September 1981 PUB-424

SEEDIS

The Socio-Economic-Environmental Demographic Information System

RECEIVED
LAWRENCE
BERKELEY LABORATORY
FEB 5 1982
LIBRARY AND
DOCUMENTS SECTION

DRAFT

For Reference

Not to be taken from this room

Computer Science and Mathematics Department
Lawrence Berkeley Laboratory
University of California, Berkeley

This work was supported by the U.S.Department of Energy under Contract W-7405-ENG-48.

PUB-424 c.1

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

THE SEEDIS PROJECT: A SUMMARY OVERVIEW

John L. McCarthy
Deane W. Merrill
Aaron Marcus
William H. Benson
Fredric C. Gey
Harvard Holmes
Carl Quong

DRAFT

Computer Science and Mathematics Department
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

September 1981
PUB-424

ACKNOWLEDGMENTS

Many people and organizations have contributed to the SEEDIS Project during the past nine years. While space does not permit listing all the individuals who have participated, we would like to acknowledge major support from the following agencies:

U.S. Department of Energy
U.S. Department of Labor, Employment and Training Administration
U.S. Army Corps of Engineers (Civil Works Planning)
U.S. Environmental Protection Agency
U.S. Bureau of the Census
U.S. Department of Housing and Urban Development

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

DISCLAIMER

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the authors and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

TABLE OF CONTENTS

1 INTRODUCTION
1 SEEDIS Project Goals and Purposes
2 History and Background
2 FEATURES
4 Data Selection and Manipulation
4 Data Entry and Loading Facilities
6 Distributed Network Facilities
7 Interactive Graphic Analysis and Display
7 User Interface
8 Data and System Documentation
8 DATABASES
10 Special Facilities for Large Numeric Databases
10 GEOGRAPHIC AREAS AND MAP FILES
12 FUTURE PLANS
13 AVAILABILITY AND SOURCES OF FURTHER INFORMATION
14 SELECTED REFERENCES

EXHIBITS

3 Exhibit 1: Schematic View of the Major Functional Components of SEEDIS
5 Exhibit 2: Examples of Geographic Level and Scope in SEEDIS
6 Exhibit 3: The Distributed Computer Network, 1981
9 Exhibit 4: Summary of Major SEEDIS Databases as of September 1981
11 Exhibit 5: Major Geographic Levels Defined in SEEDIS as of September 1981

APPENDICES

16 Appendix A: Example of How to Use SEEDIS -- an Annotated Interactive Dialogue
21 Appendix B: Interactive SEEDIS Chart Examples
22 A Special Fuzzy Bar Chart
22 A Special Line Chart Variation
23 Standard Chart Examples
24 Appendix C: Interactive SEEDIS Mapping Examples
24 An Area Choropleth Map
25 A Bivariate Area Choropleth Map
26 A Symbol Choropleth Map

INTRODUCTION

SEEDIS is a research and development project on Social, Economic, Environmental, and Demographic Information Systems at the Lawrence Berkeley Laboratory (LBL), supported by the Department of Energy, Department of Labor, and others. The SEEDIS project includes:

- a research program within LBL's Computer Science and Mathematics Department to investigate information systems spanning diverse data sources, computer hardware and operating systems
- a testbed distributed information system running on a network of Digital Equipment Corporation (DEC) VAX computers, which is used for selected applications as well as research and development
- a set of interactive information management and analysis tools in fields such as energy and resource planning, employment and training program management, and environmental epidemiology
- a major collection of databases for various geographic levels and time periods, drawn from the U.S. Census Bureau and other sources

SEEDIS Project Goals and Purposes

Policy formulation, implementation, and management depend upon accurate, timely information. Analysts, decision makers, and managers need to locate, retrieve, combine, analyze and display information from a variety of sources. While time and resources usually do not permit collecting new information, there is a wealth of publicly available data (from government and other sources) that often could meet such needs if it were quickly and easily accessible. Although computers and machine-readable data have made it potentially easier to locate and analyze information, use of information from different sources is difficult because of differing storage formats, communications procedures, coding conventions, data structures and units of analysis.

The SEEDIS Project addresses these information needs and problems through research, design, development, and demonstration of information system components. SEEDIS software provides a unified framework for data management, information retrieval, statistical analysis, and graphical display. Using SEEDIS, non-programmer users can efficiently access and manipulate large, diverse, and distributed statistical databases. In some of these respects, SEEDIS is similar to systems such as UPGRADE, developed by the Council on Environmental Quality [COUN 80], and the Decision Information Display System (DIDS), developed jointly by the National Aeronautics and Space Administration (NASA) and the Department of Commerce [DALT 79, DECI 81].

Unlike most statistical information systems, SEEDIS is a testbed for different functional components, including software developed at LBL and elsewhere. An underlying file interchange format and command language interpreter provide a "software bus" for interchange of data and data descriptions among a variety of storage and access methods, search and retrieval tools, display and analysis facilities, and user interface environments. Users need not be concerned with the detailed structure or operating requirements of individual system components, and there are a number of advanced and experimental facilities available for those who wish to use them.

The SEEDIS project works with selected applications specialists to

- test new concepts and tools using large "real world" databases
- get feedback about how information tools can be improved
- acquire and develop additional databases

History and Background

Although the VAX version of SEEDIS is only two years old, the experience embodied in its development spans nearly ten years. In 1972, the Department of Labor requested the Lawrence Berkeley Laboratory to apply its expertise with large databases from accelerator experiments to the development of storage, retrieval, and report generating software for 1970 United States Census data. This effort led to development of software on LBL's CDC computers for interactive access to a growing collection of databases. Subsequent developments included tools for mapping and graphic display [AUST 75, GEY 75, WOOD 78A], a file interchange format, and a command language "monitor" to link the evolving subsystems [KREP 77].

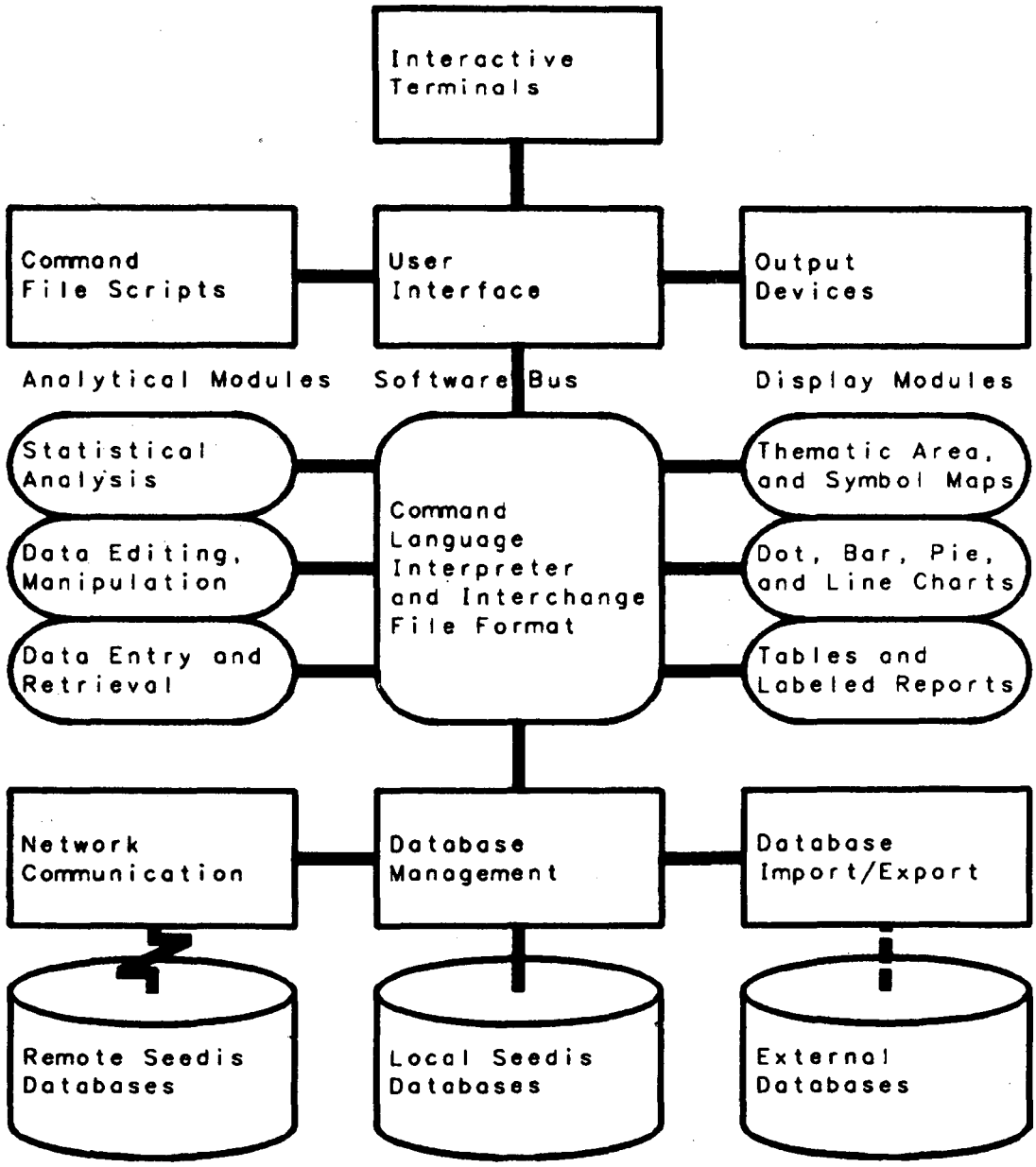
Work on the integrated VAX version of SEEDIS began in 1979, and efforts are currently underway on major enhancements necessary for incorporation of 1980 census data [CSAM 81, COMP 81, GEY 81, MARC 81, MERR 81A].

FEATURES

Exhibit 1 presents the logical structure of SEEDIS, with its underlying file interchange format, unified user interface, and various functional modules. Major features of SEEDIS, as elaborated in the subsections below, include the following:

- flexible selection and manipulation of data
- export and import of data to and from external sources
- convenient data entry and loading facilities
- distributed operation over a network of VAX computers
- interactive color chart and map making facilities
- a "user-friendly" human-computer interface
- comprehensive printed and on-line documentation
- efficient handling of large numeric databases

Exhibit 1: Schematic View of the Major Functional Components of SEEDIS



Data Selection and Manipulation

SEEDIS includes powerful and unique features for selecting

- level of analysis, such as tracts, counties, or over sixty other levels, as described in the section below on geographic areas and map files
- scope of analysis, specific units or entities within a level (e.g., all counties in Federal Region IX, selected tracts in New York City, etc.)
- data items of interest (e.g., number of Hispanic families with annual income less than $3000, death rate from leukemia, average concentration of suspended particulate in 1976, etc.)

Exhibit 2 illustrates the concepts of geographic level and scope for a few of the major geographic areas currently implemented in SEEDIS. For example, a user could request a county unit of analysis (the next to last column pictured) with a scope of either the Eastern United States, District of Columbia SMSA, or District of Columbia (the last three rows pictured, respectively). The sample SEEDIS-user dialogue in Appendix A shows actual commands used to select state level data for several states.

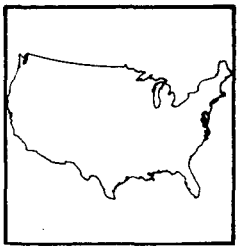
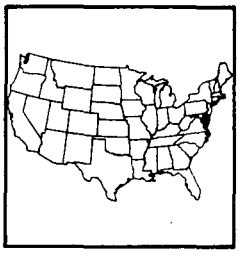
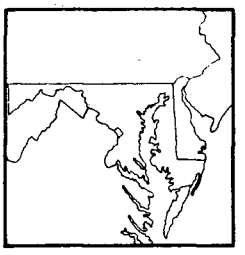
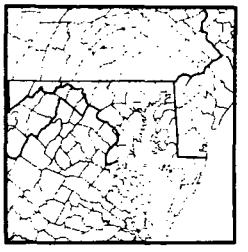
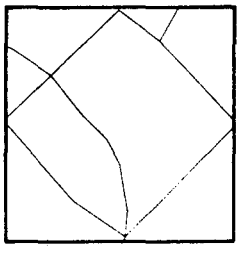
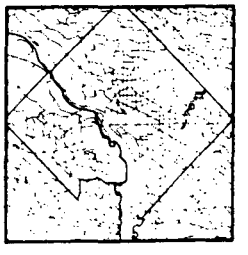
Other SEEDIS facilities enable users to manipulate data and entities to which the data pertain in a variety of ways, including the following:

- create new data items, sets of entities, etc. using logical and arithmetic functions
- automatically aggregate, disaggregate, interpolate and integrate data from different geographic levels into a single analysis file, as explained in the section on geographic areas and map files below
- select data and particular geographic entities on the basis of data item values (e.g., census tracts in which the proportion of housing units with oil heat exceeds fifty percent)
- produce self-documenting, machine-independent formatted data files for use in subsequent SEEDIS sessions or software external to SEEDIS

Data Entry and Loading Facilities

Many users want to use SEEDIS to manipulate and display their own data, sometimes in combination with data from archived SEEDIS files. For small files, a SEEDIS data entry subprogram prompts users for data input and automatically stores it in the correct format for further analysis. For larger datasets, knowledgeable users can convert and manipulate 'most external formatted files via the self-describing "CODATA" format used by SEEDIS [MERR 81A]. Files needed by more than one user can be permanently installed in SEEDIS at any node on the Distributed Computer Network [MERR 81B].

Exhibit 2: Examples of Geographic Level and Scope in SEEDIS

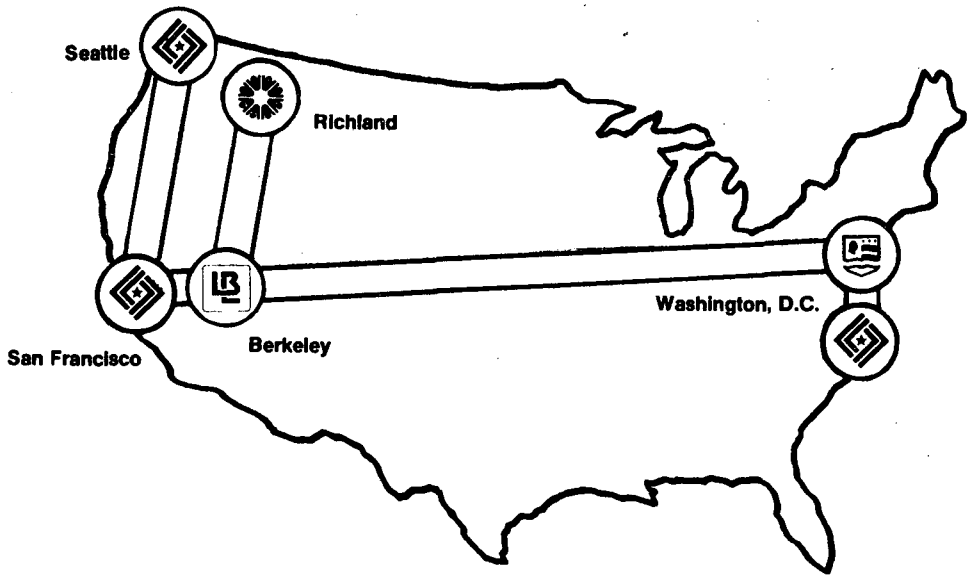
SCOPE	LEVEL			
	NATION	STATE	COUNTY	TRACT
WORLD				
U.S.				
EASTERN U.S.				
D.C. AREA				
D.C.				

Distributed Network Facilities

SEEDIS gives users shared access to a wide variety of data stored in different physical locations, while permitting each local facility to maintain control over its own databases. Underlying network facilities provided by DECNET [DIGI 79] enable users at any node to access databases anywhere on the network. Except for response time, SEEDIS appears and behaves as if all the data were stored locally. A future goal is to provide distributed data management, retrieval and analysis capabilities over networks of heterogeneous computer systems.

SEEDIS presently operates in a network of eight DEC VAX computers. This distributed computer network (DCN), which may grow with the addition of new nodes and which is linked to other networks such as ARPANET, is pictured below in Exhibit 3. Another SEEDIS system installed at the Brookhaven National Laboratory can access the network via ordinary telephone lines. Each of the current nine SEEDIS installations have 30 megabytes of SEEDIS program modules and data description files physically resident on a local disk pack. The main set of prototype data (including geographic base map files) is stored on disks connected to two VAXes at LBL.

Exhibit 3: The Distributed Computer Network, 1981



Department of Energy Research Laboratories

- **Berkeley:** Lawrence Berkeley Laboratory
- **Richland:** BATTELLE Pacific Northwest Laboratories
- **Washington, D.C.:** George Washington University

Department of Labor, Employment & Training Administration

- **Seattle:** Federal Region X Office
- **San Francisco:** Federal Region IX Office
- **Washington, D.C.:** National Office

Interactive Graphic Analysis and Display

For nearly a decade, development of high quality computer graphics facilities for data analysis and presentation has been a major thrust of the SEEDIS Project. SEEDIS mapping facilities have been used for major cartographic publications such as the Urban Atlas, which was a joint effort of LBL, the Department of Labor, and the United States Bureau of the Census [UNIT 74]. VAX SEEDIS incorporates a number of capabilities from the forefront of computer graphics research as well as standard facilities, including:

- production of graphic displays on a variety of standard monochromatic and color devices (including inexpensive devices that can be used over ordinary telephone lines)
- standard and custom labeled tables, bar charts, pie charts, line graphs, scatterplots, and other graphic output [BENS 77, EADE 81] (see Appendix B for examples)
- polygon (choropleth) and symbol mapping for predefined geographic entities [YEN 79A, YEN 79B] (see Appendix C for examples)
- color maps and charts, including bivariate displays [TRUM 80] and "fuzzy graphics" for exploring data in visual ways that may reveal unexpected patterns [BENS 81]

Interactive SEEDIS facilities give users the option of producing either standard default graphics with a single command, or highly customized output with user-specified labeling, shading, color selection, and positioning within the frame. Users can produce insets with automatic user specified magnification of map segments and multiple charts within a single output frame.

User Interface

The SEEDIS project has devoted considerable effort to designing consistent vocabulary, layout, and sequencing of system dialogue and display in order to improve the effectiveness of SEEDIS from a human-factors point of view [MARC 81]. A sample human-computer dialogue appears in Appendix A below. Current features of the SEEDIS user interface include:

- interactive operation with menu prompting (i.e., lists of command options from which the user can choose)
- on-line help, status reports, and explanation of commands
- searching aids, including on-line browsing of data dictionaries
- standard report formats (profiles) for user-defined sets of areas
- optional batch mode operation for long jobs or sample sessions
- automatic logging of all user commands, producing scripts which can be used for diagnosing problems, analyzing usage patterns, and creating batch procedures
- a comment facility which permits users to log their complaints and suggestions to SEEDIS developers "in context" during the course of an interactive session

Data and System Documentation

SEEDIS has comprehensive documentation, ranging from interactive information on data and commands for users to technical system documentation for programmers. Users can browse documentation interactively on-line or consult a printed version. Every file installed in SEEDIS has an on-line "data dictionary", containing not only a precise definition of every data element, but also details of acquisition, authors' addresses and phone numbers, citations of relevant documents, warnings about any known errors or omissions, location of source files, and specific procedures used to convert the file or calculate new data elements. A table of contents in each dictionary quickly directs the user to sections of interest, and dictionaries can be printed for continuing reference.

DATABASES

In connection with various applications projects since the early 1970's, the SEEDIS Project has acquired and developed databases containing roughly 2.5 billion individual data values for some 1 million different data items [BURK 79B]. (Data items are variables such as the number of unemployed persons, or individual cells of a table such as unemployment by age, race, and sex; data values are the numeric values of such items for individual entities such as states, counties, or households.) Excluding duplicate copies, these databases occupy approximately 25 billion bytes of storage, primarily in compressed form on some 250 high density (6250 bpi) magnetic tapes. In printed form, the data would occupy over 10 million pages. With addition of data from the 1980 census and other sources, the collection will more than double in the next two or three years.

Exhibit 4 outlines the current inventory of major SEEDIS databases, along with summary information about database contents, sizes and geographic levels. Each data item is available for one or more distinct types of entities or geographic levels (e.g. household, state, county, etc.). Each SEEDIS database contains information for one or more sets of comparable entities defined at the same level (e.g., 1980 census counties, or 1979 Standard Metropolitan Statistical Areas).

A prototype subset of the most frequently accessed data (representing about 2 percent of the total inventory) presently is stored on disk for rapid access from the VAX version of SEEDIS. This prototype set of data contains some 54 million individual data values for approximately 26,000 different data items. It includes over 19,000 different data items for each state, 12,000 data items for each of the roughly 3000 counties in the United States, and 24,000 data items for other geographic levels.

Most of the largest SEEDIS data files, including 1970 census data for small areas such as enumeration districts, block groups, tracts, minor civil divisions and census county divisions, formerly were stored on a random access photodigital "chipstore" storage device [GEY 75]. In 1979, when IBM discontinued maintenance of the chipstore, those data were transferred to high density magnetic tapes.

Exhibit 4: Summary of Major SEEDIS Databases as of September 1981

Database Description	Year(s)[1]	Items[2]	Levels[3]	Million Values[4]
SELECTED VAX SEEDIS DATABASES ON DISK		**26,018**		**53.8**
Social and Demographic Characteristics				
PL94 Reapportionment Data	80	13	nsmcdptb	4.0
County Data Book	47-77	1,022	sc	3.3
Census 4th Count, by race	70	5,890	sm	1.7
Population by age, race, sex	70-77	608	nrsc	4.4
Economy and Employment				
BLS Labor Force and Unemployment	74-79	72	smo	0.1
BEA Economic Projections	69-2030	1,820	nrsm	0.6
Census of Agriculture	74	1,200	sc	3.8
Employment by Industry	71-76	284	sc	0.9
Energy and Environment				
Air Quality Indicators	74-76	500	cto	2.3
Electric Generating Capacity	60-95	18	nsmrc	0.1
Epidemiology and Health				
Leukemia Mortality	69-77	960	nrsc	4.0
Age Adjusted Cancer Mortality	50-69	424	c	1.3
Cancer Incidence	69-71	352	t	1.5
Age Adjusted Mortality	68-76	1,816	nrsc	6.3
Area Resource File	77	889	c	1.3
Other Databases on Disk	**Misc**	**10,151**	**Misc**	**19.2**
MAJOR SEEDIS GETTAPE-STOTAPE SYSTEM FILES		**17,742**		**1,104**
1970 Census of Population				
First Count Special Extract	70	150	dtb	46
Second Count	70	3,500	smcdpt	280
Fourth Count	70	6,600	smcdpt	485
Fifth Count Special Extract	70	900	dtb	275
Other Censuses Data				
Censuses of Agriculture	49-74	2,400	c	7
Census of Manufactures	72	2,200	c	7
1960 Census of Population	60	1,000	c	3
Consolidated City Data Book	47-77	992	p	1
OTHER MAJOR TAPE FILES IN EXTERNAL FORMATS		**951,220**		**1,335**
United States Census Bureau				
1970 Sixth Count	70	150,000	smcp	195
1970 Public Use Samples	70	400	ih	200
County Business Patterns	64-77	30	nsc	40
Current Population Surveys	70-80	300	ih	300
Employment by Industry and Occupation	70	800,000	ns	100
Survey of Income and Education	76	490	ih	100
Other Miscellaneous Databases				**400**

1. For information on specific years and items available for each, see [BURK 79B, COMP 81].
2. Distinct data cells or variables for summary data; items are not double counted for multiple levels.
3. Major types of entities to which data pertain, coded from largest to smallest as follows:

n nation	m stnd metro stat areas	p places	h households
r regions	c counties	t tracts	i individuals
s states	d minor civil divisions	b block groups	o other

4. Millions of individual data values (variables or cells per entity times number of entities).

Special Facilities for Large Numeric Databases

Because storage costs and access time can escalate for large databases, the SEEDIS project has devoted considerable attention to methods for efficient compression, storage, and retrieval of numeric data. SEEDIS currently employs a special LBL computer-independent binary storage technique based on run length encoding of zeros and missing data, which typically compresses data to about one-third of the original number of characters [MERR 81A]. New techniques developed by LBL staff may yield even greater storage and retrieval efficiency. [EGGE 81]

Long-range SEEDIS development plans call for fast, interactive access to the complete set of databases using optical disk mass storage devices and distributed data management techniques. In the meantime, the project is developing a network link to the LBL Computer Center's tape robot and Gettape-Stotape System (GSS) to provide automatic access to the entire collection of tapes for interactive extraction of moderate size subsets of data (up to 10 million characters in average times of 2 to 30 minutes).

GEOGRAPHIC AREAS AND MAP FILES

Data collected by different agencies usually are not directly comparable. Because of this common problem, SEEDIS has developed unique data conversion capabilities as well as mapping for different sets of geographic area definitions.

SEEDIS presently includes definitions for over 60 different types of entities to which data pertain. Most of these represent different geographic levels, from nations down to census tracts and block groups. Some represent slightly different versions of the same basic type of entity, such as counties as defined by five different agencies or Standard Metropolitan Statistical Areas as defined in six different years. Even small differences must be considered in order to properly compare data across time or agencies. New levels are defined as needed to accommodate new files.

To facilitate combination of data from different levels of analysis, special geographic files in SEEDIS define each geographic unit in terms of the larger entities of which it is a part. For example, every county (1970 census definition) is identified as belonging to a particular EPA Air Quality Control Region, a particular Bureau of Economic Analysis Area, etc. Where necessary, counties are divided into smaller undivided units whose assignment to larger areas is uniquely defined. Exhibit 5 summarizes the major SEEDIS geographic levels, the number of individual entities (areas) in each, type of mapping available, and related data holdings on disk and tape. SEEDIS also provides facilities for users to browse such lists interactively for easy reference.

Exhibit 5: Major Geographic Levels Defined in SEEDIS as of September 1981

Geographic Level Description	Year(s)	Units [1]	Map [2]	VAX DISK FILES Items	VAX DISK FILES KValues[3]	TAPE FILES KValues[3]
International						
Nations, FIPS definitions	80	223	p	6,271	6	
Large Interstate						
Bechtel energy model regions	78	14	p	18	<1	
Federal regions	78	10	p	18	<1	
Census regions	78	9	p	18	<1	
Coal supply regions	78	12	-	0		
Bureau of Economic Analysis regions	77	9	p	560	4	
Petr. alloc. district refinery regions	78	7	p	18	<1	
Water Resources Council regions	72	22	p	0		
Standard consolidated stat. areas	79	13	-	0		
Small Interstate						
Standard metropolitan stat. areas	71-81	323	s	5,932	1,443	43,000
New England county metro. areas	77,81	310	p	577	154	
Bur. Labor Statistics labor mkt. areas	79	437	p	72	31	
Water Resources Council subareas	72	222	-	0		
Bureau of Economic Analysis areas	69,77	183	p	598	110	
Public Use Sample county groups	70	408	p	1,586	647	
EPA air quality control regions	72	247	p	1,970	479	
State and Large Intrastate						
States and territories	70	55	p	19,504	996	111,000
Congressional Districts	62-82	436	-	0		
Labor Market Projections Model areas	79	1,280	-	470	600	
Bureau Labor Stat. Prime Sponsors	80	469	-	72	28	
State economic areas	70	510	p	2,274	1,160	
Single-state portions of interstate areas		various	v			
County						
Counties, U.S. Census Bureau	70	3,255	p	8,220	23,513	172,000
Counties, U.S. Census Bureau	80	3,253	p	21	57	
Counties, Nat Center Health Statistics	70	3,082	p	2,256	6,953	
Counties, National Cancer Institute	70	3,061	p	0		
Counties, Johns Hopkins Mort Surv Prog	70	3,075	p	1,568	4,822	
COUNTY70/COUNTY80 pieces	80	3,265	p	18	57	
Place						
Places, Environmental Protection Agency	72	9,745	s	0		
Census places with population over 1000	70	11,970	s	7	143	178,000
Census places	80	19,144	-	18	344	
Places, Bureau of Labor Statistics	79	1,565	-	0		
MCD80/PLACE80 pieces	80	53,032	-	15	665	
Other Subcounty						
Census Minor Civil Divisions	70	35,198	s	23	671	390,000
Census Minor Civil Divisions	80	35,197	-	17	598	
Census Tracts	70	34,869	p	828	2,994	390,000
Census Tracts	80	48,475	-	17	815	
Census enumeration dists & block groups	70	249,189	s	7	1,744	247,000
Precinct parts of census ED/BG's	80	319,105	-	7	2,234	
EPA air quality stations active 74-76	72	6,625	s	200	1,325	

1. number of individual units (entities or areas) in the level; numbers may differ from year to year
2. p=polygon outlines; s=symbol point locations; - = none; v = depends on level
3. thousands of data values--numeric values of items for individual units

SEEDIS can automatically combine data from different geographic levels for display and analysis, so a data file need only be stored at its most detailed geographic level if data are complete for every geographic entity and there is no suppression. For additive data like population, aggregation from small to large areas is straightforward; disaggregation from large to small areas assumes proportionality with some user-specified proxy variable. For non-additive data like average family size, aggregation requires a user-specified weighting variable (e.g., number of families); disaggregation assigns the value from the larger area to all smaller units [MERR 80]. Without such capabilities, other systems and users usually have to resort to methods that involve questionable assumptions, loss of detailed information, or both.

For mapping purposes, SEEDIS includes a set of cartographic base map files, essentially one for each type of geographic entity [BURK 79A]. These include the complete set of 1970 census tract base maps (available from the Census Bureau), which were originally produced at LBL. Each geographic entity within a given level is associated with a series of latitude-longitude pairs which define a polygon representing its boundaries. Some polygons are aggregates of county polygons; others, corresponding to subcounty areas, were carved out of county polygons. Point locations such as air quality monitoring stations are identified by a single coordinate pair. Map files are archived in latitude-longitude coordinates, in order to permit overlaying of different geographic entities. Projection (e.g. to conic coordinates) is performed at run time as required for display purposes.

FUTURE PLANS

SEEDIS will continue as a research project and experimental information system, with major enhancements to facilitate incorporation of 1980 census data. Current plans for future developments include:

- facilities to describe and handle complex data structures, including hierarchical records, multiply occurring data items and repeating groups of items, variable length data, transposed (vertically partitioned) and partially transposed data records
- automatic creation of metadata information such as dictionaries and cross-database indexes from augmented data definition files
- selection of data items by keyword, subject, and/or item values
- statistical routines that automatically read and write an underlying self-describing file structure so that the output of any routine can immediately serve as input for any other routine
- a "report by example" subprogram to create standard and custom reports which can be subsequently run with either lineprinter or phototypesetter output for any combination of geographic areas
- integration of user interface functions to facilitate experimentation with pictorial and touch-sensitive interfaces, different input and output formats for different types of devices (e.g., dumb crt's, printers, and graphics terminals), flexible invocation of any SEEDIS function from within another, and user selection of command, menu, or question and answer mode

AVAILABILITY AND SOURCES OF FURTHER INFORMATION

SEEDIS is currently used by the United States Department of Labor, Department of Energy, Environmental Protection Agency and Army Corps of Engineers. Other organizations or individuals interested in using the system have several alternatives, as follows:

- The National Technical Information Service prepares standard reports based on 1970 census data for user-designated census areas or aggregations thereof. For information, write or call:

  Marvin Wilson, NTIS
  5285 Port Royal Road
  Springfield, VA 22161
  (703) 487-4805
  (FTS) 737-4805

- The State Data Program/Survey Research Center on the University of California's Berkeley campus provides standard reports similar to those of NTIS as well as more specialized data extraction services at cost. For information, write or call:

  Ilona Einowski, Data Librarian
  SDP/SRC 2538 Channing Way
  University of California
  Berkeley, CA 94720
  (415) 642-6571

- In the future, VAX SEEDIS itself will be made available for distribution through the National Technical Information Service. Organizations interested in installing SEEDIS can contact:

  Harvard Holmes, SEEDIS Project
  Computer Science and Mathematics Dept
  Lawrence Berkeley Laboratory
  Berkeley, CA 94720
  (415) 486-5181
  (FTS) 451-5181

For further written information, see the references listed below.

SELECTED REFERENCES

AUST 75 Austin, D.M., Kranz, S.G., and Quong, C.: "An Overview of the LBL Socio-Economic Environmental Demographic Information System (SEEDIS)," Lawrence Berkeley Laboratory Report LBL-3699 (March, 1975)

BENS 77 Benson, W.H.: "Interactive Analysis and Display of Tabular Data," 2 Computer Graphics 2 (Summer, 1977), 48-53

BENS 81 Benson, W.H.: "An Application of Fuzzy Set Theory to Data Display" in R.R. Yager (ed.), Recent Developments in Fuzzy Set and Possibility Theory (Pergamon Press, forthcoming 1981)

BURK 79A Burkhart, B.R., ed., "Cartographic Base Files at Lawrence Berkeley Laboratory: 1978 Inventory," LBL-8707, January 1979

BURK 79B Burkhart, B. and Merrill, D., eds.: Spatial Data on Energy, Environmental and Socioeconomic Themes at Lawrence Berkeley Laboratory: 1978 Inventory; LBL-8744, UC-13, April 1979

COMP 81 Computer Science and Mathematics Department, "SEEDIS Release Notes version 1.2," June 1981; version 1.3 (preliminary) October 1981

COUN 80 Council on Environmental Quality, UPGRADE User's Manual (version 4.0), Washington, D.C., 1980

DALT 79 Dalton, Billingsley, Quann, and Braker; "Interactive Color Map Displays of Domestic Information," 13 Computer Graphics 2 (August, 1979) 226-233

DECI 81 Decision Information Display System, Office of Federal Statistical Policy and Standards, Department of Commerce, User's Guide, Washington, DC, September 1980

DIGI 79 Digital Equipment Corporation, DECnet-VAX User's Guide, Maynard, Massachusetts 1979

EADE 81 Eades, C., "CHART: A Graphic Display and Analysis System, A User's Guide," PUB-3015, 1981

EGGE 81 Eggers, S., Olken, F., and Shoshani, A., "A Compression Technique for Large Statistical Databases," LBL-12353, February, 1981

GEY 75 Gey, F. and Mantei, M. "Keyword Access to a Mass Storage Device at the Record Level," Proceedings of the First International Conference on Very Large Databases, Framingham, Massachusetts, September 1975

GEY 81 Gey, F., "A Beginner's Guide to SEEDIS," LBL-11198, January 1981

HOLM 78 Holmes, H., "The LBL Socio-Economic-Environmental Demographic Information System," Proceedings of the Workshop on Information

Systems for Coastal Zone Management, Center for Ocean Management Studies, University of Rhode Island, June 1978, pp. 77-83

KREP 77 Kreps, P. and Sventek, V., "SEEDIS Monitor," LBL-6440, 1977

MARC 81 Marcus, A., "SEEDIS Graphic Design Manual" (in preparation)

MERR 80 Merrill, D., "Handling Spatial Data in SEEDIS (Socio-Economic Environmental Demographic Information System)," presented at the Integrated County-Level-Data User's Workshop, Reston VA, October 15-16, 1980 National Technical Information Service publication CONF-8010139

MERR 81A Merrill, D., "CODATA Users' Manual," LBID-021, revised April 1981

MERR 81B Merrill, D., "Installing Data in SEEDIS," LBL internal documentation, revised March 1981

TRUM 80 Trumbo, B., "A Theory for Coloring Bivariate Statistical Maps," Technical Report No. 44, Study on Statistics and Environmental Factors in Health, Department of Statistics, Stanford University, 1980

UNIT 74 United States Bureau of the Census and Manpower Administration, Urban Atlas, Washington, D.C. 1974

WOOD 78A Wood, P. and Benson W., "Computer Mapping Software at the Lawrence Berkeley Laboratory," LBL-7938, 1978

WOOD 78B Wood, P., Drysdale, F., and Calef, C., "Using Computer-assisted Cartography to Map the U. S. Energy System," LBL-7937, June 1978

YEN 79A Yen, A., "VAX/CARTE User's Manual," May 1981

YEN 79B Yen, A. and Wood, P., "Moving Interactive Thematic Mapping from Mainframe to Mini: Some Design Possibilities and Development Experience", AUTO Carto IV Proceedings, vol. II (1979), 379-386

Appendix A Example of How to Use SEEDIS -- An Annotated Interactive Dialogue

The following is a brief example of a simple SEEDIS retrieval and display session. All of the dialogue in the non-proportional type font appears just as it would on a user's screen or hardcopy terminal. Boxed annotations point out features of SEEDIS which appear immediately following in the dialogue. In general, all SEEDIS user commands appear at the left side of the terminal input/output frames, following the colon prompts (:), while computer responses begin at the next major tab stop (column 20) except for very wide lists, etc.

User begins by logging onto VAX/VMS. After completion of login, user types "seedis" in response to the VMS "$" prompt in order to access SEEDIS. Welcoming message notes standard commands that can be used throughout SEEDIS.

```
$ seedis
                    ------------------------------------------------
                    Welcome to SEEDIS, version 1.1
                    ------------------------------------------------
                    At any point in Seedis, you can type the following
                    to get these services:

?                   list commands in this menu
help                explain commands in this menu
show                list or explain items to be selected
review              list current status
cancel              delete current menu selection
quit                return to previous menu

                    If you exit abnormally, type this to recover:
@restore
                    Then to restart Seedis, type this:
seedis
                    Non-Seedis symbols are now being put aside.
                    Your menu prompt will be here shortly.
```

Normal SEEDIS menu format is "SUBPROGRAM: available commands." Standard commands are not repeated in every menu, but can be invoked at most points in SEEDIS. The help command gives a two or three paragraph explanation of the current subprogram.

```
                    SEEDIS: area, data, display, profile
: help
                    LBL SEEDIS is a set of program modules for retrieving,
                    analyzing, and displaying selected portions of large data
                    bases. These include a wide variety of geographically
                    linked data on the U.S. population, economy, employment,
                    mortality, air quality, energy production and use, etc.

                    The major processes in SEEDIS are:
Area              - define a geographic study area (composed of states,
                    counties, etc.)
Data              - select data for the geographic study area chosen.
Display           - manipulate and display the data in table, chart, graph,
                    and/or map form.
Profile           - obtain a standard report of selected socio-economic data
                    for the study area already chosen.

                    Normally Area, Data, and Display are performed in the order
                    given. However, once the geographic study area is defined
                    (Area), one may alternate between Display and the selection
                    and extraction of additional items in Data.
```

Appendix A (continued)

After completing a help message, SEEDIS repeats the previous prompt. At any point, the user can type "?" to see a complete list of commands that can be invoked at that point in the session, with brief descriptions.

```
                    SEEDIS: area, data, display, profile
: ?
Input               Description
------------------- ------------------------------------------------------
area                select geographic area (level and scope)
data                select, extract, or enter data
display             display data
profile             make standard socio-economic report

?                   list available commands in this menu
help                describe how to use Seedis
show                list Seedis databases
review              list present status and history
cancel              delete files created by Seedis
quit                leave Seedis and return to operating system
```

User proceeds to specify geographic area selection, state level of analysis, and particular states from which data will be drawn.

```
                    SEEDIS: area, data, display, profile
: area

                    AREA: nation, state, county, <other level>
: state
                    AREA: <state>,us,us+,fr<nn>

:cal
                    AREA: <state>,us,us+,fr<nn>
:nev
```

SEEDIS prints error messages for any unrecognized input and then reprompts.

```
                    AREA: <state>,us,us+,fr<nn>
:hi
                    unrecognized input: HI
                    For more information type ?, show, review, or help.
                    AREA: <state>,us,us+,fr<nn>
:haw
                    AREA: <state>,us,us+,fr<nn>
:ari
```

The user now types "quit" to terminate area selection and proceeds to the next step of selecting data. Note use of "show", one of the global commands, which usually gives information about available selections for items indicated by angle brackets. Here it refers to database codes. The column labelled "access" indicates whether data are currently online or whether a disk pack has to be mounted to access a particular database.

```
                    AREA: <state>,us,us+,fr<nn>
:quit

                    SEEDIS: area, data, display, profile
: data
```

Appendix A (continued)

```
                    DATA: <database code>, extract, newdata
: show
                    DATABASE CODES FOR STATE LEVEL

Code Database Title                              Scope Vars Access
---- ---------------------------------------- ----- ---- -------
AF   1970-77 Population by Age, Sex, Race       US     608 local
AG   MEDX:1970 Population, Hshlds, Centroids    US       6 local
AP   1980 Census Population                     US       3 local
AX   BEA State Econ. Projections, 1969-2030     US    1820 local
R    Revised 1970 Pop'n by Age, Sex, Race       US     138 local
E    Electric Generating Capacity, 1960-1995    US      18 local
AA   1968-72 Age Specific Mortality, Whites     US     198 local
AB   1968-72 Age Specific Mortality, Blacks     US     198 local
AC   Expectation of Life (from 1968-72 data)    US      44 local
AK   1969-77 Leukemia Mortality                 US      36 local
AD   Labor Force and Unemployment, 1974-79      US      72 local
AJ   1971-1976 Employment by Type & Industry    US     284 local
AQ   1970 Census 4th Count, all races           US+   1178 offline
AR   1970 Census 4th Count, whites              US    1178 offline
AS   1970 Census 4th Count, blacks              US    1178 offline
AT   1970 Census 4th Count, other races         US    1178 offline
AU   1970 Census 4th Count, Spanish             US    1178 offline
AW   1980 Presidential Vote                     US       5 local
C    Selected Socioeconomic Data                US      61 local
F    1947-1977 County Data Book                 US    1022 local
G    1974 U. S. Census of Agriculture           US    1200 local
Y    1976 SIE Tabulations (14 tables)           US    4103 local
```

In order to extract data, the user begins by selecting a particular database and specifying its code (in this case "f" for the 1947-1977 County Data Book). Database selection automatically accesses the appropriate online data dictionary to facilitate selection of particular data items.

```
                    DATA: <database code>, extract, newdata
: f
                    COUNTY DATA BOOK DICTIONARY
                         CONSOLIDATED FILE
                       COUNTY DATA 1947-1977

                  SEEDIS Data Base Name - ICCNTDB77

            Source: County and City Data Books 1947-1977
                      U. S. Bureau of the Census
                         Washington, D. C.

  for further information contact

  Fredric Gey
  ph. 415-486-6208
      FTS 451-6208

  08/09/79                    -0-                 revised 09/18/80
                    County Data Book 1947-1977
```

User can browse dictionary in any order. Dictionaries begin with a table of contents, which can be used to quickly find data of interest.

```
                    DATA: <line letter(s)>,table,<page number>,CR
: table

                         TABLE OF CONTENTS

          Description                                        page
     PREFACE (COVER PAGE)                                      0
     INTRODUCTION                                              1
     GEOGRAPHIC DESIGNATIONS                                   2
```

Appendix A (continued)

```
GENERAL POPULATION CHARACTERISTICS, MIGRATION          3
   VITAL STATISTICS, FAMILY                            6
   AGE                                                 8
   INCOME                                              9
   SOCIAL SECURITY AND PUBLIC ASSISTANCE              11
   EDUCATION                                          12
   PRESIDENTIAL VOTE                                  13
   MEDICAL                                            14
   CRIME AND POLICE                                   15
EMPLOYMENT AND LABOR FORCE                            16
   EMPLOYMENT, GENERAL CHARACTERISTICS                16
      JOURNEY TO WORK                                 17
   LOCAL GOVERNMENT EMPLOYMENT                        18
   SOCIAL SECURITY COVERAGE                           19
```

User types "16" to look at page containing data on employment. Line letters appear to the left of each selectable data element displayed on the terminal.

```
                    DATA: <line letter(s)>,table,<page number>,CR
: 16

                    EMPLOYMENT AND LABOR FORCE

             Population 14 Years and over
A!CCDBC0152                          1940
B!CCDBC0153                          1950
             Labor Force. 14 Years and over
C!CCDBC0154                            1940
D!CCDBC0155                            1970 ?????
             Males 14 Years and over Percent in Labor Force
E!CCDBC0156                        1940
F!CCDBC0157                        1950
             Females 14 Years and over Percent in Labor Force
G!CCDBC0158                          1940
H!CCDBC0159                          1950
             Civilian Labor Force
I!CCDBC0160                  1950
J!CCDBC0161                  1960
K!CCDBC0162                  1970
                                Percent Male
L!CCDBC0163                              1960
```

User types a series of letters, separated by commas or a range indicator "-" to select particular sets of data items.

```
                    DATA: <line letter(s)>,table,<page number>,CR
: d,i-k
                        4 data elements have been selected
```

The "review" command displays a list of data items selected thus far. User could specify additional items if s/he wished to do so, but instead types "quit" to proceed with data extraction.

```
                    DATA: <line letter(s)>,table,<page number>,CR
: review
                        4 data elements have been selected for this database
CCDBC0155           CCDBC0160           CCDBC0161           CCDBC0162

                    DATA: <line letter(s)>,table,<page number>,CR
: quit
```

Appendix A (continued)

The "extract" command tells SEEDIS to set aside all data that have been selected and store them in a temporary workfile. Alternatively, the user could have selected additional items from other databases. The "review" command here gives status information about the current session; disk information is given because certain SEEDIS data are on mountable disk packs.

```
                          DATA: <database code>, extract, newdata
: extract

                          DATA: <database code>, extract, newdata
: review
geographic level          STATE
geographic scope
geographic area           some area has been selected

data                      some data have been obtained
data selection            none remain to be extracted

display terminal
map projection

disk packs                SEEDIS001 is on _DRC3:
                          SEEDIS002 is not on line
                          SEEDIS003 is on _DBA3:
                          (If necessary, type $dschedx to schedule disk packs
                          for future use. System mounting is required.)
```

User now proceeds to display the extracted data. User specifies "other" terminal type since s/he is not working at a graphics terminal. User then uses chart subprogram to generate a labelled list of the extracted data.

```
                          DATA: <database code>, extract, newdata
: quit

                          SEEDIS: area, data, display, profile
: display

                          DISPLAY: 4010, 4014, 4027, 4027f, admrg, other
: other

                          DISPLAY: chart, list, map, plot
: chart
 READING   4 ROWS,   6 COLUMNS FROM TABLE
 TYPE PLOT REPORT TO SEE THE DATA
: plot report
```

	FIPS.STATE	CCDBC0155 Labor Yearseand 1970 ????	CCDBC0160 Civilian Forceb1950	CCDBC0161 Civilian Forceo1960	CCDBC0162 Civilian Forceb1970
ARIZONA	4	666809	258216	453988	641000
CALIFORNIA	6	8338493	4237703	6135341	7992168
HAWAII	15	344269	0	218440	294484
NEVADA	32	218097	68206	119842	208996

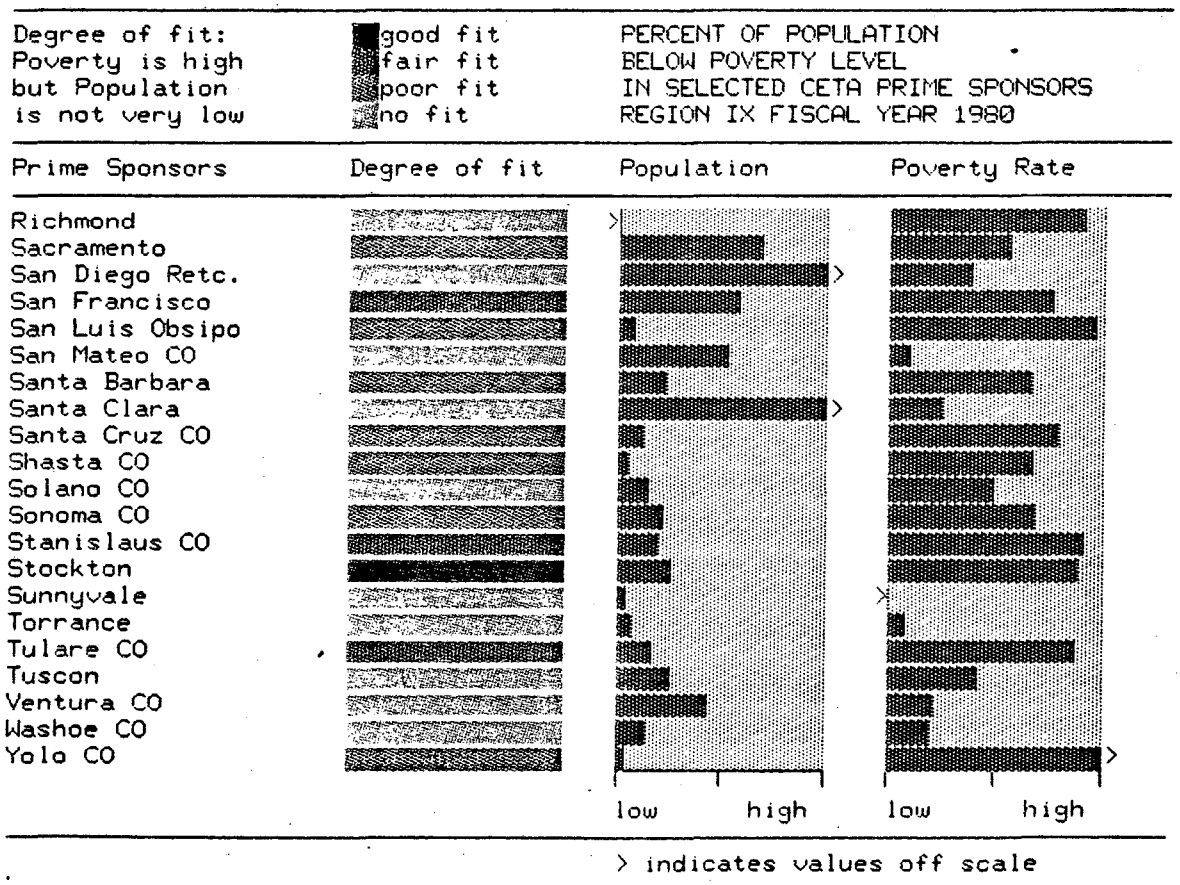
User could now proceed to make maps or charts, do further data extraction or manipulation, etc.

Appendix B Interactive SEEDIS Chart Examples

The following examples demonstrate dot, line, bar, and pie charts that SEEDIS can generate. The exact size and position of titles, figures, and other elements are adjustable. Depending upon the fineness of resolution, the typography, texture and line quality, the final size, and use of color, the images can be used as working quality or final presentation quality charts. The user can determine the ideal format in which to present the data.

A Special Fuzzy Bar Chart

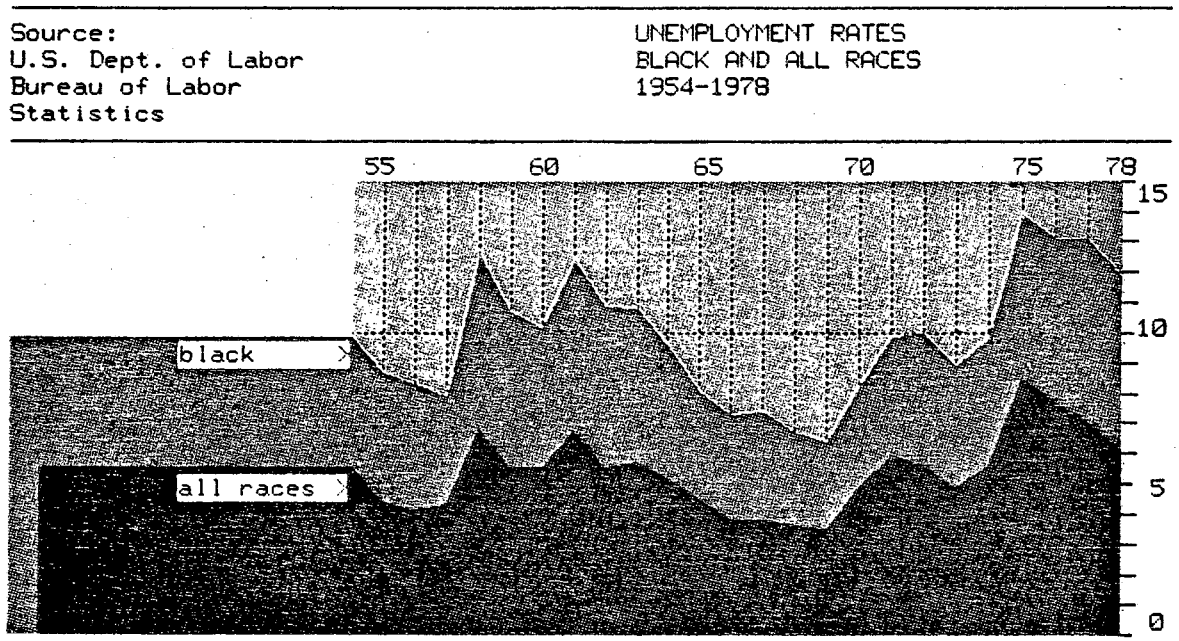
The first example shows a chart that uses fuzzy set theory to combine several imprecisely specified parameters. The degree of best fit for the desired combination is indicated by the gray value of the bars in the left-most column. The original image was displayed on a Tektronix 4027 color terminal and recorded via a laser directly onto a Xerox 6500 color copier. This image is a black-and-white reproduction of the original color xerographic print.



Appendix B (continued)

A Special Line Chart Variation

The next example shows a simple variation of a standard line chart with filled in areas below the lines. Note the overlap between different portions of the display. The original image was displayed on a Tektronix 4027 color terminal and recorded via a laser directly onto a Xerox 6500 color copier. Because the colors used were standard colors available on the Xerox printer, it was not necessary to display them as "dithered" patterns, i.e., overlayed dot combinations of standard colors. Both the previous image and this one are suitable for presentation graphics as color xerographic prints or as color slide images.

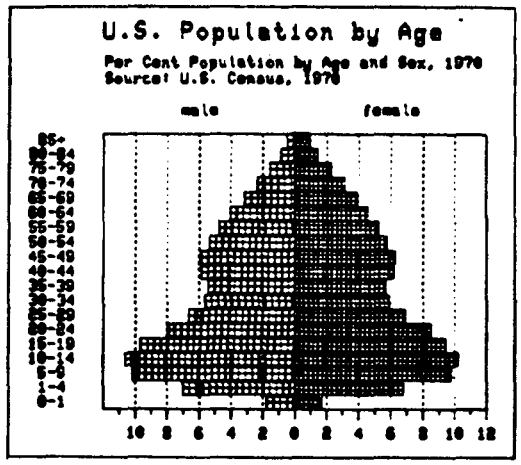


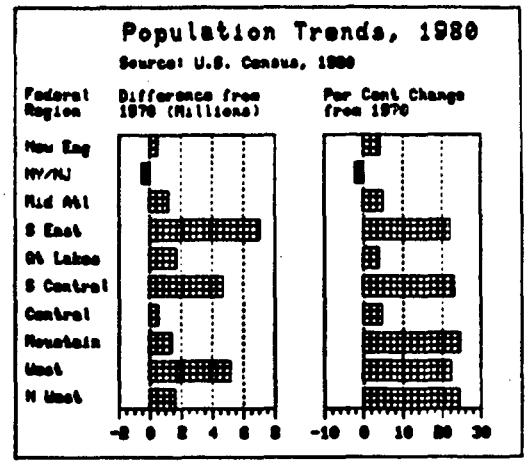
Lawrence Berkeley Laboratory

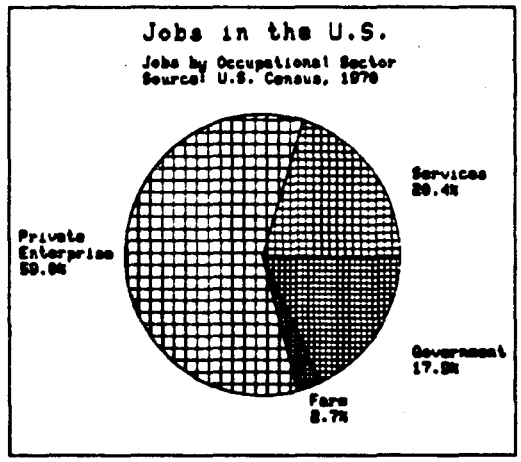
Standard Chart Examples

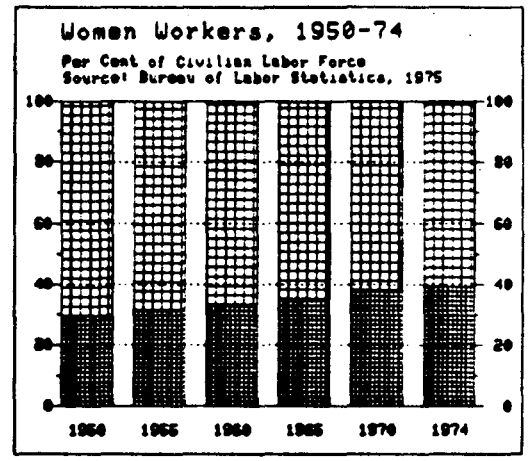
The last set of examples indicate the type of charts which SEEDIS can generate semi-automatically as working quality images. These charts have been slightly redesigned using the graphic editing capabilities of SEEDIS. The images were originally drawn on a Tektronix 4631 hard copy device. They appear here reduced to 64% of their original size. They indicate a sturdiness of typography and line quality, which is important when considering the eventual reproduction of working quality and presentation quality charts.

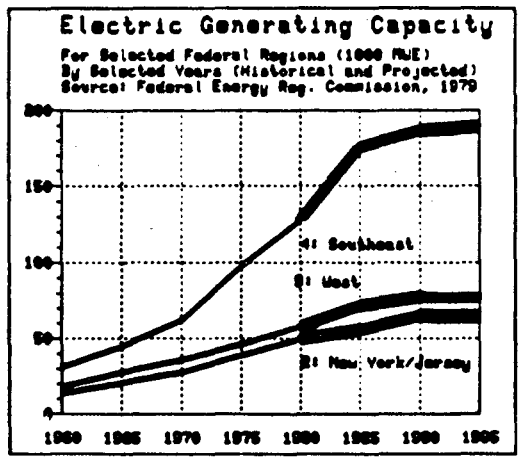
Appendix B (continued)

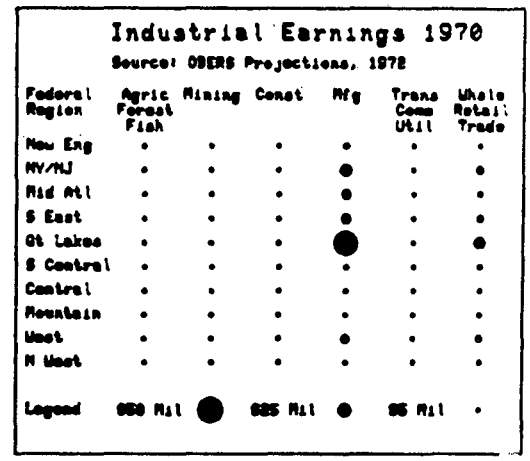










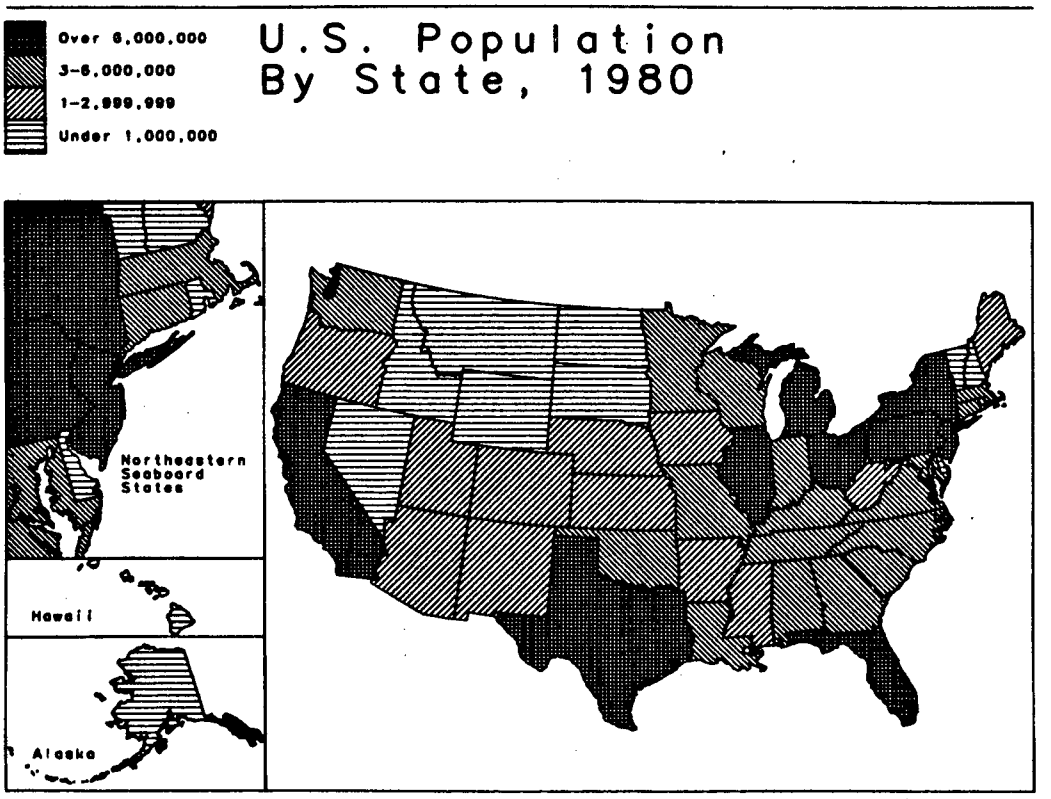


Appendix C Interactive SEEDIS Mapping Examples

The following examples demonstrate area and symbol choropleth maps that SEEDIS can generate. As with the preceding chart images, the exact size and position of titles, figures, and other elements are adjustable. Depending upon the fineness of resolution, the typography, texture and line quality, the final size, and use of color, the images can be used as working quality or final presentation quality charts. The user can determine the ideal format in which to present the data.

An Area Choropleth Map

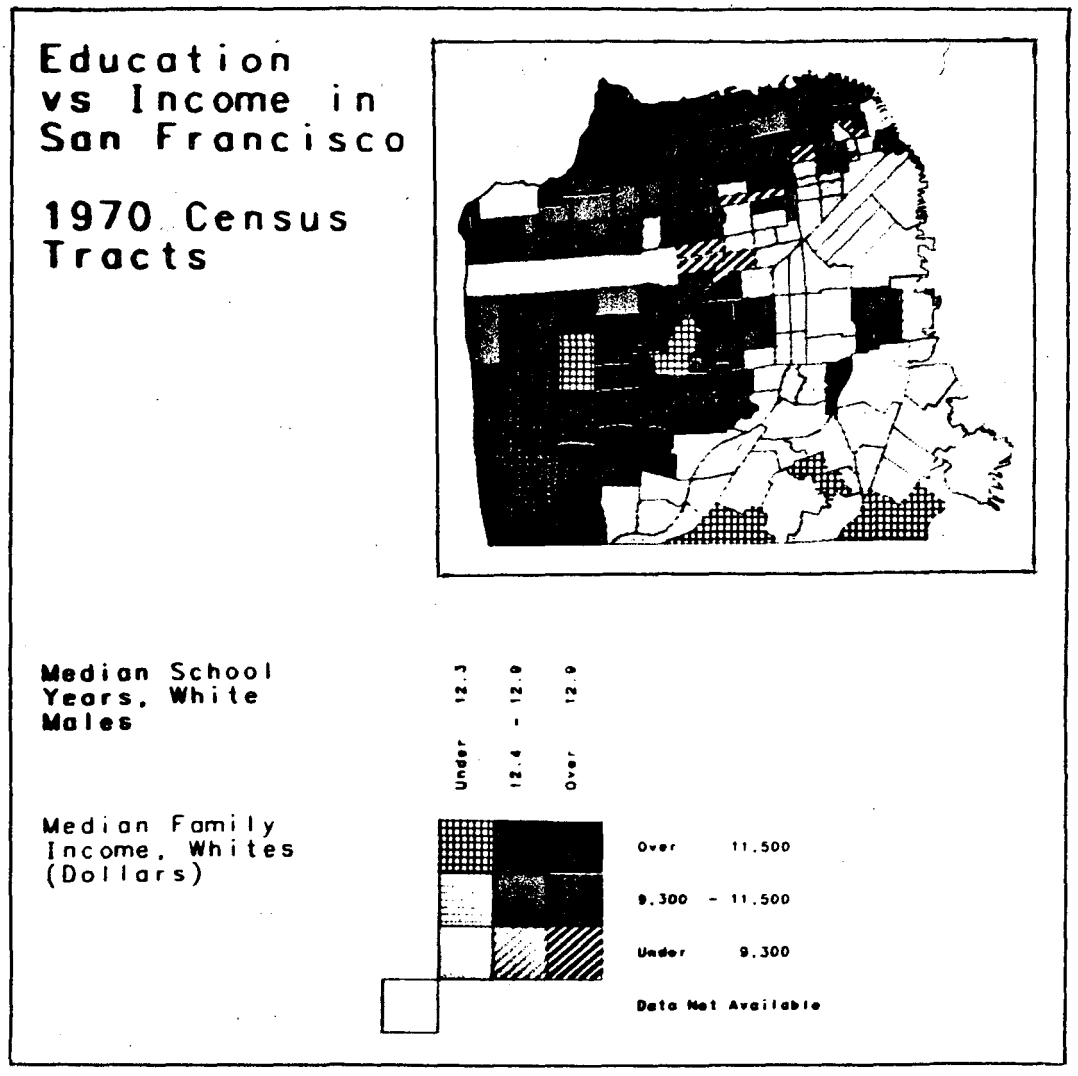
The first example shows an area choropleth map of the United States at the state geographic level. Note the use of varying gray value textures, the ability to create inserts, and software generated typographic characters. These maps were drawn on a Tektronix 4631 hard copy device and appear here reduced to 64% of their original size. The quality of this map makes it suitable for presentation graphics in slide form or in black-and-white printed publications.



Appendix C (continued)

A Bivariate Area Choropleth Map

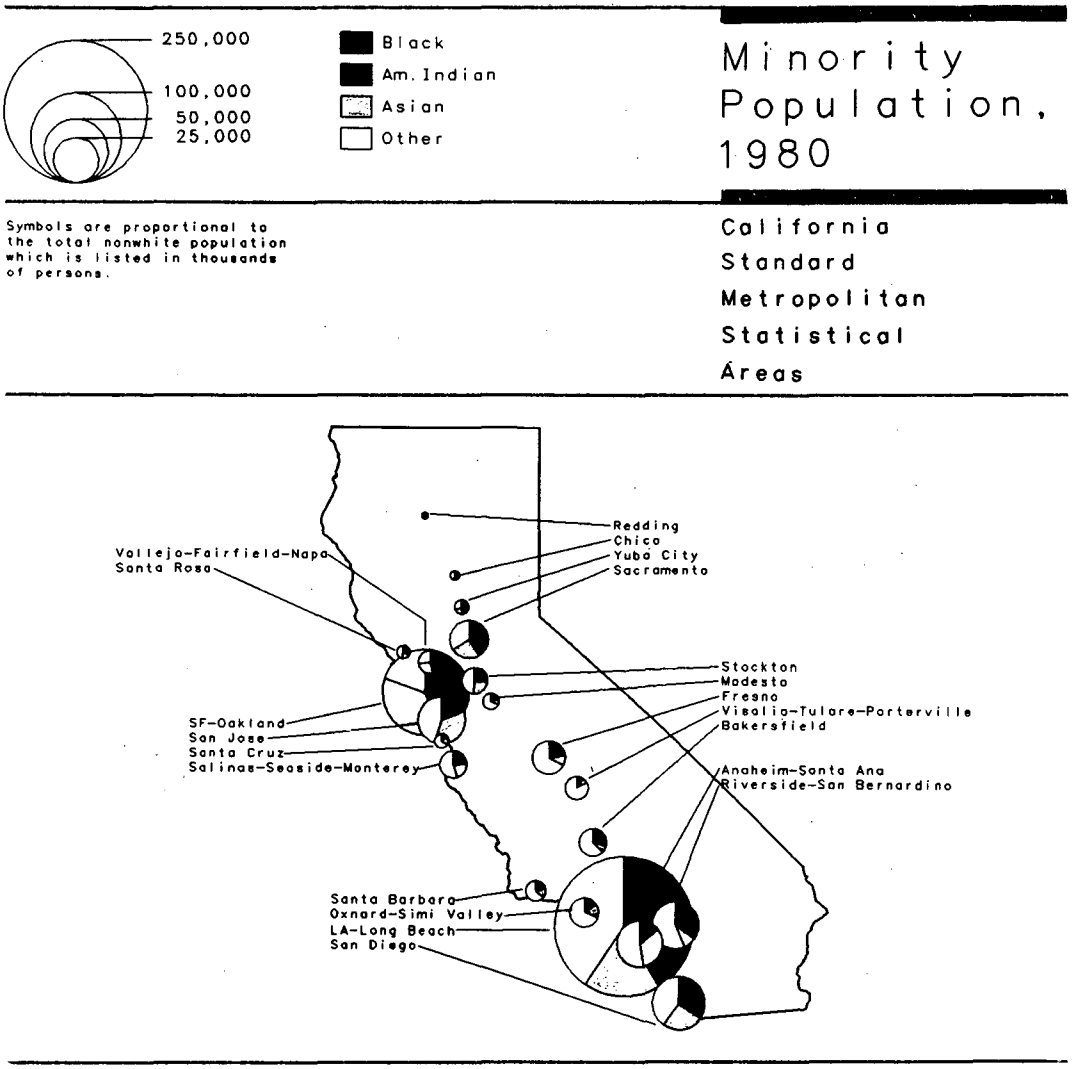
The second example shows a bivariate area choropleth map of San Francisco at a geographic level of 1970 Census tracts. The bivariate choropleth map allows the viewer to determine locations of positive and negative correlation between two parameters. These images can be drawn in color and recorded with a Dicomed film recorder or a Dunn camera on 35 mm. or 4 x 5 inch color film or on 8 x 10 inch Polaroid color prints. This example is a black-and-white reproduction of an original print reduced to 64% of its original size.



Appendix C (continued)

A Symbol Choropleth Map

The final example shows a symbol map that uses pie charts to show multiple population characteristics of California at the geographic level of 1981 Standard Metropolitan Statistical Areas (SMSA81). Note the division of the circle symbol into varying gray values and the overlapping of circles with automatic erasure of underlying circle areas. These images can be drawn in color and recorded on 35 mm. or 4 x 5 inch color film or on 8 x 10 inch Polaroid color prints. This example is a black-and-white reproduction of a print reduced to 64% of its original size.



PUB-424 250/10/81

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

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