Title
PROBLEMS IN SPATIAL DATA ANALYSIS

Author
Merrill, D.W.

Publication Date
1982-02-01

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

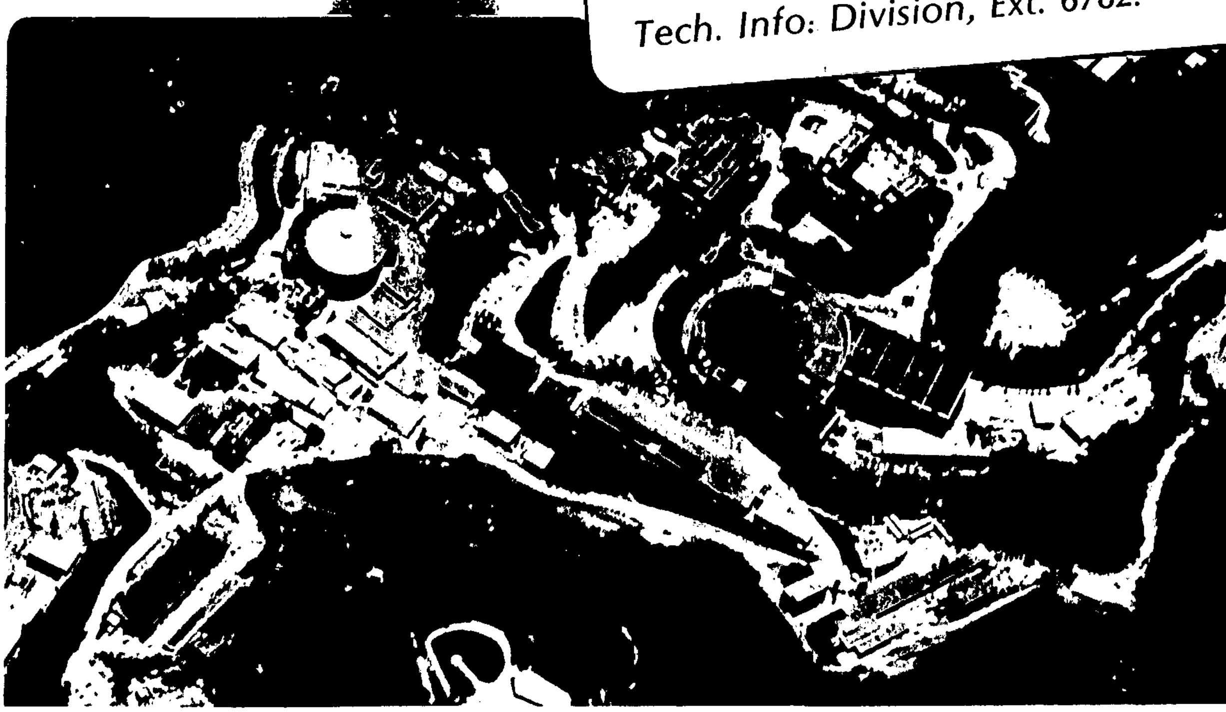
Physics, Computer Science & Mathematics Division

Presented at the Seventh Annual SAS Users Group International Conference, San Francisco, CA, February 14-17, 1982

PROBLEMS IN SPATIAL DATA ANALYSIS

Deane W. Merrill

February 1982



Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

LBL-14047

PROBLEMS IN SPATIAL DATA ANALYSIS

Deane W. Merrill

February 1982

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

This work was supported by the Office of Energy Research, U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Submitted to the Seventh Annual SAS Users Group International Conference, San Francisco, California, February 14-17, 1982.

PROBLEMS IN SPATIAL DATA ANALYSIS

Deane Merrill, Lawrence Berkeley Laboratory

ABSTRACT

A common problem in dealing with socio-economic, environmental, demographic, and health data is the need to combine data from differently defined geographic entities. For example, 1970 and 1980 Census data are not directly comparable even at the county level, due to changes in geographic definitions. More drastic problems occur in combining data from different government agencies.

In most integrated data systems, either the problem is ignored, or data are forced into consistency with undesirable effects: one must either aggregate to larger areas with loss of geographic detail, or disaggregate to smaller areas under arbitrary assumptions. A unique solution to this dilemma has been implemented in SEEDIS, the Socio-Economic Environmental Demographic Information System at Lawrence Berkeley Laboratory (LBL).

In SEEDIS, some 70 geographic levels (e.g. 1970 or 1980 counties) are defined, corresponding to archived data files. If efficiency considerations are ignored, each file needs to be stored only at the most detailed level for which the data are complete. Geocode correspondence files provide the information required to transform data from any level to any other level. Inevitably, disaggregation requires ad hoc proportionality assumptions; different assumptions are suitable for different applications. In SEEDIS, unlike other systems, these assumptions are under the the user's control.

The SEEDIS geocode conversion files and proxy variable files can be used to aggregate and disaggregate arbitrary data files, either within or outside SEEDIS. Other files now being developed describe other dimensions of the data, for example industrial or occupational classification.

THE NEED FOR GEOGRAPHIC COMPARABILITY

A difficult problem faces the analyst who tries to combine and analyze data from a number of different sources: in general the geographic units of his/her various data sets do not correspond. Fortunately, Federal Information Processing System (FIPS) standards are being used increasingly within the Federal government. Although FIPS standards can resolve differences of nomenclature for the same geographical entities, they cannot adequately handle the problem which is discussed here.

Agencies define geographic entities in response to their own needs. Water quality data, for example, are most easily described with respect to natural watershed boundaries. Socio-economic data, on the other hand, are more naturally collected for major urban centers. Pleasant as it might be for the analyst, it is unreasonable to expect agencies to relate all data to an arbitrary set of standard geographic units.

Geographic boundaries, especially political ones, change over time. Boundaries of Congressional districts change every two years in order to ensure representation proportional to population. Standard Metropolitan Statistical Areas (SMSA's) are periodically redefined according to established criteria by which they receive public financial assistance. Even relatively stable entities like counties can change; between 1970 and 1980 there were county boundary changes in (at least) Alaska, Colorado, Hawaii, Michigan, South Dakota, Virginia, and Puerto Rico. To be useful, data must describe the real world. One should not expect the Bureau of the Census to publish 1980 Census tabulations corresponding to 1970 (or 1790!) county definitions.

One must always consider one's particular application and be precise with definitions. One might like to know which cities of the United States gained or lost population between 1970 and 1980. Should one say that a city's population has increased by 30 percent if the increase is due merely to an enlargement of the incorporated area? Such an increase would be important to a city planner concerned with tax revenues, but not to an environmentalist or demographer concerned with changes in population density. In this paper we focus our interest on the more difficult second case, where one wants to study various aspects of the physical world, regardless of political nomenclature.

Methods are needed for easily comparing data sets which describe slightly different geographic entities, for example 1970 or 1980 Census counties. A related problem is the need to aggregate data to larger geographic areas (for example states to Federal regions) either to achieve meaningful sample sizes or for summary display purposes.

SOME APPROACHES TO THE COMPARABILITY PROBLEM

Analysts and government agencies have adopted various schemes in an attempt to resolve the problems of geographic incompatibility.

The simplest approach (not a solution) is to simply present the data for the particular geographic entities to which they pertain. For example, the Bureau of the Census provides 1970 data for 1970 counties and 1980 data for 1980 counties. The conscientious user soon notices that the county codes in the 1970 and 1980 Censuses are slightly different from each other, and in both cases are slightly different from standard FIPS codes. The FIPS definitions themselves changed slightly between 1970 and 1980.

Another approach (again not a solution) is to match areas by name or geocode. For example, in the Census Bureau's County Data Book, differences between county definitions in the 1960 and 1970 Census are ignored. In the cases where county boundary changes occurred, comparisons of 1960 and 1970 data give erroneous and misleading results.

A third approach was adopted for Oak Ridge National Laboratory's Geoecology Data Base [1], a collection of diverse county-level environmental data designed for use with the Statistical Analysis System (SAS). The authors defined a standard set of 3071 county equivalents, which are aggregates of counties as defined in several major data sets. For example, each independent city in Virginia is combined with one of the counties adjacent to it. By simple aggregation, the Geoecology authors calculated (or estimated) all data in terms of the 3071 standard county equivalents. The Geoecology user is thus relieved of the onerous task of achieving comparability, but at a price: finer geographic detail originally present in some of the files has been lost. Also, data from a few larger areas had to be disaggregated to the standard Geoecology county units, requiring arbitrary proportionality assumptions. Likewise, disaggregation of data from newer geographic entities like 1980 Census counties will require similar proportionality assumptions.

A fourth approach, rather opposite to that of Geoecology, was adopted for the early results of the Lawrence Berkeley Laboratory's PAREP (Populations at Risk to Environmental Pollution) project [2]. In the PAREP project, data were disaggregated to provide estimates for county pieces, i.e. the largest subcounty units that can be aggregated to form the various county units of any of the original files. Although the user loses none of the detail present in the source data files, a serious drawback of this approach is its dependence on proportionality assumptions out of the user's control.

The next section of this paper describes a fifth approach, which has been implemented in SEEDIS, the Lawrence Berkeley Laboratory's Socio-Economic Environmental Demographic Information System [3]. General features of SEEDIS are described in the final section of this paper.

THE SEEDIS APPROACH TO THE COMPARABILITY PROBLEM

Fundamental to SEEDIS is the concept of geographic level, i.e. the geographic detail of the data in question. Most geographic information systems, for example UPGRADE [4] and DIDS [5], define several geographic levels, usually including nations, states, counties, and a few others. For mapping purposes, geographic base files (GBF) are provided, which describe the boundaries of each geographic unit at each level. Presently, SEEDIS defines 77 different levels of geographic detail; 52 of these have associated GBF's for choropleth or symbol mapping.

In Table 1 are listed the 77 geographic levels in SEEDIS as of February 1982. "Level" is the name used by SEEDIS; "Description" briefly describes the geographic entities of the level; "Year" is a year for which the entities are defined; and "Units" is the number of geographic entities at that level. "Map" indicates the availability of a GBF for mapping; "p" specifies polygon boundaries for shaded choropleth mapping; "s" specifies point locations for symbol mapping; "-" specifies the absence of a GBF. Most geographic levels are defined for the United States, including the territories of American Samoa, Guam, Puerto Rico, and Virgin Islands. The NATION80 level provides the 233 nations of the world, according to 1980 FIPS definitions.

SEEDIS, unlike other systems, explicitly distinguishes between minor variants of geographic levels, for example 1970 and 1980 Census counties. Separate levels describe Bureau of Economic Analysis Areas as defined in 1969 and 1977, or Standard Metropolitan Statistical Areas (SMSA's) as defined in six different years. Each SEEDIS level corresponds to at least one data file installed in the system; every data file is installed at least at its own most detailed geographic level. A file can be installed at several different levels; for example Summary Tape File 1A (STF1A) of the 1980 Census of Population is installed at ten different levels; namely, those provided by the Bureau of the Census plus some others obtained by aggregation.

Of course it is impractical to archive every data file at every geographic level. SEEDIS achieves comparability among different files through a set of detailed geocode correspondence files. These files, still being completed, describe each geographic level in terms of the units of other levels. Where entities mutually overlap, for example 1970 and 1980 Census counties in Alaska, subcounty entities (1970/1980 county pieces) are defined whose membership in each 1970 or 1980 county is unique.

The SEEDIS user obtains data from an archived database at its original geographic level; then he/she transforms the extracted data to any other desired level for comparison with other data. As an example, suppose one wishes to map, at the Census tract level, percentage change in population between 1970 and 1980. As indicated in Table 1, SEEDIS provides polygon mapping at the 1970 census tract level (the GBF was produced at LBL for the Census Bureau). At least four private companies have produced proprietary GBF's of 1980 Census geography, but none is available in the public domain.

In SEEDIS, 1980 Census population can be obtained for 1980 Census tracts; the data are aggregated or disaggregated (usually the former in this case) to the level of 1970 tracts; then 1970 population is extracted, the percentage change is calculated, and the map is drawn. Where 1970 and 1980 tracts mutually overlap, an automatic intermediate step involves partial disaggregation to the level of 1970/1980 tract pieces. Special software, still being developed, is available in SEEDIS to make this task as simple as possible for the user. The basic tool is a enhanced and corrected version of the 1970/1980 tract correspondence file available from the Bureau of the Census.

In aggregating or disaggregating, SEEDIS must distinguish between additive data, such as land area or labor force, and non-additive data, such as population density or per capita income. Four cases are considered: (1) aggregation of additive data involves simple addition; (2) disaggregation of additive data assumes proportionality with an proxy variable such as land area or population, which is selected from a menu by the user; (3) aggregation of non-additive data is an average, weighted by a user-selected variable; (4) disaggregation of non-additive data assumes the same data value throughout the entire larger area.

IMPLEMENTATION OF DATA AGGREGATION AND DISAGGREGATION

At the present time, data aggregation and disaggregation in SEEDIS are crudely implemented for most but not all geographic levels. The user interface is primitive and the choice of proxy variables is limited. The implementation is somewhat level-specific, and response is slow. Present development is directed toward elimination of all these problems.

Later, default weights and proxy variables will be specified for all 27,000 data elements in SEEDIS. (A data element is an attribute stored for each geographic entity in a database.) At the same time, the additivity or non-additivity of each data element will be recorded, so that less user intervention will be required. The discriminating user will still be free to choose his/her own weighting and proxy variables, when desired; these can be supplied by the user or selected from any data-base in SEEDIS.

The concept of geographic level will be generalized to other data dimensions, including time, race, age, industry, occupation, and cause of death. Hierarchical descriptions, cross-level correspondence files, and proxy variable files already exist for 1963, 1967 and 1972 Standard Industrial Classification (SIC) codes, 1970 and 1980 Census industry codes, and 1972 input-output industry codes. Similar descriptions are available for the different revisions of the International Classification of Diseases and Accidents (ICDA) codes, and the various occupation codes used by different government agencies. Geography, the one dimension already studied in detail, is the most complex.

It is interesting to note that additivity can apply in one dimension and not in another. For example, population is additive over geography but not time; inches of rainfall are additive over time but not geography; population density is additive over neither geography nor time; and industrial emissions are additive over both geography and time.

INTERPOLATION OF POINT DATA

Our discussion so far has considered the transformation of areal data from one set of geographic units to another. The geographic units may be represented as polygons which cover the area being studied. The methods described so far are not appropriate for the data measured at discrete sampling points, such as weather data, air or water quality data, or geologic survey data.

Theoretical methods for interpolating point measurements are extensively discussed in the literature. A comprehensive overview is provided by Ripley [6], who identifies five general methods for the smoothing and interpolation of point data: trend surfaces, moving averages, tesselations and triangulations, stochastic process prediction (kriging), and contouring. One of these methods, a moving average, has been implemented in SEEDIS in order to allow the user to easily estimate air quality at any desired geographic level. Input to the model is a data file of 1974-1976 air quality at individual monitoring stations, along with their latitude and longitude coordinates. The model is a generalization of the method used in the PAREP project [2] to produce air quality estimates at the county and tract level [7,8]. The county level estimates have been widely circulated and are included in the Geoecology data base [1].

Numerous studies in the last decade have attempted to relate human health to air quality, using, for example, mortality rates and air quality estimates at the SMSA, county, or tract level. Such applications are concerned only with long-term effects on a mobile population, so extreme detail in time and space is unnecessary. The same model might be used to estimate, say, crop exposure to acid rain, or long-term changes in population density.

In a moving average model, a data value at any point is estimated as a weighted average of values measured at nearby points. The weight $w(d)$ is a decreasing function of the distance d; according to Ripley [6], common choices of $w(d)$ are d^{-r}, e^{-ad}, and e^{-ad^2}, where a and r are positive constants. In the SEEDIS model, $w(d)$ is of the third form, specifically:

$$w(d) = \exp^{-0.5 \frac{d^2}{d_0^2}}$$

The choice of $w(d)$ was dictated by three criteria: (1) the estimated function should be smooth in the vicinity of the measured points; (2) the estimated function need not pass exactly through all the measured points; (3) the area integral of the estimated function should be finite, so that distant points can be ignored in the calculation.

An estimate of air pollution is obtained as a weighted average at any desired point, for example at the population centroid of a county or tract. Reasonable values of d_0 are dictated by data availability and consistency, by the variability of air quality over time, and by the mobility of the population. Studies are continuing at LBL to determine the validity of the model.

For the PAREP county level analysis [7] and the Geoecology data base [1], estimates were calculated at county population centroids with d_0 = 20 kilometers. For the PAREP tract level analysis [8], estimates were calculated at tract centroids with d_0 = 10 kilometers. In the interactive SEEDIS model, the choice of centroids and the choice of d_0 are left to the user. For an area larger than a county, estimation at a single population centroid is not appropriate. In this case, a population-weighted average of county values should be calculated, using the procedure described earlier for aggregation of non-additive data.

SEEDIS OVERVIEW

SEEDIS, the Socio-Economic Environmental Demographic Information System, is an interactive data management and analysis system under development by Lawrence Berkeley Laboratory, a Department of Energy facility administered by the University of California. SEEDIS embodies some 60 person-years of integrated development under funding by the Department of Energy, the Department of Labor, and other government agencies.

SEEDIS provides researchers with easy access to data files installed in the system, including the 1970 Census of Population and several hundred other socio-economic, environmental, demographic, health- and energy-related files. The 1980 Census is partially installed, and the rest will be available a week or two after its release from the Bureau of the Census. At LBL, the data base on tape in an automatic tape library (ATL) includes 2.5 billion data values, roughly 25 billion bytes or 250 tapes at 6250bpi. The archive will double with the addition of the 1980 Census. The most frequently used data, 85 million data values, are presently installed for rapid interactive access in a disk-based system. Later this year, caching mechanisms will be implemented to provide interactive access to the entire tape-based data inventory.

SEEDIS operates in a network of nine DEC VAX-11/780 computers, located in the San Francisco Bay Area, in the State of Washington, in Washington DC, and in North Carolina. Other government agencies having similar equipment could link into the SEEDIS network or operate independent SEEDIS systems. For automatic access to the LBL data archive, a DECNET link to one of the nodes already in the network would be required. Data can be archived anywhere in the network; except for response time, SEEDIS at every node behaves as if all the data were locally stored.

Users can easily extract data required for specific applications. These data can be analyzed within SEEDIS or exported in a simple self-describing format to other computers. Conversely, users can load data from external sources for analysis and graphic display in SEEDIS. Numerous mapping and other display options are available. A variety of terminals, plotters, and film recording devices, both black-and-white and color, are supported. SEEDIS is user-friendly and is completely described on-line. Extensive printed documentation is available.

SEEDIS, as part of a publicly funded research program, is in the public domain. The State Data Program and Survey Research Center (SDP/SRC) on the University of California's Berkeley campus can provide standard Census data reports and more specialized data extraction services at cost. In the near future, it is hoped that such services, as well as tape copies of SEEDIS for installation on a DEC VAX-11/780 computer, will be available through the National Energy Software Center and the National Technical Information Service.

Even for installations not having a DEC VAX computer, portions of SEEDIS will be useful in other geographic information systems. The geocode description files, the cross-level correspondence files, and the proxy variable files used for aggregation and disaggregation are all stored in self-describing ASCII format. Data dictionaries are available for major public data files, including the 1974 Census of Agriculture, the 1977 County and City Data Book, the 1977 Area Resource File, and the 1980 Census Summary Tape File 1. Geographic base files used for mapping, although in a VAX binary format, can be readily converted with a simple FORTRAN program to any desired format.

Further information about SEEDIS, and copies of a 26-page summary overview document [3] can be obtained from: Ilona Einowski, Data Librarian; SDP/SRC, 2538 Channing Way; University of California; Berkeley CA 94720; Tel. (415) 642-6571.

REFERENCES

1. Olson, R.J., Emerson, C.J., and Nungesser, M.K.; Geoecology: A County-Level Environmental Data Base for the Conterminous United States; Oak Ridge National Laboratory Report ORNL/TM-7351, September 1980.

2. Sacks, S.T., Selvin, S., and Merrill, D.W.; Building a United States Data Base: Populations at Risk to Environmental Pollution; Lawrence Berkeley Laboratory Report LBL-9636; October 1979.

3. McCarthy, J.L., Merrill, D.W., Marcus, A., Benson, W.H., Gey, F.C., Holmes, H., and Quong, C.; SEEDIS: The Socio-Economic Environmental Demographic Information System; Lawrence Berkeley Laboratory Report PUB-424; September 1981 (draft).

4. Milask, L.J., Graves, C.G., and Feldman, M.L.; UPGRADE User's Manual (version 4.0); Sigma Data Computing Corporation, January 1980.

5. Zimmerman, E.K.; The Evolution of the Domestic Information Display System: Toward a Government Public Information Network; Review of Public Data Use Vol. 8, No. 1; June 1980.

6. Ripley, B.D.; Spatial Statistics; John Wiley and Sons, 1981.

7. Selvin, S., Merrill, D., Kwok, L., and Sacks, S.; Ecologic Regression Analysis and the Study of the Influence of Air Quality on Mortality; Lawrence Berkeley Laboratory Report LBL-12217; October 1981; submitted to Journal of Environmental Economics and Management.

8. Selvin, S., Sacks, S., Merrill, D.W., and Winkelstein, W.; The Relationship Between Cancer Incidence and Two Pollutants (Total Suspended Particulate and Carbon Monoxide) for the San Francisco Bay Area; Lawrence Berkeley Laboratory Report LBL-10847; June 1980.

TABLE 1.
GEOGRAPHIC LEVELS IN SEEDIS

Level	Level Description	Year	Units	Map
	Nation and Large Interstate levels			
BEA	Bureau of Economic Analysis regns	1977	9	p
BECHT	Bechtel energy model regions	1978	14	p
CENSUS	Census regions	1970	9	p
COAL	Coal supply regions	1978	12	-
FED	Federal regions	1970	10	p
NATION80	Nations, FIPS definitions	1980	233	p
NPC	Nat Petrol Council oil & gas regns	1976	12	-
PAD	Petr alloc district refinery regns	1978	7	p
WATER	Water Resources Council regions	1972	22	p

TABLE 1, CONTINUED
GEOGRAPHIC LEVELS IN SEEDIS

Level	Level Description	Year	Units	Map
	Small Interstate Levels			
AQCR	EPA air quality control regions	1972	247	p
BEA69	Bureau of Economic Analysis areas	1969	173	p
BEA77	Bureau of Economic Analysis areas	1977	183	p
LMA	BLS labor market areas	1979	437	p
NECMA77	New England county metro areas	1977	270	p
NECMA79	New England county metro areas	1979	275	p
NECMA81	New England county metro areas	1981	310	p
PUS70	Public Use Sample county groups	1970	408	p
SCSA79	Standard consolidated stat areas	1979	13	-
SCSA81	Standard consolidated stat areas	1981	17	-
SMSA71	Standard metropolitan stat areas	1971	246	s
SMSA73	Standard metropolitan stat areas	1973	267	s
SMSA75	Standard metropolitan stat areas	1975	276	s
SMSA77	Standard metropolitan stat areas	1977	283	-
SMSA79	Standard metropolitan stat areas	1979	288	s
SMSA81	Standard metropolitan stat areas	1981	323	-
WRASA	Water Res Council aggreg subareas	1972	106	p
WRSA	Water Resources Council subareas	1972	222	-
	State Parts of Interstate Levels			
BEAPART77	STATE/BEA77/NECMA77 pieces	1977	585	p
STAQCR	STATE/AQCR pieces	1972	314	p
STBEA69	STATE/BEA69 pieces	1969	290	p
STBEA77	STATE/BEA77 pieces	1977	285	p
STLMA	STATE/LMA pieces	1979	432	p
STNECMA77	STATE/NECMA77 pieces	1977	363	p
STNECMA79	STATE/NECMA79 pieces	1979	368	p
STNECMA81	STATE/NECMA81 pieces	1981	404	p
STPUS70	STATE/PUS70 pieces	1970	535	p
STSCSA81	STATE/SCSA81 pieces	1981	83	-
STSMSA71	STATE/SMSA71 pieces	1971	334	s
STSMSA73	STATE/SMSA73 pieces	1973	365	s
STSMSA75	STATE/SMSA75 pieces	1975	373	s
STSMSA77	STATE/SMSA77 pieces	1977	381	-
STSMSA79	STATE/SMSA79 pieces	1979	386	s
STSMSA81	STATE/SMSA81 pieces	1981	423	-
STWATER	STATE/WATER pieces	1972	110	p
STWRASA	STATE/WRASA pieces	1972	237	p
STWRSA	STATE/WRSA pieces	1972	368	-
	State and Large Intrastate Levels			
CD86	86th Congr Congressional Districts	1959	436	-
CD91	91st Congr Congressional Districts	1969	436	-
CD94	94th Congr Congressional Districts	1975	436	p
CD96	96th Congr Congressional Districts	1979	436	p
CD98	98th Congr Congressional Districts	1983	436	-
LMPM	Labor Market Proj Model areas	1979	1280	-
PRSP80	Bureau Labor Stat. Prime Sponsors	1980	469	-
SEA70	State economic areas	1970	510	p
STATE	States and territories	1970	55	p

TABLE 1, CONTINUED
GEOGRAPHIC LEVELS IN SEEDIS

Level	Level Description	Year	Units	Map
	County Levels			
CNTY7080	COUNTY/COUNTY80 pieces	1980	3265	p
COUNTY	Counties, U.S. Census Bureau	1970	3255	p
COUNTY80	Counties, U.S. Census Bureau	1980	3253	p
MSP	Counties, Johns Hop Mort Surv Prog	1970	3075	p
NCHS	Counties, Nat Center Health Stat	1970	3082	p
NCI	Counties, National Cancer Inst	1970	3061	p
	Place and MCD Levels			
MCD70	Census Minor Civil Divisions	1970	35198	s
MCD80	Census Minor Civil Divisions	1980	35197	-
MCDPL80	MCD80/PLACE80 pieces	1980	53032	-
PLACE	Places, Census, pop over 1000	1970	11970	s
PLACE80	Places, Census	1980	19144	-
PLBLS	Places, Bureau of Labor Statistics	1979	1565	-
PLEPA	Places, Env Protection Agency	1972	9745	s
TOWNSHIP	Bureau Labor Statistics townships	1979	112	-
	Small Subcounty Levels			
AQMS	EPA air qual monitoring stations	1976	6625	s
AQMSLOC	EPA air qual monitoring locations	1976	5777	s
BGED70	Census enum dists & block groups	1970	249189	s
EDBG80PT	EDBG80/URBARURL/CD96 pieces	1980	320000	-
PREDBG80	EDBG80/CD96 pieces	1980	319105	-
TRACT	Census Tracts	1970	34869	p
TRACT80	Census Tracts	1980	48475	-
TRACT80PT	MCD80/PLACE80/TRACT80 pieces	1980	75000	-

TECHNICAL INFORMATION DEPARTMENT
LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720