.

But such a system requires an extremely flexible data handling and communications technology. We use the term "technology" quite broadly. In addition to the typical hardware and software technologies that are implied, considerable theoretical and interpretative technologies would have to be constructed. Institutional mechanisms would have to be devised that facilitate the iterative use and development of the system, yet provide the necessary confidentiality of the information. Thus, very flexible access and input procedures would be imperative. This may mean that we are not talking about a single small area data system, but a series of data systems that are integrated by their common time-space coordinate systems.

The small area data system or systems implied by the scientific and applied users of small area data may seem unrealistic or even impossible. Yet the goals of nearly continuous space-time coordinates, dynamic data handling capabilities, and the inclusion of activity data specify important directions in the development of small area data capabilities. An assessment of current capabilities and the problems of pursuing these directions are the subjects of the last section of this paper.

Assessment of the Alignment of Needs and Capabilities

Five rather specific capabilities can be identified which satisfy user needs for small area data. They are: (1) development of X-Y geocoding schemes as an alternative to the polygon or area geocoding systems traditionally used by the Census; (2) development of data matching capabilities, from both diverse sources and diverse geocoding systems; (3) development of additional areal unit geocoding systems such as the transportation planners' arbitrary grid traffic zone; (4) inclusion, and in some cases generation, of activities and transactions data; and (5) development of capabilities for meeting problem definition and reporting requirements.

The extent to which the small area data capabilities have been met or will be met is quite varied.

The creation of the geographic base file system (DIME and ACG) represents a major step in developing X-Y coordinate geocoding capabilities. Currently these geographic base files cover 226 urbanized areas. While the software technology of this system is well in hand, the necessary data base (the address coding guides) face continual obsolescence. The rate of obsolescence is rather marked, as the address information prepared for the 1970 Census is to some extent already out of date. Thus, one of the major problems with the system is the maintenance of an accurate data base. Because the Bureau of the Census cannot adequately maintain geographic base files for the entire country, local participation in this process is very important.

Large scale information systems invariably contain errors, and the geographic base file system is no exception. The Bureau of the Census in cooperation with certain local areas is developing and applying software technology to detect and correct errors. Finally, while the

geographic base file system covers a large proportion of the U.S. population, complete coverage has not been attained. Work on extensions of the system is in progress, although funds are somewhat limited. All of this work is being carried out under the umbrella of the Correction, Update, and Extension (CUE) Program of the Bureau of the Census.*

The development of an X-Y coordinate geocoding scheme could be an important tool in data matching capabilities. Theoretically, most any practical level of data aggregation is possible within the X-Y coordinate framework. Diverse data files could thus be matched according to their common areal coverage. Some experience has been gained with such procedures through the Census Use Studies, but considerable work remains before the potential of integrated data systems is a reality.

Pressures for the inclusion of transactions and activity data have been building and will continue to build. They particularly stem from the Federal Highway Administration's needs for transportation planning. The needs are real, for transportation planning is mandated by legislation. Transportation professionals have a choice of using Census data or generating data themselves. Census has prepared computer routines to generate data as demanded by transportation planning, but as yet those data have not been made available. Consequently, there is no way to know at this time the extent to which the small area data capability of the Census will meet the needs of the transportation client. That capability is not necessary to the transportation client inasmuch as that client has the expertise and resources necessary to generate its own data set.

Transactions of interest to the transportation planner are mainly associated with the journey to work. But the journey to work is only one trip out of every five, and other kinds of trips (transactions) are beginning to loom larger in the transportation planning activity. While the capability could be used to handle those kinds of transactions, places of recreation, shopping, and so forth, are not within the Census transactions format, and the shifting interest of the transportation

---

*Editor's note: CUE is discussed in detail later in this volume by Mort Meyer, U.S. Bureau of the Census.

planner may also limit the utility of the Census capability. Thus the addition of the place of work question by itself in the 1970 Census may have somewhat diminished importance.

The value of transactions information to the scientific user is unknown, partly because of data content. Partners in transactions are the activities at each end of the transactions. The Census of the Population and Housing for the most part treats the household side. Other Censuses query the commercial, industrial, and other sides of the transactions. But to the extent to which these other Censuses will be compatible with small area data capability is problematical. Each of these "other" Censuses has its own tradition and its own bureaucracies. Furthermore, they have had extensive experience with mail-out, mail-back data collection and have not found the small area data capability necessary to their data collection. Too, they have confidentiality problems which shift the scale of grids upon which small area data capability can function. Obscuring the relations between data from particular sources will make for mis-match between workable geographic frames with the population data as opposed to those with say manufacturing data.

Confidentiality issues have also blocked the disaggregation of data into certain kinds of units of interest to scientific users. For some analyses, the block, the Census tract, or some alternative grid might be appropriate, but for many in the behavior sciences, the least common denominator is the household or the individual, and these are beyond the bounds of confidentiality constraints.

The necessity for data handling capabilities adequate for meeting the problem definition and reporting requirements of local areas has been much investigated, but the effects of this investigation and the knowledge gained from it upon the small area data capability are yet to be known. Largely at the instigation of Federal agencies who are sensitive to the needs of companion units at the local level, the Census has investigated small area data use and needs at several test sites. The New Haven test site was one, the Los Angeles was another, and Indianapolis has recently been added as a test site. Experience in these test sites has been extensive and well documented, yet it remains difficult to judge the impact of that experience upon maintenance and modifications of the small area data capability. Two observations are in order. First, at the Federal agency level there is some gross

understanding that local people somehow need data, but the Washington bureaucracy's view and understanding of these needs is rather obscure. Inability to reflect these needs in the report of the President's Commission on Federal Statistics may be cited as evidence for this judgment.

Second, the local definition of the "small" area data capability appears to vary with the size of the metropolitan area. The comparative experience of the Census Use Study in New Haven and Los Angeles bear this out. In New Haven, local users perceived the city block as the appropriate areal unit of analysis, whereas in Los Angeles, local users defined small area data in much larger areal units commensurate with current Census reporting practices. We find this observation about problems perception and data needs quite interesting in itself. Its inference from the point of view of development and maintenance of the small area capability is also interesting for it does not bode well if users in larger metropolitan areas have undue influence.

This discussion has highlighted some of the strengths and weaknesses of the small area data capability. The development and maintenance of an X-Y coordinate geocoding system provide the basis for a very flexible small area data system through potential matching, aggregation-disaggregation, and transaction data capabilities. Several institutional constraints have been identified which may retard the potential development of these capabilities.

Finally, to view small area data capabilities as mere information handling technologies is to miss their greatest contribution. The importance of spatial relationships for human populations has long been recognized, if only partially understood. This is reflected in the characteristics of our most pressing social and technological problems. The potential benefits of generating and applying knowledge of these basic human ecological relationships should, in our opinion, provide the impetus for future development of small area data systems.

Intentionally
Blank
Page

SOCIAL INDICATORS AND METROPOLITAN SYSTEMS: DATA REQUIREMENTS AND ANALYSIS

Robert Earickson*

Abstract

In all of the factorial ecology investigations, exemplified by those done in the United States, readily available national census data were utilized to describe the structure of selected metropolitan systems. These descriptions are occasionally referred to as *social indicators*. Census data contain almost no behavioral variables, but are devoted to describing attributes of metropolitan systems. Most of the current ideas about behavioral process within urban systems have been deduced from the study of the attribute structure of urban space. The quality of life of the various subgroups in an urban population should, however, be researched from two directions: A macro-model of the group's effect on the system, utilizing the factorial ecology approach, and a micro-model of individual daily interaction patterns, including spatial extent, mode of communication (language or dialect), and attitudes and purpose expressed in communication. The purpose of this paper is to describe some data requirements for this dual approach and to suggest that comparative macro and micro urban simulation models are necessary to accurately assess the quality of life in our dynamic metropolitan systems. A proposed investigation of this kind for the City and County of Honolulu, Hawaii, provides the geographic setting.

---

*University of Hawaii.

The role of the city planner has traditionally been one of dividing his time and his psyche between day-to-day office functions, being a client to diverse interest groups, and ostensibly heading a community decision-making organization.[1] In the latter function, the demand for any kind of "information system" was virtually non-existent until some bitter and sometimes violent conflicts erupted in many cities over redevelopment projects. Since then, it has become abundantly clear that ways have to be found to anticipate the often unwanted, indirect, social and economic effects of development decisions in both the private and public sectors.[2] Hence, the emergence of information systems technology and the application of systems analytic methods to metropolitan problems.

One of the more sophisticated applications of urban data to systems analytic techniques has been the *factorial ecology* approach. As anyone who is familiar with the term knows, factorial ecology is the offspring of early urban sociology and social area analysis. For the neophyte, a suitable point of departure in the literature is the June, 1971 Supplement of *Economic Geography*, which was edited by Brian J.L. Berry, author and sponsor of many factorial ecologies.[3]

Another familiar application of urban data is to development of *social indicators.* Given the great range of definitions for social indicators, to extend that term in such a way as to encompass the domain of factorial ecology does not appear to be unreasonable. Social indicators are usually gross measures of the well-being of one city, state, or nation relative to another similar geographic entity. Factorial ecology is also a comparison of various sub-regions to one another. The fundamental difference is that social indicators are always related to human values, goals and behavior whereas factorial ecologies are simply

---

[1] Richard S. Bolan, "The Social Relations of the Planner," *Journal of the American Institute of Planners*, XXXVII (1971), pp. 387-396.

[2] Doris B. Holleb, *Social and Economic Information for Urban Planning* (Chicago: Center for Urban Studies of the University of Chicago, 1969).

[3] Brian J.L. Berry, Guest Editor, "Comparative Factorial Ecology," *Economic Geography*, XLVII (1971).

descriptions of urban systems. I think this is an unnecessary distinction and I would like to propose an information systems methodology which would allow the two concepts to complement each other.

There are a number of problems associated with factorial ecology and a critique of the methodology by Rees covers these adequately.[4] I will address a few specific criticisms which I feel are closely related to information systems and urban ecology respectively.

First, factorial ecologies as a rule have relied on a very narrow data base—that of national censuses. Where non-census data were included, for reasons of compatibility they had to be collected on areal units identical, or nearly identical to census units. In this country, the data we may obtain by census are limited essentially to what is necessary for economic and electoral reasons. This procedure deprives us of much interesting health and social data and certainly precludes the collection of attitudinal or behavioral data. In order to upgrade factorial ecologies to the point where they are useful as social indicators, we need more data about the flows, linkages, attitudes, and perceptions spatially distinct population sub-groups in the urban system.[5]

Second, problems involving the size of areal sampling units comprise an enigma which has always been of concern to urban researchers. In most national censuses, the greatest amount of socio-economic data has been made available at the level of the census tract. Since tract sizes are a function only of their population number, socio-economic homogeneity is sacrificed in all but the smallest inner-city tracts. A simple example of this is given later in the paper. The point is that we have to avoid the mistake of attributing associations found at the census tract scale to the individual household, yet we cannot afford to study the entire population, individual by

---

[4]Philip H. Rees, "Factorial Ecology: An Extended Definition, Survey, and Critique of the Field," *Economic Geography*, XLVII (1971), pp. 220-233.

[5]Harvey S. Perloff, "Social Indicators and Social Aspects of State-of-the-Region Reporting: A Proposal" (Los Angeles: The University of California), mimeographed.

individual. We need an intermediate areal unit, such as the enumeration district or a grouping of homogeneous enumeration districts.

Finally, there is the somewhat limited amount of information which we can obtain from factorial ecologies as they are now constituted. Several factors, or dimensions typically emerge which describe structural attributes of urban systems. Out of most American urban ecologies, three or more primary dimensions are usually noted: 1) socio-economic status, 2) age structure or family type, and 3) one or more ethnic factors. Such knowledge does not constitute a satisfactory explanation of urban process, nor does it tell us anything about the short-term changes on which individuals and organizations tend to base action and policy decisions.

To move toward solutions to these problems, I would like to first join the rising chorus of voices which maintain that it is now necessary to expand our data base, and to obtain these data more frequently than at five or ten year intervals.* We need to identify the occurrence and location of stress situations in our urban systems. More specifically, we have to identify what the populace perceives to be the most extreme and most damaging of the areas of inequality and unfairness. For example, How good are the schools, hospitals and courts? How permeable is the social structure for the poor and the minorities? How many and what kinds of man-made and natural risks, stresses and abuses are the various population sub-groups exposed to? What are the various "publics" (ethnic groups, political groups, labor groups, age groups, etc.) attitudes toward each other? What is the health status of these groups? How much does each group participate in community organizations? How accessible are these groups to goods and services? What are the residential turnover rates? Information of this kind will certainly produce more meaningful social indicators.

Recent quality-of-life research in Washington, D.C., Los Angeles, Honolulu (proposed) and elsewhere provides a partial basis for

---

*Editor's note: Note that part of the Garrison-Hummon paper wherein this topic was discussed. Apparently even a deafening cacophony, as opposed to a rising chorus of voices, would not have persuaded the President's Commission on Federal Statistics to recommend in favor of a quincennial census of population and housing.

improved measures of metropolitan social process. Chapin conducted a survey of the activities of heads of households and their spouses in Washington, D.C.[6] Respondents were asked to list for the waking period of the day all their previous day's activities taking five minutes or more, but excluding details about work and personal activities relating to intimate matters. To reduce the large number of combinations and permutations of responses, interview information was coded to a classification system consisting of 11 obligatory and 27 discretionary activities. Associations of income, race, and activities showed marked differences among lower, middle, and upper income families as well as between whites and non-whites.

A group of social scientists headed by Dr. Harvey Perloff at U.C.L.A. are formulating so-called state-of-the-region reports for the Los Angeles Metropolitan Area. These reports are to be compiled by synthesizing three major data sources: (a) census data, (b) local administrative data, and (c) a metropolitan survey. The survey is concentrated on 30 selected census tracts in the Los Angeles area. The nature of the survey data is such that an analysis of attitudinal trends over time will be possible. Given the pattern of residential segregation which exists in Los Angeles, as elsewhere, differences in attitude distribution between tracts often connote ethnic and class-based differences and provide a basis for preliminary analyses of stress points within and among social groups. Other questions focus on the perceptions of the various communities, satisfaction with their environments, opportunities for participation in local administrative decisions, and their orientations toward local governmental structure.

In the proposed Honolulu investigation, we also expect to draw upon census data, local administrative data, and a community survey. Figure 1 shows these data sources and, tentatively, many of the data items we hope to obtain. In some cases we will need to merge data files through extensive computer programs, such as ADdress MATCHing, developed by the Bureau of the Census for that purpose. The object is to disaggregate, or aggregate as the case might be, data to relatively

---

[6] F.S. Chapin, Jr., "Free Time Activities and Quality of Urban Life," *Journal of the American Institute of Planners*, XXXVII (1971), pp. 411-417.

FIGURE 1. Honolulu Social Indicators Information System Structure

Demographic and Socio-Economic Attributes
[U.S. Census Files]

Local Administration Data:

Immigrant Social Welfare Patterns
Residential Turnover
Employment and Unemployment Patterns
Occurrence and nature of sickness during last month
Medicine or treatment received during last month
Effects of old accidents or injuries
Number of times admitted to a hospital during past year
Residence in nursing or rest home during past year
Time of entering hospital (last episode)
Condition for which admitted to hospital
Name of operation(s) performed at hospital
Name of hospital
Part of body affected by illness, impairments, and injuries
Reduction of activity caused by illness, impairments, or injuries
Time spent in bed during last month
Number of days kept out of school during last month
Days kept out of work during last month
Days kept in bed during past 12 months
Patient self evaluation of health status
Number of births and abortions
Location of physicians outside of hospitals

Hospital Record Survey [Health & Community Services Council of Hawaii]

Hospital
Patient's enumeration district
Age
Ethnic Origin

Sex
Religion
Dates of Admission and Discharge
Surgical Procedures Performed
Discharge Diagnosis
Method of Payment
Attending Physician Specialty
Attending Physician's Enumeration District (office location)

Metropolitan Sample Survey [University of Hawaii]

Activity Data:

Time person left home
Place visited
Direction travelled
Distance travelled
Person travelled alone or with accompaniment
Mode of travel
Activity at destination(s)

Communication Data:

Age of person contacted
Occupation of person contacted
Perceived ethnicity of person contacted
Subject of communication
Relative or non-relative?
Period of time acquainted with person
Language or dialect used in conversation

Attitudinal & Perceptual Data:

Attitudes toward Public Services
Attitudes toward other "Publics"
"Accessibility" to Goods, Services, Recreation and Jobs

homogeneous groups of census enumeration districts which can be called *community areas.* In Honolulu County, this will range upward to 150 areal units. As Figure 1 demonstrates, we expect to have extensive health and hospital statistics for our population. The local telephone company can provide information at a fine geographic level about residential turnover, and other governmental agencies are willing to serve as information sources.

The most important source of data for the Honolulu investigation will be the community survey. This survey will focus on patterns of activity, communication, and attitudes. We will be interested in the language and/or dialect used by our respondents in different encounters throughout a "typical" day—an important variable in any social analysis in Hawaii. The plan is to use, as much as possible, interviewers from the same ethnic group as the interviewee so as to relax the natural ethno-linquistic barriers which exist in such encounters.

We hope to make this information system a continuous feature in Hawaii. This will assure an archival data resource which will be comparable over time and similar enough to such data banks as are proposed by Perloff to enable productive social indicator comparison across state lines.

My second suggestion is not a new one, but it refers to a problem that has been pervasive in urban systems analyses in the past. The census *tract* level of reporting is simply not fine enough, with a possible exception being users dealing with central cities of most of our metropolitan areas. As an illustration of the importance of this point, examine Figure 2. Here, a portion of Honolulu County is shown. It is a quasi-rural, sparsely populated area which exhibits little cohesiveness in terms of associations among population attributes and behaviors. Far from being homogeneous ethnically, there are residential pockets scattered along transportation routes in which households of one area would exhibit different patterns of language or dialect and culture-based attitudes and perceptions than in another area relatively nearby. Clearly, it would be erroneous to summarize these data for the census tract. Even at the enumeration district level, a certain amount of heterogeneity exists. This geographic problem is not unique to Hawaii. I might add that I am skirting the remaining issues of data

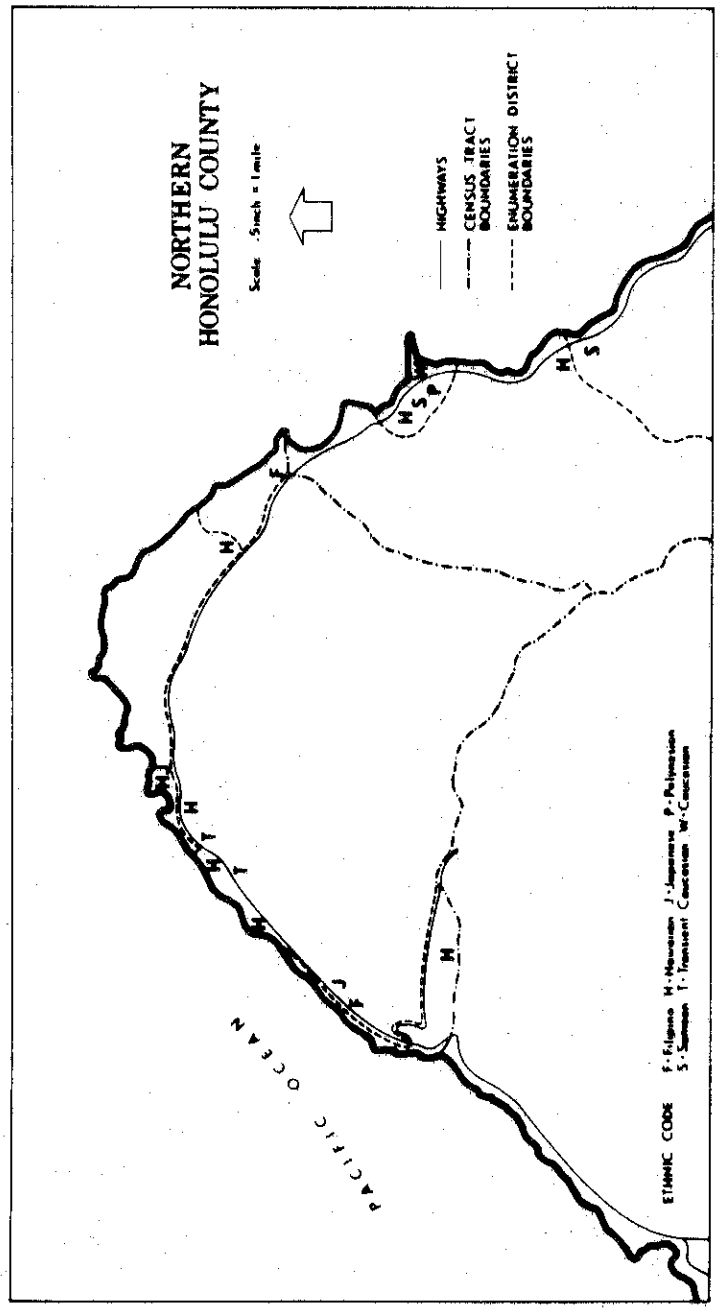


Figure 2. Part of Two Census Tracts in Northern Honolulu County, Showing Diverse Ethnic Settlement Patterns. Enumeration Districts are Ethnically More Homogeneous Than Tracts.

standardization, but these have been adequately covered elsewhere by Wellar[7] and Wellar and Parker.[8]

Assuming we obtain these data, how can we synthesize them in order to produce accurate indicators of trends in the quality of life of the diverse groups in our urban population? As a point of departure, suppose we experiment with several factor analyses and come up with some different models of factor structure and slightly variant community areas. Questions about interpretation of the dimensions and homogeneity of spatial units usually arise on static analyses. If the same problems appeared on analysis of the urban system over two or three time periods, we might also have trouble interpreting the nature of the transition that a particular geographic area was experiencing. Clearly we need a complementary form of analysis to substantiate our deductions.

The solution may lie in the detailed analysis of the social interaction data. In Honolulu, we suspect that the spatial structure of the County is determined by social processes which are heavily influenced by ethnic group associations. Thus, the circulation patterns of individuals are related to, but less important than the communications which take place there. The subjects of discussions, interrelationships of actors, and languages or dialects spoken in various interchanges will form meaningful aggregate patterns. Our objective is to abstract and simulate these ethnic group interaction patterns. These simulation models should complement our factor models, not only clarifying patterns of spatial structure, but also providing clues to the social and spatial transitions that are taking place.

Without particular attention to the problems of geographic sampling and the detail in the data bank, such research will be impossible. I think it is clear that the formulation of meaningful metropolitan social indicators is dependent upon the quality of our

---

[7]Barry S. Wellar, "An Essay on Data Standardization," *Conceptualization Themes, Wichita Falls Consortium Report, Phase II, Volume XII* (Springfield, Va.: National Technical Information Services, 1971).

[8]Barry S. Wellar and Thornton J. Parker, "Standardization: Issues and. . .?" Forthcoming in *Proceedings, Urban and Regional Information Systems Association* (1972).

information systems. Bertram Gross states that "the processes of getting better information, making forecasts, analyzing alternative futures, and establishing goal patterns are necessarily competitive ones, with any procedures of formalized planning leading at best to a partial structuring of the competitive process and at worst to monopoly power for a dominant coalition."[9] In my judgment, we have erred in the past on the side of establishing metropolitan goal patterns without being fully aware of consequences to the metropolis. Hopefully, better urban information systems will result in a positive change in the planning process.

---

[9] Bertram M. Gross, "Urban Mapping for 1976 and 2000," *Urban Affairs Quarterly*, V (1969), pp. 121-142.

Intentionally
Blank
Page

THE INTEGRATIVE, COORDINATIVE, FACILITATIVE, AND ADMINISTRATIVE ROLES REQUIRED OF STATE GOVERNMENTS IN THE DEVELOPMENT OF INFORMATION SYSTEMS WITH A SPATIAL COMPONENT

Daniel B. Magraw*

Abstract

Because spatial components of information systems cannot be viewed apart from all other aspects of those systems, the discussion is aimed in general at information systems—and at the need to think and plan in terms of comprehensive integrated information systems (and related decision systems) both for political units and for functions that cut across jurisdictional lines.

Pressure on existing information systems and demands for improved information systems have been increasing sharply. This phenomenon reflects greater accessibility and reliability of data, and most importantly, the almost desperate efforts by legislative and executive branches to find out "what's happening."

But can we really get to integrated systems within and across governmental jurisdictions? Are we ever going to be able to cope with the system known as a city, with the state water resources system, and with the interrelationships of all such systems?

---

*Assistant Commissioner for Information Services, Department of Administration, State of Minnesota.

> Those are a number of key considerations in viewing that broad question. This presentation assumes that the most important factor relates to systems logistics—and is discussed in the context of what must be done to develop an integrated information system for a state including all functional sub-systems and all political sub-divisions of a state. What kind of environment must be created to maximize the likelihood of the integrated system?

First, please let me apologize for the overly wordy and rather presumptuous label given to my presentation in the program. I got carried away thinking about all the leadership that state government could and should be exercising in statewide information systems, and oftentimes is not. Perhaps a better title—speaking in today's vernacular would be "If you don't know where it's at, Man, forget it." Since spatial components of information systems cannot be viewed apart from all other aspects, we are really concerned today with information systems in a broad sense and by implication, decision systems. Although I will not talk about decision systems, it is obvious, I think, that information systems have relevance only as they feed decision systems. Increasingly, decision systems are being computerized at all levels of the decision-making structure—a good topic for a future AAG meeting.

I am not sure whether my limited sampling of geographers is typical, but on the basis of what I see in Minnesota, they seem to have a broader view of the spectrum of problems and opportunities facing the country and the world than do many of their professional colleagues in academia or in governmental administration. If this is true, it may be traceable to the training geographers receive or their need, occasionally at least, to grapple with a real place or thing.

In any event, the concepts that we choose these days to lump under the heading "information systems" are pretty much "old hat" to the geographers of my acquaintance. They are used to dealing with physical systems and with the data bases related thereto. It appears relatively difficult for them to view an isolated concrete physical fact without trying to get a look at its total environment.

To make a long story short, the pressure on existing information systems and demands for improved and more integrated information

systems have been accelerating sharply—in all areas. We have speculated as to the reasons for the increase and have concluded that the principal factors, in Minnesota at least, are the improving availability of data in terms of quantity and accessibility, the greater reliability of data in terms of validity and temporality, and more importantly, an almost desperate effort to find out "what happened" and "what's happening"—in welfare, ecology, economics, manpower, health, land use, education, etc.

For purposes of this discussion, I am assuming the desirability of developing integrated systems. The crucial question I want to address is whether we *really can get* to the goal of integrated information systems. For example, are we ever really going to be able to cope with the system that is known as a city? With the water resources system of a state, a region, a county, etc? With the interrelationships inherent in the foregoing examples—i.e., the city water resources system is a sub-system both of the total city system and of the county, region, and state water resource systems. How do we determine and evaluate the consequences to each of these systems?

There are a number of key considerations in responding to the question—as to technology (it is essentially here), information theory (in process), organization structure (not crucial), and systems logistics. Time permits discussion only of systems logistics. In other words, assuming all else is "go," can we actually pull together from a logistical standpoint what is needed to build the system?

In the context of a specific framework, what do we have to do to develop an integrated information system for a state including all functional sub-systems and all political sub-divisions of the state? To put the question another way, what kind of environment must be created to maximize the likelihood of the integrated system?

I have addressed this problem under these headings: Planning, Control, Intergovernmental Coordination, Resource Utilization, Accessibility, Data Collection, Priorities, Evolutionary Approach.

Examples of what might be done to organize the environment are largely from experiences with Minnesota state and local governments.

Planning

First, of course, there has to be concrete evidence of the kind of objective we are talking about. Preferably, this should be a legislatively stated objective. A plan, or more likely a series of plans, should be developed to meet the objective. For example, five sets of plans might be appropriate: human resources, education, physical resources, fiscal resources, and administrative management. But if a series of plans exists, there must be some overall coordinating authority to assure basic compatibility of the plans.

In Minnesota, for example, the information systems statute reads:

> The Commissioner (of Administration), after consultation with the state information services advisory council and the intergovernmental information services advisory council, shall design and maintain a master plan for information systems in the state and its political subdivisions and shall report thereon to the governor and legislature at the beginning of each regular session.

This includes, incidentally, final authority on planning for computerization in higher education.

Control

Implementation of any plan requires close monitoring and certain levels of authority to assure consistency with the plan. This becomes both particularly important and particularly delicate when intergovernmental relationships are involved. In Minnesota, at the state level that problem has been met head on in the statutes:

> . . .all plans and programs for systems and procedures analysis, information systems, and related computer efforts of all state agencies shall be submitted to the commissioner (of administration) prior to implementation for review and approval, modification or rejection.

The same official is required by statute to "establish standards for information systems." This is an enormously important authority since more than any other single control, it can lay the base for integrated systems.

Intergovernmental Coordination

One can speak of the state controlling its own agencies, including higher education. Coordination, however, is the word as well as the fact of intergovernmental relations. Even where state authority over local units of government exists, its exercise is most often gentle. But it is absolutely essential that coordinative machinery be established, and where control exists over information systems that it be effectively used. Much of the data necessary to any statewide integrated information system must originate with local government.

Again using Minnesota as an example, the Governor pursuant to statute has appointed:

> . . .an intergovernmental information services advisory council. . .which shall assist the department (of administration) in the development and coordination of an intergovernmental information services master plan to coordinate and facilitate services, techniques, procedures and standards for the collection, utilization and dissemination of data by and between the various spheres of government.

Further, the state Commissioner of Education has been given absolute veto power over school district expenditure for computerization of any kind.

The computer bill also requires the state Commissioner of Taxation to develop a statewide fiscal information system. The statute directs the Intergovernmental Council to recommend how the local governments shall report to the Tax Commissioner. This particular requirement was passed in the special session as an amendment to the computer statute and is an example of the almost desperate cry of "what's happening."

Resource Utilization

Attainment—or even partial attainment of the objective we are speaking of on a cost effective basis—is an enormously complex task. Available resources for the purpose must be pooled and effectively administered. Important among these resources are computer capabilities, communication lines, technical personnel, and data collection efficiencies.

In state after state, the earlier proliferation of computer equipment has been halted. Computer control authority has been centralized, and consolidation of computer centers has begun. In Minnesota, for example, the Commissioner of Administration is "charged with the integration and operation of the state's computer facilities serving the needs of the state government." At present, Minnesota has a single computer center serving all agencies of the state government, except the Department of Manpower Services. In local government, it appears likely that regional centers will ultimately be established around the state. It is also likely that school districts will be serviced by an educational computer utility servicing all levels and institutions of education.

Machinery must be established so that systems design and systems software developmental efforts are closely coordinated. A pivotal central research group for that purpose should be established, ideally as part of the state group involved with standards. The transfer of systems design and/or programs from other jurisdictions (such as USAC projects) must be given careful consideration. The time has come for the burden of proof to be shifted from those who say a particular transfer is feasible to those who say it is not. In this regard, a provision of the Minnesota statute says that the Commissioner of Administration may

> join with the federal government, other states, local governments and organizations representing such groups either jointly or severally in the development and implementation of systems analysis, information services, and computerization projects.

In this spirit, the University of Minnesota and the state Department of Administration proposed a joint budget at the 1971 session to develop certain phases of the state information systems. Funded were the RAFT (Rapid Analysis Fiscal Tool) project, the MLMIS (Minnesota Land Management Information System), River Mile Index System, extension of the MAPS (Minnesota Analysis and Planning System), and a specialized data base management project.

Accessibility

The strongest support to furthering information systems development and integration comes as the value of existing data bases is proven. And increasingly, this value is a function of how accessible are the data. How fast can a user's request be served? Sometimes this is a question of real time response. Other times it may be a question of prompt response to a special report request.

In Minnesota, we have adopted a philosophy generally expressed as "all on-line files," and are gradually implementing that philosophy. To facilitate use of data we are using Qwik Qwery, Mark IV, and STRATA. Further developments in data base management systems will increase our ability to service users and also assist in the integration of sub-systems.

Data Collection

Perhaps the most important single problem facing comprehensive integrated intergovernmental information systems is that of data collection. Even after questions of data element standardization have been resolved, the economics of the situation will require an intergovernmental approach to collecting data one time only.

Vast sums are still expended at and between all levels of government in redundant collection of data and redundant entry of such data into partially or wholly redundant data systems. Within and between city and county systems, as the USAC project has so well illustrated, redundant real property records exist for planning, assessing, and tax collection purposes. One time entry is needed for cost reduction and accuracy purposes.

The Minnesota Land Management Information System, developed under the direction of John Borchert, faces a large problem of updating. Much of the data needed for that purpose is routinely collected at the county level for property tax and related land purposes. But it is not yet done in such a way or with such completeness as to handle the MLMIS update. Until this problem is solved, MLMIS will have unnecessary financial and data temporality difficulties. An example at the state level on this point is the development of a master business file (just now completed) to be used as the central index on all organizations that "do business" with the state. The index has pointers to each operational file in which a particular organization is included. An address change, for example, need be made only once. The relationship of such a file to local government needs is interesting, to offer an understatement—but time prevents comment.

Priorities

The significant costs of information systems have been referred to earlier. The point has not been made explicitly that the managerial and social costs of not developing comprehensive information systems will be devastating, if not fatal, to this democracy.

Were it possible, which it is not, to consolidate all funds budgeted for public information systems in the geographical boundaries of a state into one budget and vest the controller of the budget with full control of information systems development, funding would not be a problem. But funding *is* a problem.

The basic approach required to optimize use of scarce resources is to move as rapidly as possible with improving the environment along the lines discussed above and also to give priority to information systems which have the most rapid pay-off. I do not wish to engage in a debate on whether computerization saves money. I rest my belief that money is generally saved by computerization on many specific examples in my experience and on the assumption that management world-wide, both public and private, generally know why they are rushing pell-mell into more and more computerization.

To attain pay-off usually means visibility in the day-to-day operational activities of government. The data bases developed to serve operational needs will also serve nearly all research and planning needs.

Equally important is that operational data bases are kept current, and using of such data bases for planning has obvious advantages.

Evolutionary Approach

Obviously, it is not going to be possible to arrive quickly at the point of having full scale comprehensive integrated information systems for political units or functional areas crossing political units. It will be a long, slow process.

It is equally obvious, I think, that it will be totally impossible to get there if we fail to establish the destination and the road map to get there—constantly updating the plan as we progress, as technology advances—both hardware and software, as funding fluctuates.

We need only look on every side to see frantic, fragmented approaches to the problems of the day based on inconclusive, out-dated, and non-integrated data. The mass transit-highway-ecology-natural resources-land development clamor is one of several modern Towers of Babel. The screams about the state of the economy can be considered in total little more than rantings and ravings of special interests until we have meaningful ways to develop and integrate information systems relating wage rates, labor supply, unemployment statistics, vocational training and rehabilitation, taxation, etc., etc. Similarly, at present, how is Congress or any state legislative body able to establish funding priorities among welfare, mental health, unemployment compensation, vocational rehabilitation, corrections, and all the related human resource activities? And what about education? And on and on ad nauseum, including all the interrelationships of all of the above and many more.

In conclusion, it should be noted that there is simply no way we can luck out or stumble into or onto an integrated information system. There is also no way rational decisions can be made on the basis of fragmented sub-systems. If we are serious about addressing the problems of the day, we must set an objective for the long range development of integrated information systems and decision systems, and then build the environment necessary to that development.

Intentionally

Blank

Page

THE UTILITY OF GEOCODING TO LOCAL GOVERNMENT AUTOMATED DATA PROCESSING

William H. Mitchel*

Abstract

A major problem of local government is the relatively segmented delivery systems which interact with our citizens. "Improved local government" means more than increased efficiencies. It means improved institutional capacity to analyze the nature of urban problems, explore alternate solutions, allocate resources on the basis of prioritized needs, and assess the impact of selected programs. It means improving local government's capacity to improve those conditions of life over which they can have an effect.

One major aspect of programs designed to achieve improved living conditions is that they are interrelated in their impact. Transportation impacts land use; land use impacts educational requirements; educational programs impact mental health; mental health impacts family planning, and so on in both a simultaneous and a sequential fashion. As the goals of our delivery systems become more complex a greater need for their integration in their operation and in their planning modes develops. This emphasis on integration is probably essential if an acceptable level of local government performance is to be achieved in our urbanized society.

The processes by which improved program integration can be achieved appear to include heavy emphasis on

---

*Former Chairman, Urban Information Systems Inter-Agency Committee (USAC).

information. To be effective, however, this information must possess peculiar characteristics including the potential for aggregation in a wide variety of ways, and on the basis of geographic areas. USAC research is deeply involved in generating methods and techniques whereby automated data processing in local governments will have this capability. The ability to integrate operational data for operational control, problem identification, and cybernetic feed-back rests upon "linking mechanisms." USAC research indicates that the most powerful of these is the individual land parcel when identified in a geocoded environment. Further, many of the problems of local government are best perceived when related to geographic areas or depicted in geographically oriented, visual techniques. The areas, however, vary by problem, time, function, and problem or issue. Most of these areas do not conform to census districts and must be enrichened with data more detailed than those which the census provides. The ability of a local government to aggregate data by empirically defined areas is essential.

"Geocoding" as a term in local government automatic data processing has a variety of meanings and there is no consensus. USAC goals include a substantially increased reliance on "geocoding" with a precise definition yet to be determined. In part, this definition will depend upon cost benefit studies, utility to city operations and management, success in integrating its concepts into a broad spectrum of functions and jurisdictions, and modification of state statutes.

USAC is financing research in two areas: expansion of the street address/intersection node; and de novo creation of ortho-photographic maps with digitized values for significant artifacts and provision for the incorporation of such data into operational files.

CORRECTION, UPDATE AND EXTENSION (CUE) OF THE CENSUS BUREAU'S GEOGRAPHIC BASE (DIME) FILE

Morton A. Meyer*

Abstract

The development of mail enumeration techniques for the 1970 Census of Population and Housing required the corresponding development of new geographic tools; these included the Census Bureau's Metropolitan Map Series and the Address Coding Guides (ACG's). After preparation of the ACG's was well underway (and the census date was too close in time to permit change in the system), it was recognized that these files should have been developed utilizing a set of techniques which were originally developed by the New Haven Census Use Study under the title of "DIME" (Dual Independent Map Encoding). These techniques would have improved the accuracy of the geographic identifiers in the original files, and immediately increased the utility of the files for a variety of other purposes.

Establishing DIME features in what is now referred to as the Geographic Base Files was, however, not long delayed as the Census Bureau and other Federal and local agencies had recognized that it would be most desirable to add DIME features to the already existing files. Local agencies in 194 out of a potential 230 areas participated in Census Bureau programs which resulted in the development of Geographic Base (DIME) Files with x-y coordinates.

These files, however, reflect local geography as it existed immediately prior to or immediately after the 1970 Census.

---

*Chief, Geography Division, U.S. Bureau of the Census.

The importance of having an "in-house" functioning GBF is being recognized by local areas as an essential management as well as research tool, and numerous efforts are now being made locally to update and maintain these files.

The Census Bureau is encouraging these activities through its CUE program (Correction, Update and Extension of the Geographic Base [DIME] Files). The approach being taken recognizes the strong interest of the Census Bureau in a standardized GBF (so that we may use the address coding portion of these records in the continuing activities of the Bureau). It also recognizes the equally strong needs of local areas for file structures suited to their own particular environments. The procedures being proposed, therefore, do not embrace the establishment of a rigid, inflexible system, identical in format and in use, in every area throughout the United States. Rather the local GBF is thought of as being constructed in two parts: one part to include certain standard elements which will apply to all areas, and a second part to contain local information and geographic elements which will vary from area to area, depending upon the local use of the file and local requirements.

In addition, the "E" in the CUE program provides for the extension of the GBF. This phase of the program has begun by the extension of the Metropolitan Map Series (which are not limited to the urbanized core of SMSA's) out to the boundaries of SMSA's.

The correction, update and extension of the GBF on a continuing basis represents a new phase of the Census Bureau's activities in this area. In the coming year, the Census Bureau, with the help and cooperation of local agencies, will be engaged in efforts to optimize the efficiency of procedures, computer programs, and quality and management controls which are being developed for the CUE program.

In 1970 the Bureau of the Census conducted the Census of Population and Housing by a combination of two methods: a mail canvass in the larger urban areas of the country, and a house-to-house

enumeration in the remainder of the country. In 145 of the then 230 SMSA's (not including Puerto Rico) and in certain adjoining areas, the mail canvass procedure was used. Approximately 60 percent of the nation's population received census questionnaires by mail instead of receiving visits by enumerators. Householders were asked to complete the questionnaire in the privacy of their own home, and to mail it back to a local Census Bureau office. The questionnaire addresses were based on a computerized mailing list derived from commercial sources, corrected and updated by the Post Office Department. The geographic areas covered by this procedure are referred to as the *Computer List* areas.

Computer List areas could, of course, be developed only for city delivery service areas. For those portions of the mail canvass SMSA's which were not covered by city postal delivery service, address lists were prepared by Census Bureau field personnel. This file of addresses, which was corrected and updated by the Post Office Department at the time of delivery of the census questionnaire, was *not* computerized. The mailing pieces were addressed manually in the local field offices of the Census Bureau prior to the mail-out. These areas are referred to as the *List-Mail* areas.

The remainder of the country was enumerated by traditional house-to-house canvass procedures.

Geographic Tools

The development of mail enumeration techniques required the corresponding development of new geographic tools; these included the Census Bureau's Metropolitan Map Series and the Address Coding Guides (ACG's).

A. Metropolitan Maps

Maps provide an underlying base for virtually all census activities, and immediately after the 1960 Census of Population and Housing (long before the decision to have a mail out/mail back enumeration) the Census Bureau began a concerted effort to improve maps available for census purposes.

The basic map compilation system that was developed placed heavy reliance upon the "quadrangle" or "quad" maps published by the U.S. Geological Survey (USGS). The "quads" are excellent maps, and

although not directly usable by the Census Bureau, they are essential to provide the frame within which all other mapping data are fitted.

As the maps were developed it became imperative to provide for an updating procedure to ensure that the maps would be reasonably accurate at census time. The decision was made to involve county and regional planning and transportation agencies in this program both because of their metropolitan-wide coverage and their obvious interest as well as expertise in the field. The appropriate agencies were contacted, and all cooperated in the local review, update and correction process.

It should be noted that the Bureau's Metropolitan Mapping program was targeted at producing standardized maps for urbanized areas of SMSA's for use in the 1970 Census. Admittedly, much of the "census" information added to the quads falls short of the accuracy of those basic maps. However, minor inaccuracies have been accepted as an inherent part of the system inasmuch as the maps are intended to show *relative positions* rather than provide engineering accuracy. Our goal was "statistical" accuracy, and by and large it was achieved. The Census Bureau's Metropolitan Mapping Program also resulted in the creation of the first set of standardized urban maps for the United States.

There are approximately 200 map sets (one or more urbanized areas may be in a map set) comprising approximately 3200 map sheets (each including an area of 5 by 7 miles) covering a total area of approximately 110,000 square miles. This coverage includes about two-thirds of the nation's population. With few exceptions, the maps cover only the "census defined" urbanized areas of SMSA's.

B. Address Coding Guides

In the non-mail enumeration areas, following procedures used in previous censuses, geographic codes for each household were determined on the basis of the enumeration district (ED) in which it was located. That is, the enumerator was given a map on which the boundaries of the ED were delineated, and instructed to list and enumerate every address and household located within the ED. In areas scheduled for "block" tabulations (generally cities of 50,000 or more persons), the enumerator also entered on each questionnaire during the course of the enumeration process, the number of the block (as assigned by the Census Bureau) in which the household was physically

located. The codes for these "lowest common denominator" areas (ED's and blocks) formed the base from which code groupings for larger areas such as census tracts, townships, counties, and States were prepared.

However, in the 145 SMSA's in which the mail out/mail back procedures were to be used, a different method was needed—one which could assign a mailing list address to a specific geographic area. The solution decided upon called for the development of a master computer file for each area which would contain the information necessary to "geocode" the addresses. The file developed for this purpose was called the Address Coding Guide (ACG).[1]

To create the file, the Bureau contracted with commercial firms for mailing lists and city directories. From these, plus Bureau source files, computer records containing street names with address ranged for each block side were prepared, and their geographic codes (such as State, county, congressional district, municipality, ZIP code, and census tract and block) were established.

Local agencies were again called upon to assist the Census Bureau and to review, update, and correct the information being developed for the Address Coding Guide. This local review took place between the spring of 1967 and the summer of 1969.

DIME Files

After preparation of the ACG's was well underway (and the census date was too close in time to permit a change in the system), it was recognized that these files should have been developed utilizing a set of techniques which were originally developed by the New Haven Census Use Study under the title of "DIME," an acronym for "Dual Independent Map Encoding." These techniques would have improved the accuracy of the geographic identifiers in the original files, as well as immediately increasing the utility of the files for a variety of other purposes.

---

[1] One hundred forty-seven ACG's were actually prepared; however, only 145 were used since two were in non-mail Census areas.

A. Concept and Utility

The concept of the DIME program is derived from graph theory. Each street, river, railroad track, municipal boundary can be considered as one or more straight line segments; curved lines can be divided into a series of straight line segments. When streets or other features intersect or when line segments change direction, vertices (node points) are generated. Then, by uniquely identifying each line segment and node point and their relationships, and applying the DIME algorithm, a technique of geographic description which can be checked by computer for internal topological consistency is made possible.

The DIME concept also serves a further and equally important function. By digitizing the node points, that is, assigning x-y coordinates, Geographic Base Files were produced from which (by applying straightforward computer techniques) graphic outputs, either in the form of geographic data displays or map images can be readily produced. DIME files with coordinates may, in fact, be best described as a computer oriented geographic system containing the potential of accepting as well as producing graphic information.

B. Establishing Geographic Base (DIME) Files

Establishing DIME files for the mail census SMSA's (called the ACG Improvement Program) was not long delayed, as the Census Bureau and other Federal and local agencies had earlier recognized that it would be most desirable to add DIME features to the already existing ACG's. Federal agencies, including the Department of Housing and Urban Development and the Department of Transportation, participated in this program and assisted with its financing.

Each of the 147 areas included in the original ACG program were again contacted and asked if they wished to participate in the development of a DIME file; 115 SMSA's agreed to do so and re-coding of the files began early in 1970. The entire local phase of this operation was completed by September 1971.

One of the questions asked during the 1970 Census related to an individual's place-of-work. In order to relate a place-of-work address to place-of-residence, place-of-work addresses needed to be geocoded. To do this, an address coding guide was required. The Bureau intended, initially, to create only census tract address coding guides for this purpose as tabulations were not planned for any finer level of geographic detail. However, the rationale for development of DIME

files for the mail area SMSA's obviously held for the non-mail SMSA's, and the Department of Housing and Urban Development and the Department of Transportation agreed to support this program as well.

All 83 of the non-mail areas were contacted, and 79 of them agreed to participate in the program. This work began early in the spring of 1969 and was completed by the spring of 1970.

In summary, 194 areas (plus the new SMSA's of Appleton-Oshkosh and part of San Juan, Puerto Rico) have participated in Census Bureau programs which resulted in the development of DIME files. An additional 32 areas participated only in the original ACG programs, and four eligible areas did not participate in either of these programs. DIME files with coordinates are now available for almost all 194 areas.

C. Information in File

Appendix A contains a detailed description of the geographic elements contained in the GBF/DIME files. Briefly stated the files contain for each street segment (a street segment includes both sides of the street) census geographic location codes (block number, tract, place, State, etc.), name of the street including its type and direction, range of addresses within the segment, identification numbers of node points at each end of the segment, and coordinates of each node point expressed in State plane coordinates, latitude and longitude, and map set miles.

For the 194 areas in total, the files identify over 3,000,000 segments, each containing the full set of geographic elements as described above, and their accompanying 3,000,000 plus node points, each with geographic coordinates. Unfortunately (but, perhaps, to be expected in an undertaking of this magnitude), the files contain various types of residual errors, even though several edits have already been performed.

Before describing some of the residual errors it is noted that the update phase of the GBF program (see below) includes provision for their correction. (Persons interested in a more detailed statement regarding types and quantities of errors remaining in the file, either in general or for specific areas, should write to the Geography Division, U.S. Bureau of the Census.)

Errors exist in the following areas:

1. *High level codes*–High level codes are considered to be the State, county, minor civil division (or census county division), congressional district, place, ward and annexation codes. Up to 1½ percent of the records may have one or more erroneous or missing codes in these fields.

2. *Block chaining rejects*–Each census block was chained from node number to node number around the block to ensure that all sides of the block were coded. If a block side was not coded, was coded incorrectly, or if extra records were coded to the node pair or block number, all records for that block were rejected even though not all records for the block were in error. Up to 5 percent of the blocks in the file could not be completely chained for one reason or another. The number of error records involved will be well below this level, however, probably closer to 2 percent.

3. *Segment identification*–There are a variety of errors existing in the street name and type, and the street prefix or suffix direction indicators:

   a. Truncation of street name (Mail census areas only)–During the initial processing of the file an error in the program logic for shifting street "types" into a field separate from the street "name" led to the truncation of the last six characters of some street names. For example, "BROADWAY" would be converted to "BR" street with "WAY" in the street type field. If this occurred to a name with six characters or less, it resulted in the complete blanking of the name field. The percentage of this error is not known. However, all areas were screened, and any that were particularly bad were rerun.

   b. Variant spellings–No major effort was made on the part of the Bureau to check out and correct the variant spellings or abbreviations (either coded locally or resulting from computer error) which might exist for the same street or non-street feature. For example, the name GREEN might also appear as GREENE or even GREN, resulting in the file identifying one street as two or three separate streets.

4. *Address range*—In the non-mail census areas, street address ranges in the DIME files contain on the average a 5-10 percent rate of error. The rate for the mail areas is unknown, but is not likely to be any lower. Errors occurred either because of erroneous input or, in some cases, processing failure. Types of errors found include overlapping address ranges between adjacent street segments, gaps between address ranges of adjacent street segments, parity errors where odd and even number ranges appear on the wrong side of the segment, or where both "odd" and "even" addresses appear on the same segment side.

5. *ZIP code*—ZIP code errors were corrected in non-mail census areas until the file contained 5 percent or less residual error. No corrections could be undertaken in the mail census areas, and the level of ZIP code error in these DIME files is unknown. A further deficiency is that the new segment records added during the conversion of mail-census ACG files to DIME files do not contain ZIP code information.

6. *Coordinates*—Each node point identified in a segment record should have an X-Y coordinate assigned. If, however, a node point was overlooked, or if the information identifying the node point (map sheet number, tract number, or node number) was recorded incorrectly during the original coding, or during the digitizing, the node point was not assigned coordinates. While most of the areas have less than 5 percent of the node points to which no coordinate values were assigned, this percentage may be as great as 10 percent in some mail census areas.

It should also be pointed out that the accuracy of the coordinate readings in relation to the earth's surface is dependent upon (a) the accuracy of the drafting of the features on the map, (b) the placement of the node dot on the feature (whether it was "right on" the street intersection or slightly off), and (c) the accuracy of the digitized reading. In addition, there were occasional random malfunctions of the electronic digitizing equipment which caused incorrect coordinate values to be assigned to some nodes.

Continuation and Maintenance of the GBF

A. New Phase

We now have a set of DIME or ACG files covering 226 urbanized areas. They have some errors in them, and to cover a point not yet

made, they and the associated metropolitan maps are to some extent out-of-date. They reflect local geography as it existed immediately prior to the 1970 Census. To be of most use the files must be updated as well as corrected.

"Use" is defined, of course, in terms of the greater potential for use at the local level and not in terms of the use made of the file by the Census Bureau. The reasoning behind this statement is that the organization of address information into meaningful geographic units is becoming more and more an essential requirement of effective participation in Federal, State, and local programs. The home-interview surveys of urban transportation planning programs, housing condition surveys, the location of school children by school district and the allocation of police personnel for maximum patrol effectiveness are examples.

Although normally considered as a planning device, the DIME File is, basically, a management tool in the sense that it provides the "framework" upon which a comprehensive information system can be built. And it is now beginning to be used for this purpose—mostly for summarizing and analyzing the large volume of local data which includes a street address as part of the record. Stated simply, data which in the past have been too voluminous or geographically complex to work with can now be "geographically" organized; and by so doing, a major step can be taken to make local data more usable and more understandable to those in decision-making positions—whether the decision-makers are the mayor, the councilman, the director of the Department of Health, the president of the bank or the head of the local department store.

After a somewhat slow start, the importance of having an in-house functioning DIME system is being recognized by local areas, and numerous efforts are now being made locally to update and maintain these files. The Census Bureau is encouraging these activities by developing for local use a standardized approach to the correction, update and maintenance of the file.

B. General Approach

The program we now envision includes, as a first step, establishment of technical and data standards designed to produce, for both the Census Bureau and the local community, a reliable and usable product. Equally important, it undertakes to furnish the local agency

with computer edit programs needed to correct and maintain local files. Some of these programs are already available. We anticipate, perhaps optimistically, that all of them will be available—at least in a production test version—during the late summer of 1972.

The Bureau, of course, has available the edit packages it used to detect and correct errors during preparation of the basic files. Unfortunately, these programs can function only on the Census Bureau's UNIVAC equipment, and the Bureau has only limited resources available at the present time to reprogram into other languages. However, the documentation and logic for various of these programs is immediately available should some local agency wish to program similar edits for their own use.

The approach we are taking recognizes the strong interest of the Census Bureau in a standardized DIME File (so that we may use the address coding portion of these records in the continuing activities of the Bureau). It also recognizes the equally strong needs of local areas for a file structure suited to their own particular environment. The procedures being proposed, therefore, do not embrace the establishment of a rigid, inflexible system, identical in format and in use, in every area throughout the United States. Rather we think of a local GBF as being constructed in two parts. One part will include certain standard elements that will apply to all areas. The second part will contain local information and geographic elements which will vary from area to area, depending upon local use of the file and local requirements.

Standardization is being required only for those fields whose content must have a fixed definition, such as street name, direction and type; potential address range; block and tract codes; minor civil division and place codes; and ZIP codes. Otherwise, the Bureau would be unable to mesh with local files, and participation in joint update and maintenance efforts and the eventual development of a nationwide system of DIME files would remain forever an impossibility. Also, there are various important programs which the Data User Services Office of the Census Bureau is developing, or has developed, for which record standardization of selective geographic items is a requirement. Among these are:

UNIMATCH—(Universal Matcher Program.) A generalized record linkage system which compiles, assembles and executes an address based file matching system tailored to the user's specific tasks; and

DACS—(Dime Areas-Centroid System) which calculates areas and locates centroids of blocks, census tracts, or other areas defined in a DIME file.

C. Initial Phase

The initial phase of the Bureau's maintenance program is geared to the specific situation existing at the local agency. Factors determining the type of maintenance system an agency can initially undertake include:

1. Technical capabilities available to the agency;
2. Extent to which local use is made of the file, particularly as regards the anticipated frequency of the update operations; and,
3. Geographic coding program or programs in which the area originally participated.

Certain preliminary processes, however, will be standard operating procedure applying equally to all local agencies who participate with the Census Bureau in the CUE program.[2] To these agencies we provide, along with a copy of the digitized DIME File, the following two listings, "Segment Name Consistency Listing" and "Coding Limit Line/Unmatched Segment Listing," together with necessary instructions for their use. The purpose of these listings, both of which are compiled from the file, is to identify certain types of residual file errors.

The Segment Name Consistency Listing is a complete list of all unique names, both street and non-street features, appearing in the file. Names are listed in alphabetic sequence, with numbered streets appearing first. This listing can be used to identify and correct inconsistencies in feature names; for example, the GREEN ST problem referred to earlier. Each name listed includes identifications which

---

[2]CUE is the acronym used to identify the programs and operations required to Correct, Update and Extend the coverage of the DIME files.

permit localizing the error to a specific portion of the Metropolitan Map sheet and determining, thereby, what the correct name should be.

The Coding Limit Line/Unmatched Segment Listing is a list of one-sided segments. The only permissible one-sided segments are those which define the outer limit of the DIME file coded area. Any other (or the existence of gaps in the chain of boundary segments) represent errors which must be corrected.

The Geography Division has developed a computer program (written in COBOL) which will enable local agencies to replace the file errors uncovered by these listings with corrected information. The program is called FIXDIME. A detailed description of its content is available upon request. A copy of the corrected information will be provided to the Census Bureau so that it can insert the correction into Bureau files.

The next step, limited at first to areas actually planning to update the files, is to provide the local cooperating agency with an Address Range Edit listing. The Address Range Edit listing is based on the address relationships of all the segments appearing in the file for each unique street name. All segments for each street name are arranged in sequence by the computer by "node chain"; that is, each street is constructed following the same order in which the nodes for that street are delineated on the Metropolitan Map sheets.

Streets which contain segments with one or more discrepancies, such as gaps or overlaps in the address number sequence between adjoining street segments, or odd-even address mixtures on one side of a segment, are flagged for review. Address Range Edit corrections will also be inserted into both the local agency and the Census Bureau files following an operational cycle similar to that just described for the Segment Name and Coding Limit Listings.

Additional FIXDIME programs specifically designed to edit "address range" data (and certain topologic features) and enter the corrections into the file are now being developed and, as noted earlier, will be available to local agencies by late summer.

Once the files have been corrected updating processes can begin. As with correction processes, file updates can only be carried out by the local agency, as only the local agency has the necessary sources of information to identify changes in the political and physical

characteristics of the area (e.g., new streets, annexations, etc.) which must subsequently be reflected in the DIME files.

To assist the local areas in this effort, the Bureau is preparing a FORTRAN IV program called UPDIME (which will soon be available in a production test version). A description of this program is beyond the scope of this paper. In general, however, the program will make possible corrections and additions to the file, including the performance of all necessary topologic edits.

It is important to note that since the DIME files are a computer image of the Metropolitan Maps, updating the computer file must be preceded by an updating of the MMS map sheets. Every cooperating local agency will, therefore, receive a reproducible set of the node dotted and numbered map sheets for their area on which they will add new street development, delete paper streets still appearing in the file, correct street names, add or delete node dots and numbers, etc.

So that the Bureau will also be able to maintain an updated set of map files, each time a local agency completes a CUE cycle and notifies us that they are sending the Bureau a correction tape (or when we request correction information for our use), the Bureau will supply the local agency with blank diazo mylar material, upon which the updated map sheets would be reproduced locally and then forwarded to the Census Bureau. The original map sheets will be maintained locally for continuing use in the CUE program.

D. Testing the Program

To test the CUE system, the Bureau selected some nine areas and classified them into two groups as follows:

System "A"—System "A" areas will be provided with update and maintenance procedures designed for agencies which will regularly utilize computerized techniques for correcting and updating their DIME files. Such agencies will be supplied with all available Bureau edit and correction programs and outputs and will, in turn, supply the Census Bureau with tape copies (in standardized format), of all corrections and additions made to their DIME files. The System "A" areas are Albuquerque, Dallas, Fort Worth, Tulsa and Columbus, Ohio.

System "B"—System "B" areas will utilize update and maintenance procedures which have been designed to accommodate agencies whose currently planned use of the file does not envisage an immediate computerized update operation. For these agencies, clerical correction and addition techniques are being developed. Correction and updated information will be held locally for subsequent submission to the Census Bureau as the need arises for the computer files to be updated. The System "B" areas are Evansville, Memphis, Cincinnati and Hamilton-Middletown.

E. Extension of the Mapping Program

As noted earlier, the CUE program also provides for an extension of the DIME File System. We have begun this phase of the program by extending the Metropolitan Map Series, which is now limited to the urbanized core of SMSA's, out to the boundary of the SMSA. For this new series of maps, the data shown are again being taken directly from the USGS quadrangle maps. However, because of low population densities and large physical areas, the maps are being drawn to one-half the current Metro Map scale, 1" = 3200'. For places which contain a complex or dense street pattern, enlarged insets will be provided. As with the original map series, the extension maps are being sent to the local cooperating agencies for review and editing.

To date we have been able to prepare a total of 431 extension map sheets covering the complete SMSA's of: 1) Los Angeles-Long Beach, California; 2) Anaheim-Santa Ana-Garden Grove, California; 3) Albuquerque, New Mexico; 4) Dallas, Fort Worth and Houston, Texas; 5) Tulsa, Oklahoma; 6) Dayton, Springfield, and Akron, Ohio; and 7) South Bend, Indiana.

It should be pointed out that funding for the extension program is, as yet, somewhat limited and available funds must also be used for preparation of Metropolitan Maps for the 34 SMSA's newly established by the Office of Management and Budget. This latter program is also underway and map sheets for several of the new SMSA's have already been prepared and submitted to local planning agencies for review and editing.

While the extension of map sheet coverage has begun, the corresponding extension of the DIME files has not yet been undertaken by the Census Bureau. This is due primarily to lack of funds. But it should also be pointed out that many problems must be solved before the DIME files can be utilized in rural areas. How, for example, can residences with Rural Route and Star Route delivery addresses be assigned a geographic location code. Certainly more than just a mailing address appears to be needed. Perhaps the only viable solution would be the development of a nationwide, city type, mail address system. While such a system is not yet around the corner, its eventual development seems highly probable.

Summary

The correction, update, and extension of the DIME Files on a continuing basis represents a new phase of the Census Bureau's activities in this area. It is a major undertaking encompassing all SMSA's and including more than 70 percent of the population of the United States.

In the coming year, the Census Bureau, with the help and cooperation of local agencies, will be working and testing to optimize the efficiency of clerical procedures, computer correction and edit programs, and quality and management controls which are being developed for the CUE program. Our goal is to provide local government with the framework for a much needed geographically based planning and management information system capability and, by so doing, to provide the Census Bureau with a complete and up-to-date set of DIME files and Metropolitan Maps. With these the Bureau can improve the efficiency and effectiveness of its own surveys and censuses, and extend the range of services it can offer to the public.

I would like to note in conclusion, that the Census Bureau will be utilizing DIME files, supplemented by Address Coding Guides for small cities and the Post Office ZIP Code Directory, to assign geographic location codes based upon mailing address to the approximately 6.2 million establishments covered in the 1972 Economic Censuses. Coding of the single unit establishment file will take place this coming July. Coding for all files will be completed by August 1973. The techniques being used and results obtained will be the subject of a future report.

Appendix A

DESCRIPTION OF GEOGRAPHIC ELEMENTS CONTAINED IN THE DIME FILE

The following is a description of the geographic elements which appear in EACH record of the digitized Geographic Base Files. Many of these geographic elements contain the descriptions *left* and *right* as part of the name; for example, block number right, congressional district left in order to identify the geography on each side of a segment.[3] The left-right designation is made in direct relation to the placement of the From Node number on the map; that is, "looking" in the direction of increasing addresses from the From Node towards the To Node, the census geographic codes relating to the left side of the segment are designated as left, and those to the right side are designated as right.

The geographic elements which contain the left and right designations are as follows:

1. 1970 ED (Enumeration District)
2. Local Identifier
3. Address Range
4. 1970 Census Tract
5. ZIP Code
6. 1970 Place Code
7. 1970 State Code
8. 1970 County Code
9. 1970 MCD (Minor Civil Division) or 1970 CCD (Census County Division)
10. CD (Congressional District)
11. Area Code
12. Ward Number

---

[3] A "segment" is that portion of a street or non-street feature, shown on the Bureau's Metropolitan Map Series (with node numbers), located between any two node points; it includes the geographic and address codes pertaining to both sides of the street or non-street feature.

13. 1970 Block Number
14. 1960-1970 Annex Code

Segment Identification Fields:

Street Prefix Direction. A two-character alphabetic field for street direction which precedes the name, such as N, S, SW.

Street or Non-Street Feature Name. A twenty-character field that can be alphabetic, numeric, or mixed for the street or non-street feature name. A street feature is defined as a vehicular thoroughfare. A non-street feature is any other map feature, such as a railroad, shoreline, or corporate limit.

Street Type. A four-character alphabetic field for street type abbreviation. The more commonly used street type abbreviations have been standardized for consistency.

Less commonly used street types have not been standardized and may appear in the file with multiple abbreviation spellings or even as part of the Name field. Examples of street types which may contain multiple abbreviations are: Crossing, Gardens, Motorway or Skyway.

Street Suffix Direction. A two-character alphabetic field for street direction which follows the name, such as NE, SW, N, etc.

Non-Street Feature Code. A one-character field, numeric or alphabetic, identifying the type of non-street feature, such as railroads, water features (river, creek, lake shore, etc.), political or census boundaries (corporate limit, township boundary, census tract boundary, etc.), other non-street features (park boundary, street extension, etc.), paper streets (that is, a street which had been proposed, but was not actually on the ground at the time the file was prepared), etc.

ED (Enumeration District). Non-mail census areas only. A five-character alpha-numeric field for the identification of enumeration districts used in the 1970 Census. Enumeration districts are always unique within county.

Local Identifier. A six-digit numeric field containing entries which were inserted into the Address Coding Guides by the local agency at the time of original preparation. Local identifiers appear in mail census areas only.

From Map. A two-character alpha-numeric field identifying the MMS map on which the From Node is located.

To Map. A two-character alpha-numeric field identifying the MMS map on which the To Node is located.

Coding Limit Flag. A one-digit field which identifies the segments which bound (enclose) the geographic area within the SMSA covered by the GBF.

Address Range—Low and High. Two six-character fields, numeric with rare exceptions. The first six characters identify the lowest address of a street segment, and the last six characters identify the highest address. Both are either odd or even within a record and are never mixed, except through error. If there was only one address on a block side, the high and low address generally have been coded with the same number. Non-street features by definition can contain no entries in these fields.

Record Identification Number. A permanently assigned numeric identification for each record consisting of the following elements.

> *File Code.* A four-digit numeric identification assigned to all records. The file code will be unique for each SMSA.
>
> *Record Number.* A unique six-digit numeric identification for each record.
>
> *Check Digit.* A one-digit numeric suffix to the combined file codes and record number described above, mathematically derived from this number, and used to minimize errors of transcription or punching of the number.

Census Tract. A six-digit numeric identification of the census tract as defined for the 1970 Census.

ZIP Code. The five-digit numeric identification of regions defined by the United States Postal Service.

SMSA (Standard Metropolitan Statistical Area). A four-digit numeric identification of each SMSA listed in the *Federal Information Processing Standard Publication (FIPSPUB),* Number 8.

Street Code. A five-digit code, always numeric, for the street name. Street codes appear in mail census areas only.

From Node. A four-character numeric field containing the node number found at the *low* address end of the street or at the beginning point of a non-street segment.

To Node. A four-character numeric field containing the node number found at the *high* address end of the street or at the ending point of a non-street segment.

Place. A four-digit numeric code identifying the place in which the record is located. This term is used by the Bureau of the Census to identify both incorporated places and census defined unincorporated places. (Unincorporated place codes appear in mail census areas only.) The numeric code is assigned to places in alphabetical sequence within the State.

State Code. A two-digit numeric code. The codes used are those defined in the *Federal Information Processing Standard Publication (FIPSPUB)*, Number 5 and are unique within the United States.

County Code. A three-digit numeric code. The codes used are the 1970 county codes as defined in the *Federal Information Processing Standard Publication (FIPSPUB)*, Number 6 and are unique within State.

MCD (Minor Civil Division)/CCD (Census County Division). A three-digit numeric code identifying a township or equivalent census defined area. The codes are assigned in alphabetical sequence within each county, and are unique within county.

CD (Congressional District). A two-digit numeric identification of congressional districts as defined on January 1, 1970.

Area Code. A three-digit numeric code providing an abbreviated identification combining both MCD and place. Area Code is unique within county.

Ward. A two-digit numeric code given to political subdivisions of a place. Ward numbers appear only for mail census areas and only where inserted at the local level during the original preparation of the ACG.

Block Number. The three-digit identification of "blocks," as defined by the Bureau of the Census. (In general, they are equivalent to city blocks.) The left-most digit is 1 or greater except where "pseudo-block" numbers were used to identify large water features during local coding of the GBF. In such cases the left-most digit of the block number is a zero (0). Block numbers are unique within census tract and never cross a tract boundary; however, a block may straddle other boundaries within a tract, such as city limits.

Annexation. Annexation information appears in mail census areas only. A one-digit code of "5" identifies segment sides annexed to places with 2000 or more inhabitants between April 1, 1960, and the time of the preparation of the Address Coding Guide. A zero or blank indicates that the segment side was not part of an annexation made during this period.

State Plane Code. A two-digit numeric code identifying the State Plane grid system in which the node is located. This code is necessary for two reasons: several of the larger States have multiple State plane systems, and SMSA's often cross State boundaries.

Coordinate System. (Note: If coordinate information was not inserted into a file record, the particular coordinate field involved will be blank or contain a spurious one-digit number.)

Map Set Miles. Two six-digit coordinates expressed in miles and thousandths of a mile measured from an arbitrary point at the lower left of the Metropolitan Map Series coverage for each SMSA. The decimal point is implied between the third and fourth digits.

Latitude, Longitude. Six and seven-digit coordinates, respectively, expressed in degrees and ten thousandths of a degree. The decimal is implied between the second and third digits for latitude and between the third and fourth digits for longitude. Latitude is defined as the angular distance north or south from the earth's equator measured through 90 degrees. Longitude is defined as the arc or portion of the earth's equator intersected between the meridian of a given place and the prime meridian.

State Plane. Two seven-digit coordinates expressed in feet relative to the State plane system assigned to the area.

Intentionally
Blank
Page

THE POTENTIAL FOR LINKING ALL TYPES OF GEOGRAPHIC BASE FILES

James P. Corbett*

Abstract

This paper characterizes virtual maps and the general structure of geographic files. The characteristics of simplicity or complexity relative to a fidelity standard are explained. The structural characteristics of geographic files are explicitly stated.

There is a discussion of the problems of source material and its evaluation. The material leads up to a discussion of file collation or file linking. The nature of file linking and some indication of the complexity of the linkage is afforded by examples.

A particular point is made of the utility of the itinerary form of encoding of urban structures. A worked out example of an itinerary is given.

An actual map is a graphic representation of a geographic concept. Although such maps are commonly textually annotated, the graph conveys essential information, particularly information about fundamental geometrical or spatial relations among the elements represented on the graph. Graphic structure is an essential part of an actual map, and in fact a document bearing no conventional graphic representations would not ordinarily be referred to as a map.

When all the relations, both explicit and implicit, conveyed by the graphic structure are reduced to a textual description, we refer to the

---

*Mathematical Statistician, Statistical Research Branch, U.S. Bureau of the Census.

result as a virtual map. The reason for this is that the description clearly is not an actual map, since it contains no literal graphics. However, an actual map can be produced from the description.

The equivalence of an actual map to a virtual map can be tested by translation and retranslation. In practice we will encounter both virtual and actual maps which have only a certain degree of equivalence. One of the oldest forms of textual description is the itinerary. Ancient itineraries often permit us to reconstruct substantially detailed local maps. In modern times, notes, courses, bearings, and textual notes of a competently documented survey exemplify systematic and relatively accurate itineraries. The extent to which such textual descriptions permit graphic reconstruction is a matter of the skill with which the itinerary is written, and the degree and reliability of detail provided.*

A worked out itinerary for a small American town is given in Appendix A. This itinerary is given in abridged form, and the rules for this abridgement are also specified in this appendix.

There are several factors which determine the basic quantity of information which is contained in a graphic structure. In the planar case, the structure is composed of three kinds of geometric entities, points, linear segments, and bounded areas. The relations of order and contiguity among these elements must be specified. The usual formal device for this specification is the incidence matrix, which tabulates the boundary points of each particular segment, and the boundary segments for each bounded area.

When we consider the requirements of metrical accuracy in connection with the basic topological structure still further information is required. A line segment can be thought of as having a simple or complex form, depending upon the total variation of its curvature. In general, representation of the line segment to a given degree of fidelity requires that points be specified to within a given accuracy, and that the number of such points increase with the complexity of the form of the line. Regularity or structural simplicity, where it exists, can be used to shorten the description, either in that fewer points need to be

---

*Editor's note: Corbett's itinerary is applicable, in effect, to any documentation effort associated with data base development, and has obvious transferability implications.

explicitly specified, or that some simple rule can be given whereby geometric shapes can be determined. Street addresses are a form of information subject to analysis from the point of view of simplicity of structure. The most complex assignment of such addresses would be essentially random. In such a case, each individual address would have to be specified separately. On the other hand, there are cases in which street addresses are formed in an extremely simple and regular way. Here the number of individual addresses to be specified may be quite small.

A further example of the simplicity-complexity principle is to be found by comparing the Public Land Survey descriptions for such states as Illinois with descriptions of less regularity. Regular descriptions require explicit mention of far fewer points and areas than do irregular ones. It should be emphasized however, that simplicity or complexity is only determinable relative to a fidelity standard. The apparent simplicity of the Public Land Descriptions begins to disappear as requirements for precision are increased.

We should distinguish between a virtual map and a machine readable code. Historically, a great deal of effort has been expended on devising particularly clever ways of enumerating or identifying geometric entities. Important purposes for which codes are devised are secrecy, semantic compression or compactness, and protection against transcription error. Redundancy is ordinarily an important causative factor in the selection of a code, for example, use of check digits to check transcription errors. Codes not intended to provide secrecy usually have a simple structure, such as substitution cyphers. Approaching the subject of geo-coding from this narrow technical point of view of coding, tends to obscure the more important aspects of the information content of the basic descriptive text of a virtual map. It is a common misconception that incompatibility of codes has something to do with the structure of the code itself. This is rarely the case, the usual predominating cause being the incompatibility or ambiguity of the basic descriptions.

The Structure of a Geographic Base File

The laws of geometry impose a natural structure on a geographic base file. A particular file may be described, encoded and stored in an arbitrary manner, but the basic structure will nonetheless be subject to the general laws.

The file can conveniently be divided into five major parts, of which the first, the index file, can be thought of as inessential, although extremely convenient. The index file consists of lists of entry points into the remaining four geometric files.

The vertex or point file consists of descriptive records pertaining to individual points. Such information as coordinate numbers forms a part of the descriptive record. In an analogous manner, the linear segment file and the bounded area file contain the descriptive records of the corresponding geometric entities.

The three descriptive files are linked by the topological file. This file relates the point, area and linear segment elements to each other through the relations, "bounds," or by the converse relation (dual), "is bounded by" ("cobounds"). The records in the topo-file consist of the segment identifier, the two bounding points of the segment and the two areas separated by the segment. The record may be degenerate in several ways, for example, fields may be absent, or the point or area fields may be identical.

The three descriptive files may contain images of the topo-file. The usual form of the topo-file is as a set of segment records. If this topo-file is imaged for each area code, we obtain an area-keyed image of the topo-file, and if the topo-file is imaged for each point code, we obtain a point keyed image of the topo-file. These images of the topo-file are properly parts of the descriptive files of corresponding geometric elements. Suppose that the element files for a city contain blocks, street segments and street intersections only. Suppose further that the congressional district designation is appended to the block identifiers as a descriptive element. Then by calling up the blocks having one and the same district identifier, we can by reference to the topo-file obtain the list of segments and vertices incident to these blocks. The boundary elements relative to the districts can then be simplified to provide an augmentation of the topo-file which will specify contiguity relations among congressional districts.

The graph represented by this augmentation to the topo-file is a subgraph of the original block graph. This procedure permits us to augment the area file by the list of congressional districts and such descriptive material as may be desired to append to the district identifier. A similar augmentation to the point and segment files can be made relative to the same district identifiers. Such area files can be built up for any kind of administrative area. There are of course many other ways in which the element file or the topo-file may be enlarged.

Data Sources and their Reliability

One of the primary tasks in geo-coding is the location and identification of useful data sources. Sources can be classified in a number of ways: official sources, commercial and industrial sources, and private sources. Sources differ in authenticity, in that maps generally depend upon prior maps, and surveys upon prior surveys. This situation results in chains of references sometimes extending through several prior sources. Maps are usually conflations from varied sources, of variable authenticity and variable consistency.

Sources differ in form. A source may be a map, a textual description, a simple oral description available only from someone familiar with the locality, or a first-hand observation. Much of the value of a source depends on the skill and completeness with which the documentation is written. Finally, machine readable forms are gradually coming into use and can be expected gradually to displace the traditional forms. This will become increasingly true as our abilities to encode graphic material improves, and as these abilities are more widely perceived.

All sources require competent documentation, and should be edited regardless of the authority of the source. This last precaution must be taken if for no other reason than to guard against errors of transcription and interpretation. Edits consist generally of tests of consistency. Where possible tests of self-consistency should be applied. Where common coverage from several sources is available, and the data have been determined to be self-consistent, further tests of mutual consistency among the sources should be made.

Collation, the Process of File Linking

The collation of two geographic files can be made in a variety of ways. The matter is best understood by studying examples.

A file consisting of names and addresses can be collated with an address-ward file. This is a simple collation which results in a conflated file containing names and addresses within a ward. The end result of a collation is usually a conflation. In addition to the conflated file we obtain two difference files, known as the symmetric differences. These consist of the addresses in the name file not found in the ward file, and the addresses in the ward file not found in the name file. These differences are an important result of the collation.

The situation is not always so simple. The Bureau of the Census maintains two types of geographic files. The first, known as the ACG (Address Coding Guide) formed the basis of geographic coding for the 1970 Census. The second, the GBF (Geographic Base File) represents a conflation of the ACG with a map encoding. It is important to be able to collate these two files, if only as a test of compatibility. Effective linking requires matching the files by block-pairs. In order to do this, the ACG must first be processed to obtain the common address ranges for pairs of blocks, and then the set of elements must be matched against the corresponding set of elements from the GBF. This is a matching of sets of elements from one file against sets of elements from another, and is more complex than the simple matching of two address lists.

A collation is not confined to two files having a similar record structure. For example, it may be required to test the assignment of segment records in the GBF to CD (Congressional Districts) by comparing the legal description with the actual assignments appearing in the files. Here the situation is much more complicated. The legal description must first be reduced to a boundary and content description, and this boundary and content extracted from the Geographic Base File by topological methods.

In all these cases a conflation is, at least implicitly, the end product, and in each case a symmetric difference file must be created to indicate any lack of consistency between the two files.

The extent to which two files are collatable is determined by their common geometric content. Failure to collate may be wrongly ascribed

to the way in which the files are encoded. As an example of this consider the so-called MEDLIST file. This is a file of coordinates of centroids of ED's. Consider the problem of collation of this file with the Bureau of Public Roads County file. The only geometrical basis for collation is that the centroids must each be interior to their assigned counties. A similar observation applies to such point codes as PICIDAD.[1] The centroids can give a rough estimate of tract geography on the basis that the tract should contain centroids proper to the tract. We can therefore perform a conflation of the county outline file, and MEDLIST to obtain an estimate of mutual consistency. The ambiguity of the linkage is not a result of the symbology adopted in the coding but a consequence of the fundamental character of the geometric elements identified and described in the two files.

These few simple examples afford some idea of the nature, purpose and techniques of file collation. The examples illustrate that the analytical problems may be simple or complex, and that it is essential to understand the underlying structure of the geographic files in order to perform an effective collation. The particular symbology adopted has only the effect of increasing or decreasing the convenience of establishing file linkage.

Appendix

The following itineraries are abridged. The idea for encoding places by means of abridged itineraries is of unknown origin. There is at least one prior coding of this kind represented by the zip-code map of the City of Chicago appearing in the "Yellow-pages" for that city. The method may be the most efficient method for the encoding of urban structure. It is under test in at least two major coding programs. The

---

[1] PICADAD is a system for machine processing of geographic and distance factors in analysis of transportation and marketing data, where: PI = place identification step, CA = characteristics and area of each place stored in the memory or master deck of cards, and DAD = machine procedure for computing distance and direction of movement.

key to abridgement is to order the itineraries in terms of intersections. The simple example of Montpelier sufficiently illustrates the principle of ordering.

Itineraries of Montpelier, Blackford County, Indiana.

North-South list.

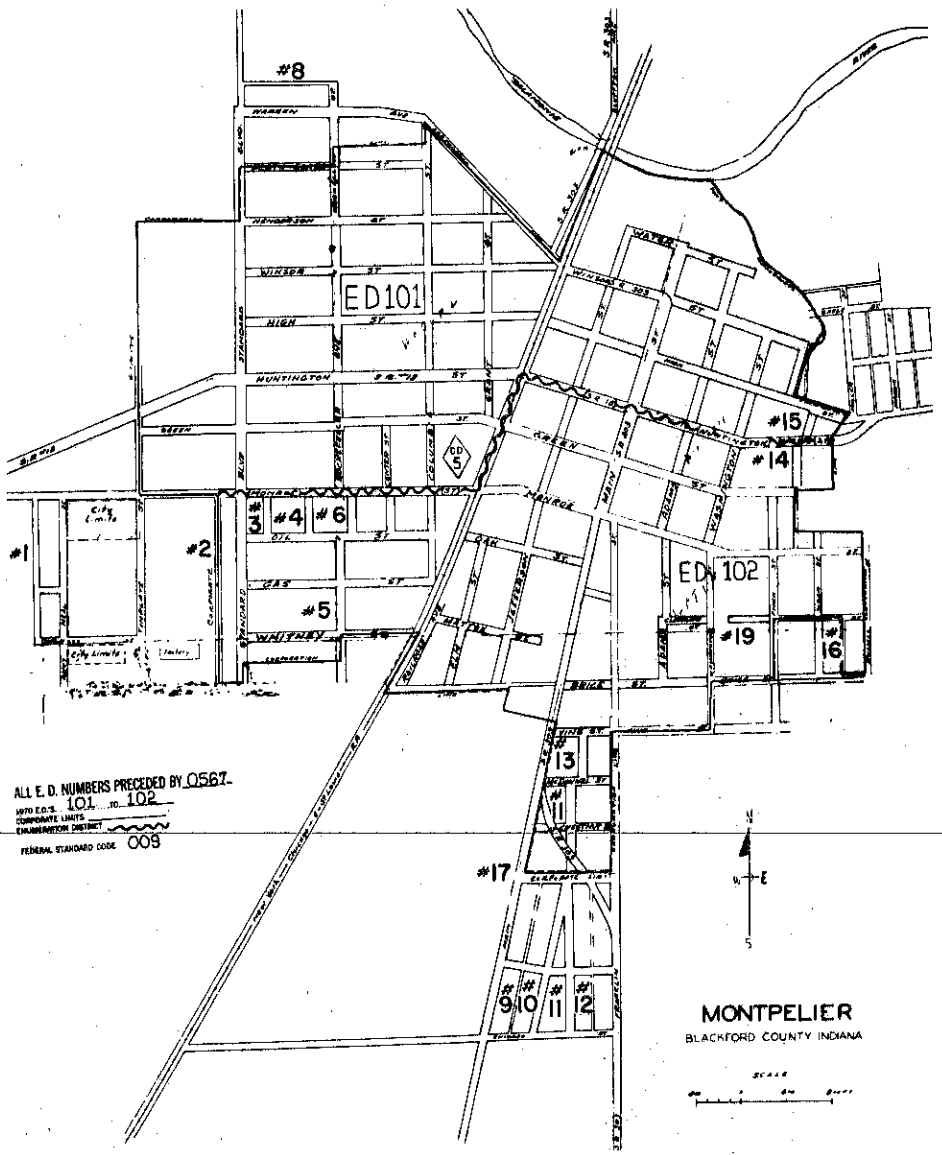
Street	South limit	North limit
No. 1	Whitney (ext. W.)	Monroe St.
Neal St.	Ext. S. of Whitney (ext. W.)	"
Tinplate St.	Whitney (ext. W.)	Ext. N. of Huntington St.
No. 2	Ext. S. of Whitney (ext. W.)	Monroe St.
Standard Blvd.	Ext. S. of Whitney	No. 8
No. 3	Oil St.	Monroe St.
No. 4	"	"
No. 5	Ext. S. of Whitney	Oil St.
Rockefeller Ave.	Monroe St.	No. 8
No. 6	Oil St.	Monroe St.
Center St.	Oil St.	Green St.
Columbia St.	Gas St. & R.R. (1)	Warren Ave.
Grant St.	Green St.	"
N.Y. Chicago St. Louis R.R.	Ext. S. of Chicago St.	N. of Salamonie R.
Railroad Ave.	Brice St.	Ext. N. of Mattox St.
Elm St.	"	Oak St.
Jefferson St.	"	Water St.
Main St.	Ext. S. of Chicago St.	"
No. 9	Chicago St.	No. 17
No. 10	"	"
No. 11	"	McDonnel St.
No. 12	"	No. 17
No. 13	McDonnel St.	Vine St.
Franklin St.	Ext. S. of Chicago St.	Monroe St.
Adams St.	Brice St.	Ext. N. of Monroe St.
Washington St.	Vine St.	Winsor St.
Finch St.	Vine St.	Green St.
No. 14	Green St.	Huntington St.
No. 15	Huntington St.	High St.
Sloan St.	Brice St.	Monroe St.
No. 16	"	No. 17
Russell St.	Brice St.	Ext. N. of Monroe St.

Itineraries of Montpelier, Blackford County, Indiana.

East-West list.

Street	West limit	East limit
Water St.	Jefferson St.	Ext. E. of Adams St.
No. 8	Standard Blvd.	Rockefeller St.
Warren Ave.	Standard Blvd.	Winsor St. & R.R. (1)
Plate Glass St.	"	Columbia St.
Henderson St.	"	Warren Ave.
Winsor St.	"	Ext. E. of Washington St. & S. to High St.
High St.	"	E. City Limits
Huntington St.	Ext. W. of Tinplate to Monroe St.	"
Green St.	Tinplate St. ext. N.	"
Monroe St.	Huntington St.	"
Oil St.	Standard Blvd.	Columbia St.
Oak St.	R.R. (1)	Main St.
Gas St.	Standard Blvd.	Columbia St.
Mattox St.	R.R. (1) & Railroad Ave.	Ext. E. of Jefferson St.
Cleveland St.	Adams St.	Washington St.
No. 19	Ext. W. from Finch St.	Russell St.
Whitney St.	Standard Blvd.	R.R. (1)
Brice St.	R.R. & Railroad Ave.	Russel St.
Vine St.	Main St.	Ext. E. of Finch St.
McDonnell St.	Main St.	Franklin St.
Chestnut St.	Main St.	Franklin St.
No. 17	Main St.	Franklin St.
Chicago St.	R.R. (1)	Franklin St.

1970 CENSUS PLACE MAP



Intentionally

Blank

Page